The integration of paralinguistic information from the face and the voice

Rebecca Watson

School of Psychology, Institute of Neuroscience and Psychology

University of Glasgow

Submitted for the Degree of Ph.D. to the

Higher Degree Committee of the College of Science and Engineering,

University of Glasgow

September 2012

Abstract

We live in a world which bombards us with a huge amount of sensory information, even if we are not always aware of it. To successfully navigate, function and ultimately survive in our environment we use all of the cues available to us. Furthermore, we actually *combine* this information: doing so allows us not only to construct a richer percept of the objects around us, but actually increases the reliability of our decisions and sensory estimates. However, at odds with our naturally multisensory awareness of our surroundings, the literature addressing unisensory processes has always far exceeded that which examines the multimodal nature of perception.

Arguably the most salient and relevant stimuli in our environment are other people. Our species is not designed to operate alone, and so we have evolved to be especially skilled in all those things which enable effective social interaction – this could be engaging in conversation, but equally as well recognising a family member, or understanding the current emotional state of a friend, and adjusting our behaviour appropriately. In particular, the face and the voice both provide us with a wealth of hugely relevant social information - linguistic, but also non-linguistic. In line with work conducted in other fields of multisensory perception, research on face and voice perception has mainly concentrated on each of these modalities independently, particularly face perception. Furthermore, the work that has addressed integration of these two sources by and large has concentrated on the audiovisual nature of speech perception.

The work in this thesis is based on a theoretical model of voice perception which not only proposed a serial processing pathway of vocal information, but also emphasised the

similarities between face and voice processing, suggesting that this information may interact. Significantly, these interactions were not just confined to speech processing, but rather encompassed all forms of information processing, whether this was linguistic or paralinguistic. Therefore, in this thesis, I concentrate on the interactions between, and integration of face-voice paralinguistic information.

In **Chapter 3** we conducted a general investigation of neural face-voice integration. A number of studies have attempted to identify the cerebral regions in which information from the face and voice combines; however, in addition to a large number of regions being proposed as integration sites, it is not known whether these regions are selective in the binding of these socially relevant stimuli. We identified firstly regions in the bilateral superior temporal sulcus (STS) which showed an increased response to person-related information – whether this was faces, voices, or faces and voices combined – in comparison to information from objects. A subsection of this region in the right posterior superior temporal sulcus (pSTS) also produced a significantly stronger response to audiovisual as compared to unimodal information. We therefore propose this as a potential people-selective, integrative region. Furthermore, a large portion of the right pSTS was also observed to be people-selective and heteromodal: that is, both auditory and visual information provoked a significant response above baseline. These results underline the importance of the STS region in social communication.

Chapter 4 moved on to study the audiovisual perception of gender. Using a set of novel stimuli – which were not only dynamic but also morphed in both modalities – we investigated whether different combinations of gender information in the face and voice could affect participants' perception of gender. We found that participants indeed combined both sources of information when categorising gender, with their decision being

reflective of information contained in both modalities. However, this combination was not entirely equal: in this experiment, gender information from the voice appeared to dominate over that from the face, exerting a stronger modulating effect on categorisation. This result was supported by the findings from conditions which directed to attention, where we observed participants were able to ignore face but not voice information; and also reaction times results, where latencies were generally a reflection of voice morph. Overall, these results support interactions between face and voice in gender perception, but demonstrate that (due to a number of probable factors) one modality can exert more influence than another.

Finally, in **Chapter 5** we investigated the proposed interactions between affective content in the face and voice. Specifically, we used a 'continuous carry-over' design – again in conjunction with dynamic, morphed stimuli – which allowed us to investigate not only 'direct' effects of different sets of audiovisual stimuli (e.g., congruent, incongruent), but also adaptation effects (in particular, the effect of emotion expressed in one modality upon the response to emotion expressed in another modality). Parallel to behavioural results, which showed that the crossmodal context affected the time taken to categorise emotion, we observed a significant crossmodal effect in the right pSTS, which was independent of any within-modality adaptation. We propose that this result provides strong evidence that this region may be composed of similarly multisensory neurons, as opposed to two sets of interdigitised neurons responsive to information from one modality or the other. Furthermore, an analysis investigating stimulus congruence showed that the degree of incongruence modulated activity across the right STS, further inferring neural response in this region can be altered depending on the particular combination of affective information contained within the face and voice. Overall, both behavioural and cerebral results from this study suggested that participants integrated emotion from the face and voice.

Contents

List of Tables

List of Figures

Acknowledgements

There are a number of people who have been of particular importance throughout my Ph.D, and I would like to thank them here.

This thesis would have been in no way possible without the help, guidance and understanding of my supervisor Professor Pascal Belin. I will always be grateful for the support he has shown me.

Additionally, my opportunity to conduct research on this scale is due to the financial aid of the Biotechnology and Biological Sciences Research Council (BBSRC), and so to them I am indebted.

The Voice Neurocognition Laboratory has acted as a 'cocoon' within the School of Psychology: with the assistance, kindness and patience of all its members, past and present, I have learnt a tremendous amount. Special thanks have to go to Dr. Ian Charest and Dr. Marianne Latinus - the former who took me under his wing as an undergraduate student, and helped build the foundations for this thesis; and the latter who has become a much valued 'sounding-board' for an increasingly endless stream of questions and ideas.

I would also like to thank the School of Psychology and the University of Glasgow for providing the outstanding facilities required to achieve this thesis, and for never making me think twice about undertaking my PhD here.

Finally, I thank my friends, both within and outwith university; and also my family, for their support.

Author's Declaration

This thesis has been composed by the undersigned. It has not been accepted in any previous application for a degree. The work, of which this thesis is a record, has been completed by myself, unless otherwise indicated in the text. I further state that no part of this thesis has already been, or is concurrently, submitted for any such degree or qualification at any other university.

.....

Rebecca Watson

1. General introduction

Imagine you are reading a book in the library, and two people start loudly whispering next to you. You may skip a line, have to re-read a paragraph, or perhaps think you have read some of the words that you have actually heard in the conversation. The conversation catches your interest, and you start to eavesdrop. Whilst doing so, you'll probably glance up from your book once or twice. Why? Such an ordinary example illustrates the strong interactions that exist between our visual and auditory systems - interactions that, for the most part, go unnoticed. Multisensory information can help us better perceive the surrounding world, which is why you were most likely tempted to look up from your book while listening in on that conversation – you were looking for visual cues to help you understand more clearly what you were hearing. However, information from other sensory modes can also be distracting in some cases, which could also explain why your reading became interrupted when the conversation started beside you.

Multisensory processing can be defined as the influence of one sensory modality on activity generated by another modality (Stein and Meredith, 1993; Meredith and Clemo, 2010). Although sensory processing and perception have been studied for decades in both psychology and neuroscience, and multisensory behavioural illusions and effects were reported as early as the '60s and '70s, traditional studies on perceptual processes primarily investigated sensory modalities in isolation. This research focus contrasts somewhat with our usual perceptual experience, where events around us nearly always stimulate several of our senses concurrently.

Fundamentally, perception is a multisensory phenomenon. In the everyday environment we are exposed to a constant flow of sensory information and the human brain is organised so that it can combine information from various sensory channels in order to enhance detecting, identifying, and responding to objects and events. We are programmed to combine relevant and complementary cues from different modalities - separating these pertinent events from unrelated background noise filters out irrelevant information and allows us to increase the reliability of our sensory estimates. This task should not be underestimated: the brain has to try and keep signals from different events separate, but yet combine the signals of different modalities that originate from a common event. Ultimately, to truly understand sensory perception we must work out how information received from one sense can be modulated by information concurrently processed via the other senses.

The traditional focus on sensory modalities as independent entities has become increasingly extended to the study of multisensory processing: huge inroads have been made over the past few decades in our understanding of multisensory processing at the behavioural, neurophysiological and cerebral levels. Furthermore, research on audiovisual person perception – specifically, face-voice integration, the manner of multisensory integration that this thesis focuses upon – has, in particular, exploded in the last several years. This research has enabled us to not only gain a clearer picture of how these processes work in the brain, but also to relate the results to realistic, everyday situations in which auditory and visual events hardly ever occur in isolation.

1.1 Stein and Meredith and their three founding rules of multisensory integration

The Merging of the Senses (Stein and Meredith, 1993) was arguably the real instigator of growth in multisensory processing research, especially in neuroscience. This book describes a series of experiments – mainly conducted by the authors - that investigated the multisensory nature of neurons in the superior colliculus (SC) of the cat, a structure mainly concerned with orientation and attentive behaviours. Subcortically, deep layers of the SC receive input from not only the visual cortex, but also auditory and somatosensory areas (e.g. Meredith and Stein, 1983; Stein, 1976): approximately 60% of the neurons tested in this subcortical structure were found to respond to visual, auditory and somatosensory stimuli. This is in contrast to other areas of the brain specialised to process information from a specific sensory modality, which receive projections mostly from the sensory modality in question (for example, the primary visual cortex (V1) receives information from the retina via the lateral geniculate nucleus (LGN), and the primary somatosensory cortex (S1) receives information from skin receptors). The authors reported that for there to be a true synthesis of sensory information, the response to a multisensory stimulus must differ from all of those elicited by its modality specific components. Thus, at the single neuron level, multisensory integration is defined operationally as a statistically significant difference between the number of impulses that are evoked by a crossmodal combination of stimuli and the number of impulses evoked by the most effective of these stimuli individually (Meredith and Stein, 1983; Stein and Stanford, 2008).

At this point, it might be useful to define the different types of neurons that are found in multisensory brain regions. Contrary to what one might expect, these areas generally do not consist of a homogenous set of multisensory, integrative neurons; rather, multisensory

regions are usually composed of a mix of neurons with different selective properties. The first class of neuron is ***unisensory***. These neurons produce significant neural activity (measured as an increase in spike count above baseline) with only one modality of sensory input, and this response is not modulated by concurrent input from any other sensory modality.

The second class of neuron is ***bimodal*** (or indeed, trimodal). They produce significant neural activity with two or more unisensory inputs (Meredith and Stein 1983; Stein and Stanford 2008). In other words, if the neuron produces significant activity with both modalities, then it is bimodal. Importantly however, bimodal activation only implies a convergence of sensory inputs, not an *integration* of those inputs (Stein et al., 2009). Bimodal neurons can be further tested for multisensory integration by using multisensory stimuli. When tested with a multisensory stimulus, most bimodal neurons produce activity that is greater than the maximum activity produced with either unisensory stimulus. The criterion usually used to identify multisensory enhancement in neurons is called the 'maximum criterion' or rule (Audiovisual (AV) > Maximum(Auditory(A),Visual(V)). A minority of neurons produce activity that is lower than the maximum criterion, which is considered multisensory suppression. Whether the effect is enhancement or suppression, a change in activity of a neuron when the subject is stimulated through a second sensory channel only occurs if those sensory channels interact. Thus, multisensory enhancement and suppression are indicators that information is being integrated.

The third class of neurons is ***sub-threshold***. These neurons have patterns of activity that look unisensory when they are tested with only unisensory stimuli, but when tested with multisensory stimuli they show multisensory enhancement (Allman and Meredith, 2007; Allman et al., 2008; Meredith and Allman, 2009). For example, a sub-threshold neuron

may produce significant activity with visual stimuli, but not with auditory stimuli. Because it does not respond significantly with both, it cannot be classified as bimodal. However, when tested with combined audiovisual stimuli, the neuron shows multisensory enhancement and thus integration. The three classes of neuron are illustrated in Figure 1.1

A majority of bimodal and sub-threshold neurons show multisensory enhancement (i.e., exceed the maximum criterion when stimulated with a multisensory stimulus); however, neurons that show multisensory enhancement can be further subdivided into those that are super-additive and those that are sub-additive. Super-additive neurons show multisensory activity that exceeds a criterion that is greater than the sum of the unisensory activities (AV > (A + V); Stein and Meredith, 1993). In the case of sub-threshold neurons, neural activity is only elicited by a single unisensory modality; therefore, the criterion for super-additivity is the same as the maximum criterion. This is in contrast to bimodal neurons, in which the criterion for super-additivity is usually much greater than the maximum criterion. Thus, super-additive bimodal neurons show extreme levels of multisensory enhancement. Although bimodal neurons that are super-additive are, by definition, multisensory (because they must also exceed the maximum criterion), the majority of multisensory enhancing neurons are not actually super-additive (Perrault et al., 2003; Stanford et al., 2007). In single-unit studies, super-additivity is not a criterion for identifying multisensory enhancement, but rather is used to classify the *degree* of enhancement. This is in contrast to the super-additive criterion applied to fMRI, where it used to define brain regions as integrative. This criterion, within the context of fMRI research, is described further in **Chapter 2** of this thesis. Sub-additivity refers to the converse, where the multisensory activity is less than the unisensory responses. Experiments in monkeys, cats and rodents have all identified neurons which show response depression, an effect that is sometimes so

powerful that, for example, an auditory stimulus can suppress even a robust visual response (e.g., Meredith and Stein, 1983, 1986; Wallace and Stein, 1996).

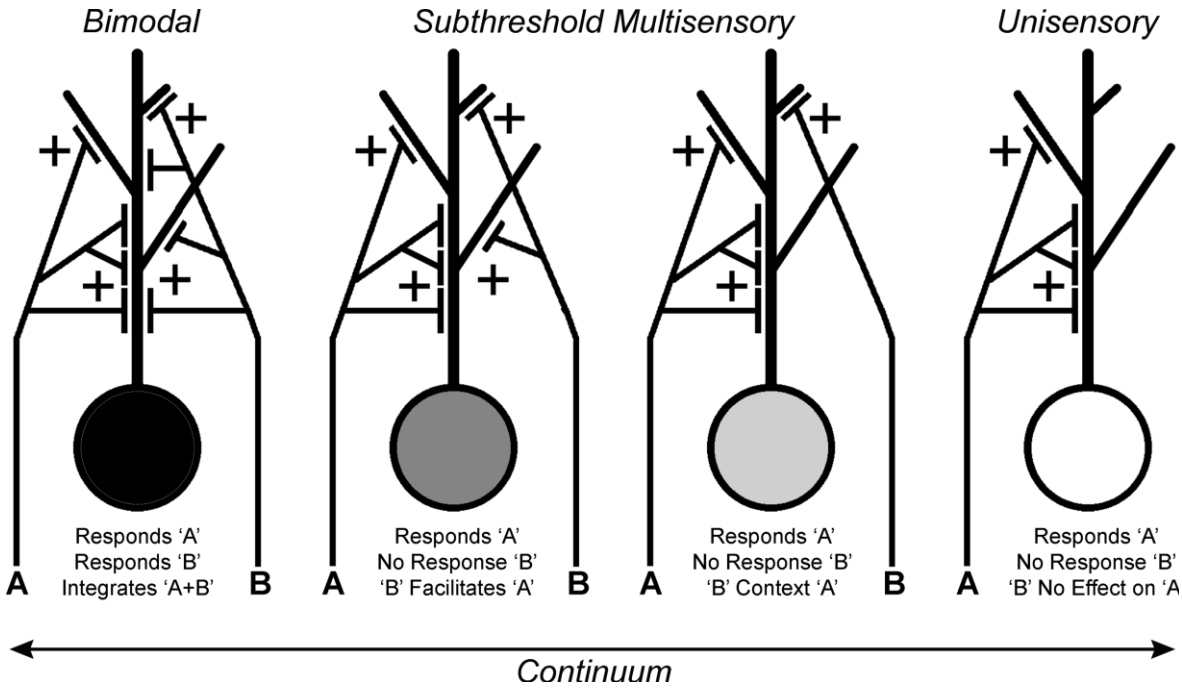


Figure 1.1. Examples of different patterns of sensory convergence onto individual neurons. The black neuron responds to inputs from 'A' and 'B', the definitive property of a bimodal neuron. In addition, this neuron integrates the two modalities. Inputs from modality 'B' are not capable of generating suprathreshold activation in the dark grey neuron; however, it can significantly influence the activity induced by 'A' revealing the sub-threshold multisensory nature of the neuron. For the light grey neuron, the inputs from 'B' are only apparent when combined with inputs from modality 'A' only under specific contexts or conditions. On the right, the white neuron is affected only by inputs from modality 'A,' indicative of a unisensory neuron. Figure taken from Meredith and Clemo (2010).

Throughout their studies, Stein and Meredith also observed that successful integration of multimodal stimuli was also dependent on certain relationships between the unimodal sources, which led them to form three founding rules of multisensory integration – the *spatial rule*, the *temporal rule* and the *law of inverse effectiveness*.

The ***spatial rule*** states that multisensory stimuli are more likely to be integrated when their sources come from approximately the same location, partly due to the overlapping of receptive fields. Multisensory stimuli in spatial correspondence (i.e., from the same event) evoke an increase in firing rate in these multisensory cells and vice versa. On the other hand, cues derived from different events provide inputs from different spatial locations. In such cases, if one of the stimuli falls within the receptive field of a given multisensory neuron, the other stimulus is likely to fall outside the receptive field of the same multisensory neuron. This often results in no interaction between their inputs, and thus no enhancement in the neural signal.

The ***temporal rule*** states that multisensory stimuli are more likely to be integrated when they occur approximately simultaneously, and temporal proximity of multisensory input results in stronger neural activity. The temporal principles regarding multisensory integration are very powerful: changing the interval between two stimuli affects not only the likelihood that an interaction will occur, but its magnitude and quality (enhancement/depression) as well. Meredith et al. (1987) demonstrated that, for example, response enhancement decreased gradually as the visual stimulus preceded the auditory stimulus at progressively longer intervals: a strong multisensory integration effect was seen when the time window between the onsets of auditory and visual events was less than 100ms, with a clear decline of this integration effect when the time windows become progressively larger than 100ms. A further increase in temporal disparity can even result in the cells becoming inhibited.

Finally, the ***law of inverse effectiveness*** states that multisensory stimuli are more likely to be integrated when the best unisensory response is relatively weak. Furthermore, as information obtained from more than one modality is transformed into an integrated

product that differs from the unimodal input, the enhancement in cellular activity induced by congruent multisensory cues is often super-additive (greater than the sum of the individual inputs) (Wallace et al., 1998; Calvert et al., 2000). Spatial, temporal, and other associative relations appear to be critical in constraining selective combination of related subsets from multiple inputs to multiple senses – they help us judge which particular inputs from one sense should be jointly weighted together with a particular selection of inputs from other senses.

1.2 Mechanisms of multisensory integration

The exact means by which one modality is able to influence another is as of yet not fully resolved; however, there is evidence for at least two possible (and co-existing) mechanisms (for a review of this issue see Ghazanfar and Schroeder, 2006). The first involves multisensory integration taking place in higher-order association cortices and structures. Indeed, a traditional view was that sensory integration took place *only* at a late stage of the cortical hierarchical processing scheme. Although multisensory cells have been best characterised in the SC, early electrophysiological research within the animal brain highlighted a number of 'supramodal' regions or convergence zones (Damasio, 1989) where neurons responded to inputs from more than one modality. These areas included both cortical structures such as the insula (Mesulam and Mufson, 1982) and the orbitalfrontal cortex (Jones and Powell, 1970), but also subcortical regions such as the basal ganglia (Chudler et al., 1995), claustrum (Pearson et al., 1982), amygdala (e.g. Turner et al. 1980; McDonald, 1998), thalamus (Mufson and Mesulam, 1984; Blum et al., 1979) and the ventral and lateral intraparietal areas (Lewis and Van Essen, 2000; Linden et al., 1999).

Turning to primate cortical regions, the upper bank of the superior temporal sulcus (STS) has emerged as a crucial integrative region, particular the posterior part (pSTS). This region is known to have bidirectional connections with unisensory auditory and visual cortices (Cusick, 1997; Padberg et al., 2003), and to contain around 23% of multisensory neurons (Barraclough et al., 2005). Ghazanfar et al. (2005) showed that the STS was involved in speech processing when monkeys observed dynamic faces and voices of other monkeys and consistent with findings from animals, the human STS also becomes active when processing multisensory speech information. Furthermore, the human STS also responds to multisensory presentations of letters, tools, and faces and voices (see Hein and Knight (2008) for a review).

However, there has been increasing realisation that interplay between different senses affects not only established multisensory convergence zones, but also brain regions, neural responses and perceptual judgements putatively unisensory (i.e., within the primary visual and auditory cortices). Indeed, animal physiology (Schroeder et al., 2001; Schroeder and Foxe, 2002; Fu et al., 2003; see also Stein and Stanford, 2008), human electrophysiology (e.g., Molholm et al., 2002; Talsma et al., 2007) and human imaging studies (e.g., Calvert et al., 2000) have provided evidence that multisensory integration is not restricted to higher multisensory brain areas, with some traditionally recognised 'modality-specific' brain regions, or early ERP modulations, being influenced by multisensory interplay (e.g. Ghazanfar, 2005; Kayser et al., 2007; Macaluso and Driver, 2005).

Multimodal influence on sensory cortices is proposed to be implemented via two distinct mechanisms. The first of these is direct anatomofunctional coupling between unimodal cortical processing modules. This has been supported by recent work in humans which has observed direct connections between primary auditory and primary visual cortex using

probabilistic tractography (Beer et al., 2011), and previous tracer studies in animals that found direct connections between auditory and visual cortices (e.g., Falchier et al., 2002; Rockland, 2003). The second is a more indirect pathway through areas of AV multisensory convergence (e.g., the STS) to earlier 'unisensory' areas via feedback connections. For example, effective connectivity analysis of fMRI data indicated that visual cortex was influenced by somatosensory cortex via feedback projections from the parietal cortex (Macaluso, 2000). However, it should be noted that these two possible anatomical mechanisms are not mutually exclusive and the anatomical connectivity data actually highlights direct connections between primary sensory cortices *and* higher-order association regions (e.g. Beer et al., 2011). Similarly, multisensory integration can occur at both an early (i.e., unisensory cortices) and late ('supramodal' regions) stage of stimulus processing. All in all, the anatomical and functional instantiation of multisensory processes would appear to involve connections between many nodes of a large network consisting of primary-sensory and higher-order regions of the brain, as illustrated in Figure 1.2.

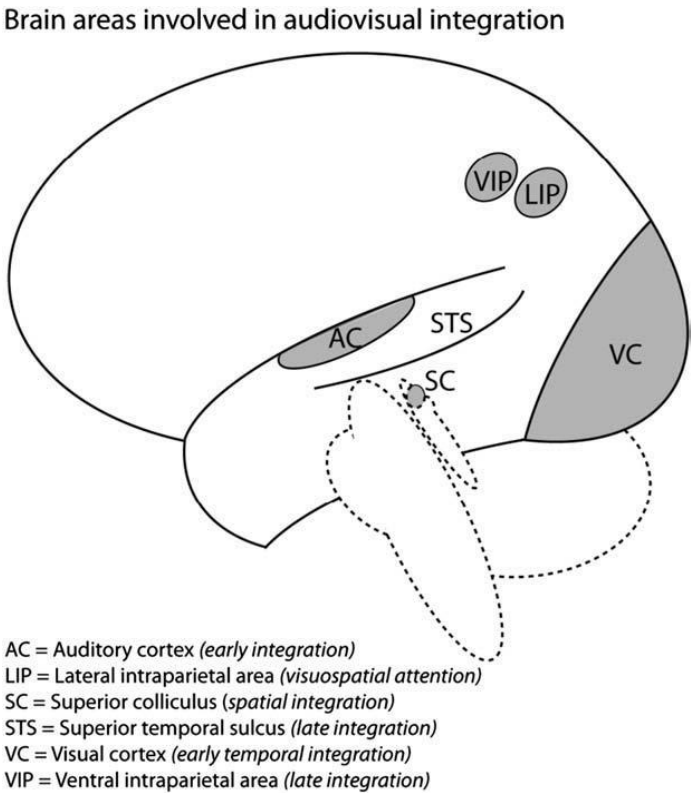


Figure 1.2. Brain areas that are involved in audiovisual integration (one type of multisensory processing). Figure taken from Koelewijn et al. (2010).

Furthermore, an important point to note is that different approaches have used different criteria in order to define a region as a multisensory convergence zone, including different statistical approaches, which are described in more detail in **Chapter 2** of this thesis. Anatomical studies generally test for traceable connections with sensory-specific areas, for multiple modalities. Physiological, single-cell studies examine the presence of responses to more than one modality when each is stimulated separately, and also the responses during multisensory stimulation compared to a unimodal baseline. Finally, neuroimaging studies - for example, functional magnetic resonance imaging (fMRI) and positron emission topography (PET) - inherently assess a more 'general' level of large-scale neuronal populations. These differences could, in part, contribute to the variation of proposed multisensory regions and mechanisms within studies in this field.

Before moving on to discuss integration of face and voice information – the type of multisensory processing this thesis focusses upon – I will briefly discuss attentional aspects of multisensory integration. Relevant to this thesis, in **Chapters 3** and **5** we used paradigms which allowed for any modality dominance (a much researched aspect of crossmodal attention) to be highlighted; furthermore, in **Chapter 3** we also explicitly manipulated attention. Thus, I believe it is valuable to outline the main aspects of attention research with regards to multisensory processing.

1.3 Attention and multisensory processing

Multisensory interactions can also exist at an attentional level in which, for example, a sound draws our visual attention to a certain location (e.g., Spence and Driver, 1997, Driver and Spence, 1998). Attention refers to those processes that allow for the selective processing of incoming sensory stimuli. Selective attention is the mechanism that allows us to focus on an important input whilst ignoring unimportant events, and this helps us to prioritise those stimuli that are most relevant to achieving our current goals and/or to performing the task at hand.

Attentional processing can occur in a bottom-up (exogenous) manner – in this case, an object gets selected even though the person was not planning to select it, often by the unexpected occurrence of an particular event (e.g., someone calling our name across the street). In other cases, attention occurs in a top-down (endogenous) manner in which the person voluntarily controls what is attended and what is not (e.g., listening to a particular individual whilst at a noisy party) (see Koelewijn et al, 2010).

Although attention research has traditionally considered selection among the competing inputs within just a single sensory modality at a time (most often vision; reviewed in Driver, 2001), the last couple of decades have seen a burgeoning of interest in the existence and nature of crossmodal constraints on our ability to selectively attend to a particular event object or sensory modality (Spence and Driver, 2004). In fact, crossmodal interactions in attention have now been demonstrated between most combinations of different sensory stimuli (Calvert et al., 2004). In crossmodal attention research, attention can be directed to not only one modality in the broader sense, but also specific features – for example, shape and colour in the visual modality, and pitch and amplitude in the auditory modality.

1.3.1 Modality dominance

One of the most fundamental questions in crossmodal attention research concerns the extent to which people can selectively direct their attention toward a particular sensory modality such as, for example, vision, at the expense of the processing of stimuli presented in the other modalities. Indeed, the way we perceive multisensory events reveals that our brain may not give equal weight to the information coming from the different sensory modalities. Rather, one sensory modality can dominate the other.

Over the years, it has frequently been claimed that humans preferentially direct their attentional resources toward the visual modality (e.g., Spence et al., 2001). An everyday example of visual dominance over audition is the 'ventriloquism' effect experienced when watching television and movies, where the voices seem to emanate from the actors' lips rather than from the actual sound source (e.g., Howard et al., 1966; Pick et al. 1969; Alais and Burr 2004; also reviewed in Bertelson and de Gelder, 2004). Another famous example that supports the notion of an attentional account of visual dominance is the ***Colavita effect***

(reviewed in Spence, 2009). Colavita reported that while people find it easy to make speeded modality discrimination/detection responses to auditory and visual stimuli when they are presented in isolation, they often fail to respond to auditory stimuli when they are presented at the same time as visual targets (Colavita, 1974; Colavita et al., 1976).

The Colavita effect has proven to be a robust phenomenon that endures many experimental manipulations. For instance, the visual dominance persists despite matching the subjective intensity of the two stimuli, or doubling the subjective intensity of the tone relative to that of the light (Colavita, 1974). The effect remained regardless of whether unisensory auditory responses were slower than unisensory visual responses or vice versa, and also continued when the probabilities of the uni- and bisensory trials, within experimental blocks, were varied (although higher probabilities of bisensory stimuli reduced the magnitude of the effect; Koppen and Spence, 2007). Furthermore, the visual dominance also persists irrespective of the semantic congruence/incongruence between the auditory and the visual stimuli in the bisensory trials (Koppen et al., 2008).

Overall, there are now many examples in the literature demonstrating vision's dominance over audition. Audition has, however, been shown to dominate over (or modulate) vision in several other tasks. For example, a single flash of light accompanied by multiple auditory beeps is perceived as multiple flashes (Shams et al., 2000). Similarly, in a study by Bresciani et al. (2008), participants presented with simultaneous sequences of flashes, taps and beeps were instructed to count the number of events presented in one target modality, and to ignore the stimuli presented in the other 'background' modalities as the number of events presented in the background sequence could differ from the target sequence. Results showed that vision was the most susceptible to influence from background information, and the least efficient in biasing the other two senses. By contrast, audition was the least

susceptible to background-evoked bias and the most efficient in biasing the other two senses. In general, it seems that spatial tasks typically result in visual dominance whereas temporal tasks more often result in auditory dominance.

Ernst and Banks (2002) propose *maximum likelihood estimation* to account for many of the findings from sensory dominance research. In this account, which sense dominates the other in any given situation depends on the variance associated with each perceptual estimate. They suggest that such perceptual estimates may be 'optimal' in the sense that each modality's estimate is weighted by its reliability/variability. Thus, our brain appears to integrate noisy sensory inputs such that the variance associated with the multisensory estimate is maximally reduced.

1.3.2 How do multisensory integration and attention interact?

The fact that multisensory integration can occur in a number of different brain areas at different processing stages (i.e., sub-cortical areas (e.g., SC), early cortical regions (i.e., primary visual and auditory cortices), and higher cortical areas (e.g., STS and intraparietal areas)) raises the possibility for interactions with attention at different levels. Yet, how – and whether – this interaction occurs is still under debate. This is partly due to the fact that this is dependent on the stage at, and the mechanism by which multisensory integration itself takes place, which remains not completely understood.

One option is that integration occurs at a later stage (the 'late integration framework' (see Koelewijn et al, 2010)). In this model, unimodal attention affects the individual sensory inputs and integrates them at a late stage into a single percept at a higher heteromodal level (e.g. Busse et al., 2005). Thus, attention is required in order for multisensory integration to occur. This model is supported by results from Talsma and Woldorff (2005), who showed

that multisensory integration effects - in the form of enhanced frontal positivity 100 ms after stimulation - was only present for visually attended stimuli (see also Talsma et al., 2007). Similarly, Macaluso and Driver (2001) suggest that similar areas or even similar cells in subcortical areas or primary sensory cortices are responsible for both multisensory integration and crossmodal attention. Furthermore, supramodal areas like the STS are known to play a role in both multisensory integration (Benevento et al., 1977; Bruce et al., 1981) and crossmodal attention (McDonald et al., 2003).

Another alternative is that information is combined at an early, pre-attentive stage, and at a later stage amodal attention is captured. This suggests that multisensory integration is not only independent of attention, but actually drives it (e.g. Vroomen et al., 2001; the 'early-integration framework' (see Koelewijn et al, 2010)). This idea is consistent with work that has shown that attention needs some time to engage before it affects other processes (Woodman and Luck, 1999). The early integration framework is also in line with findings such as the ventriloquism effect, which appears to occur at a pre-attentive stage (Vroomen et al., 2001). Within this model, quickly processed unimodal information influences processing in other sensory modalities via direct connections, and further influences the bottom up processing by enhancing co-occurring information. This enhancement by multisensory integration at a pre-attentive stage can further lead to attentional capture in a situation where the individual events would not capture attention (Santangelo and Spence, 2007).

A third option is that multisensory integration occurs at multiple stages, in a more parallel fashion (the 'parallel integration framework'; Calvert and Thesen, 2004). This model is more flexible, stating that integration can occurs at an early or a late stage, depending on the resources that are available. Koelewijn et al. (2010) suggest the parallel integration

framework may best explain the interactions between multisensory integration and attention. In this framework, some events (e.g. near-threshold) might need attentional resources for integration to occur whilst others do not. If that is the case, integration can only occur at those stages that are sensitive to top–down influences and may occur relatively late in time because it takes time for top–down control to have an effect. However, supra-threshold events may integrate automatically (thus without attention) at an early stage of processing. Overall, this framework appears to draw together best seemingly conflicting results from multiple studies – studies which show that attention is needed for multisensory integration to occur (Talsma et al., 2007) and studies which show that multisensory integration is independent of attention (Vroomen et al., 2001). For a visual representation of each of the attention-integration frameworks, refer to Figure 1.3.

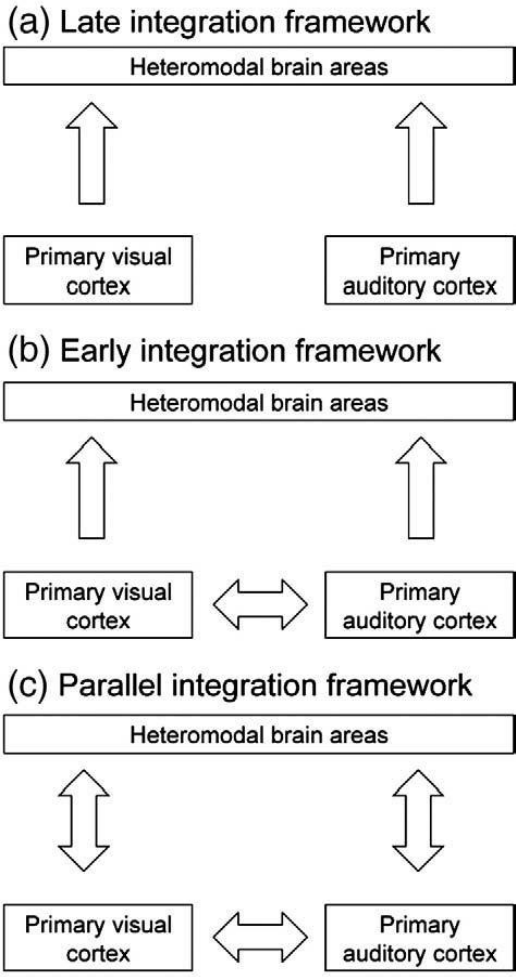


Figure 1.3. Schematic representation of attentional integration frameworks: a) a late integration framework, b) an early integration framework, and c) a parallel integration framework. Figure taken from Koelewijn et al., 2010.

1.3.3 The automatic nature of multisensory integration

This research prompts one further question: to what degree are multisensory interactions automatic? One important criterion a process has to meet in order to be called automatic is the *intentionality criterion* (Jonides, 1981). This criterion states that an automatic process is not affected by voluntary control, suppression, or 'top-down' influences, and is insensitive to the load of task demands (e.g. other competing events). If this were the case, voluntarily or top-down directing of attention to a certain modality or feature would not affect multisensory integration. As described above, some studies have shown that

multisensory integration is indeed modulated by attention (e.g. Talsma et al., 2007; Fairhall and Macalusco, 2009) which would suggest that in general it was not an automatic process; conversely, there is also evidence that early multisensory integration takes place without requiring attentional resources – or in spite of a conflicting attentional demand (e.g. de Gelder and Vroomen, 2000, Vroomen et al., 2001). When seeking to resolve these conflicts it is helpful to refer back to the parallel integration framework: looking at this, we can perhaps think of early and late integration as different processes, of which early integration is automatic and late integration is not. Koelewijn et al. (2010) conclude, based on their review of the literature on crossmodal attention that audiovisual interactions are not pure automatic processes as they do not occur under all circumstances. However, they also point out that multisensory illusions show that when these interactions do occur they can have a strong impact.

1.4 Face-voice integration: a special case of multisensory integration

A special case of multisensory (specifically, audiovisual) integration is when we integrate facial and vocal information of the people around us. Indeed, most of our social interactions involve perceiving, understanding and responding to facial and vocal information, and it is reasonable to assume that at a certain point we must combine this information in some way. By far the majority of work on human subjects has studied integration of linguistic, or speech information. This is understandable, as language - the capability to produce and perceive speech - is a unique capability which sets us apart from other animals, and the exceptional nature of this skill has lead researchers to question whether the multisensory processing underlying perception of speech is somehow different from other domains of pattern recognition. Although this thesis does not focus on audiovisual perception of linguistic information, below I provide a brief overview of the

research conducted on face-voice speech perception: not only has much of this work inspired further research on paralinguistic information integration, but it has provided more insight into possible mechanisms of audiovisual integration. Thus, I believe this description provides a good framework for research described later in this introduction.

1.4.1 Audiovisual speech perception

Most verbal communication occurs when we can both see and hear the speaker. However, speech has often been regarded as a purely auditory process. Throughout the '60s and '70s, the view was that for vision to affect speech perception, acoustic information needed to be suboptimal. Such an effect was demonstrated in an early study conducted by Sumby and Pollack (1954), who found that the integration of audiovisual cues in speech helped understanding of linguistic information when the auditory signal was degraded; a study subsequently supported and extended (e.g. Grant et al., 1998; Ma et al., 2009; Ross et al., 2007). However, this facilitative effect was not believed to occur under normal listening conditions.

This impression changed when it was shown that the perception of certain speech segments could be strongly influenced by vision even when acoustic conditions were normal, and that additionally, some audiovisual pairings (again, with neither signal degraded) could lead to illusory percepts, especially in unusual situations. These were elicited by exploiting the assumption of unity when auditory and visual events simultaneously occur. A classic example of such an effect can be witnessed in the previously described ventriloquist illusion (see Bertelson and de Gelder, 2004), where a sound source is perceived as coming from the same spatial location of approximately time-synchronised visual motion, although the sound is in fact generated by a different source at a slightly different location. Illustrating the example of how the phenomenon got its name, the ventriloquist speaks

without moving his lips, whilst the puppet's mouth moves in approximate synchrony with the heard speech. In the absence of another possible perceptual source of the heard voice, the movements of the puppet's mouth and the heard voice are perceptually combined.

Perhaps the most famous study to demonstrate the potentially illusory nature of speech integration was that of McGurk and MacDonald (1976). The authors highlighted the importance of facial movements for speech perception, showing that particular speech information from the voice and concurrent presentation of incompatible speech information from the face led to illusory percepts. They dubbed a number of syllables (i.e. 'ba', 'ga', 'pa' and 'ka') onto the lip movements of a woman mouthing incongruent syllables. In some cases, subjects reported hearing sounds that were provided neither by the voice alone nor by the movements of the face alone, but involved some combination of the two. In an intriguing example, an auditory 'baba' combined with a visual 'gaga' was often perceived as 'dada'. This phenomenon – the now famous 'McGurk effect' - is known as an auditory-visual 'fusion illusion', an illusion in which the perception is different from information presented in either modality.

The McGurk effect has been shown to be robust, so that even in cases where the face and voice are of different gender, the strength of the McGurk illusion is not affected (Green et al., 1991). Furthermore, it is traditionally considered to be relatively independent of voluntary control, as the illusion remains robust even when participants are informed of the effect (van Wassenhove et al., 2005) and what is more, recent work has shown that this interference is actually bidirectional (Baart and Vroomen, 2010). However, another study found the strength of the McGurk illusion to be reduced when familiar faces and voices of different speakers were combined, suggesting that audiovisual integration in speech

perception may not necessarily be independent of speaker recognition (Walker et al., 1995).

Approximate time-synchronisation of visual and auditory stimuli is important to achieving integrative effects, and synchronisation is often a significant contributor to the percept of 'unity'. Research manipulating asynchrony to test its influence on the McGurk effect (Munhall et al., 1996; van Wassenhove et al., 2007) suggests that there is a small time-window for integration within which the McGurk illusion is most likely to be perceived, with a seemingly greater tolerance for asynchronous presentation when the auditory stimulus lags behind the visual stimulus, in comparison to the auditory stimulus leading the visual stimulus. Audiovisual processing may be predisposed to tolerate such slight asynchronies due to the differing velocities of sound (~340 m/s) and light (~ 300,000,000 m/s). Furthermore, synchrony of the stimuli in the two modalities is naturally variable depending on the distance between the observer and the stimulus: usually the same audiovisual event stimulates the sensory organs with a certain degree of time offset.

There are physical visual clues that offer powerful additional information regarding speech production, and this is partly the reason that visual information can enhance auditory speech processing. For example, Munhall et al. (2004) report that head movements are well temporally aligned with the onset and offset of voicing – thus, there are correspondences between head movements and the dynamic sound pattern over the period of the utterance. In woman and children, the vocal tract is generally shorter than in men, and these gender and age differences are easily identifiable by sight. Additionally, lip reading is useful to both those with defective and normal hearing, as the visible articulators – primarily the lips, teeth and tongue – determine the resonances of the vocal tract. Visible configurations

of the lips, cheek and tongue can allow us to distinguish different speech sounds from one another.

Neuroimaging (typically fMRI) and anatomical evidence has highlighted a network of cerebral regions that are assumed to be involved in audiovisual speech perception. Specifically, results from a number of studies suggest that the audiovisual integration of speech is achieved by processing in a number of key cortical areas, which are also closely connected: the pSTS (Callan et al., 2004; Callan et al., 2003; Calvert et al., 2000; Sekiyama et al., 2003; Skipper et al., 2005; Wright et al., 2003); the auditory (Callan et al., 2004; Calvert et al., 1999; Möttönen et al., 2002) and visual cortices (Calvert et al., 1999); and the speech motor regions (Callan et al., 2004; Callan et al., 2003; Jones and Callan, 2003; Sekiyama et al., 2003; Skipper et al., 2005).

The pSTS has been identified as a brain area involved in audiovisual integration of speech in a number of imaging studies. The pSTS responds to audiovisual speech stimulation (Callan et al., 2004; Callan et al., 2003; Calvert et al., 2000; Sekiyama et al., 2003; Skipper et al., 2005; Wright et al., 2003) as well as auditory (e.g., Binder et al., 2000) and visual speech stimulation (Calvert and Campbell, 2003; Calvert et al., 1997), typically showing a higer degree of activity in response to audivoisual, as compared to unimodal stimulation. Additionally, differential activity between congruent and incongruent information presentation has also been observed in this supramodal region.

Calvert et al. (2000) conducted one of the first studies investigating audiovisual perception of speech. The authors contrasted the response to semantically matching and conflicting audiovisual speech against that to unimodal acoustic and visual speech heard separately. It was presumed that only matching, or congruent sensory inputs would bind together, and

thus the congruent condition was presumed to lead to multisensory integration. Here, the left pSTS alone exhibited significant super-additive (AV > A+V) response enhancement to matching audiovisual speech and sub-additive (AV < A+V) response to conflicting audiovisual speech, leading the authors to propose it as a site for audiovisual speech integration. This result has been supported by Wright et al. (2003), who found both enhanced and suppressed activations in bilateral STS region during observation of matching meaningful audiovisual words in comparison to the unimodal responses; and Skipper et al. (2005) in which the left pSTS was found to be more active during the observation of continuous audiovisual than auditory spoken stories.

Other studies have explored the way in which the brain enhances perceptibility of degraded auditory speech by pairing this with concordant visual speech (Callan et al., 2003; Sekiyama et al., 2003). For example, Callan et al. (2003) found that increased activity in the middle temporal gyrus (MTG) and superior temporal gyrus (STG)/STS was observed when audiovisual speech was presented with acoustic noise in comparison to audiovisual speech with no noise. Similarly, Sekiyama et al. (2005) observed increased activation with added noise in the left pSTS. These results suggest that the responses in the STG/STS region display the principle of inverse effectiveness, with the enhancement of STG/STS activity being greatest when the unimodal acoustic stimulus is the least effective. Another approach has been to investigate how synchrony between audio and visual speech inputs affects audiovisual integration. For example, in a PET study, Macaluso et al. (2004) specifically manipulated the temporal and spatial synchrony of auditory and visual information within audiovisual word presentation. Synchronous versus asynchronous audiovisual speech yielded increased activity in the STS region, but the spatial location of the sound source had no effect on STS activation. This suggests that temporal but not

spatial synchrony of matching auditory and visual speech is critical to integrative effects in STS.

However, it should be noted that some studies have failed to show audiovisual speech integration effects in the STS region. For example, in an fMRI study of the McGurk effect (Jones and Callan, 2003), greater responses in the STS were not observed for matching as compared to conflicting audiovisual stimuli; in fact, the conflicting stimuli activated larger areas of STS region. Furthermore, Olson et al. (2002) did not find enhanced activation in STS region when they compared the BOLD (blood-oxygen-level-dependent) responses to synchronised over desynchronised conflicting audiovisual words producing the McGurk effect.

The differences between studies that support STS as an audiovisual speech integration area and those that do not suggest that the nature of the stimuli (e.g., sentences, compared to brief words or syllables), contrasts between unimodal and audiovisual combinations (e.g., A+V vs. AV, congruent vs. incongruent) and the way integration is manipulated (e.g., , acoustic SNR (signal-to-noise ratio), temporal synchrony) might be important factors in determining whether or not activation is detected in the STS. Additionally, the fact that audiovisual integration studies have reported enhanced activation in superior temporal regions to other stimuli such as sounds and images of tools and animals (Beauchamp et al., 2004; Fuhrmann Alpert et al., 2008) questions whether the STS is speech *specific*, or simply just plays a more general role in multimodal perception.

Audiovisual speech integration has also been observed in 'lower level' areas, particularly within the auditory cortex. fMRI studies have reported response enhancement of the auditory cortex activity by visual speech in the presence of acoustic noise (Callan et al.,

2003) and in comparison to varying levels of degraded visual speech (Callan et al., 2004) during audiovisual speech observation. Furthermore, during audiovisual speech perception, BOLD responses in the auditory cortex as well as in the visual motion cortex are enhanced in comparison to responses during auditory or visual speech stimulation (Calvert et al., 1999). Interestingly, visual speech appears to also be processed in the auditory cortical areas of the STG (e.g., Calvert and Campbell, 2003; Calvert et al., 1999; Calvert et al., 2000; Campbell et al., 2001; Olson et al., 2002; Sekiyama et al., 2003; Wright et al., 2003), and some studies have reported primary and secondary auditory cortex activation by silent lip-reading (Calvert et al., 1997; MacSweeney et al., 2000). Furthermore, an early crossmodal effect was also demonstrated in a study by Besle et al (2004), which found that the behavioural facilitation allowed by audiovisual presentation of stimuli was associated with shorter ERP latencies; and MEG studies have shown that visual speech modifies activity in the auditory cortices during audiovisual speech observation ~50-200 ms after stimulus onset (Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991). Combined, this evidence suggests that audiovisual integration of speech can occur early in the cortical auditory processing hierarchy. With regards to the way the visual speech input is projected to auditory processing areas, it has been proposed that visual speech has access to sensory specific auditory cortex through feedback projections from multisensory neurons in the pSTS (Calvert et al., 2000). In support, there is evidence that responses to visual stimuli in auditory cortex neurons are projected from higher cortical regions (Schroeder and Foxe, 2002; Schroeder et al., 2003).

Finally, activity in brain regions involved with planning and execution of speech production (e.g., Broca's area, premotor cortex (PMC)) has also emerged in studies of audiovisual speech perception (Callan et al., 2004; Callan et al., 2003; Calvert et al., 2000; Jones and Callan, 2003; Olson et al., 2002; Sekiyama et al., 2003; Skipper et al., 2005).

Evidence of the roles of Broca's area and PMC in audiovisual integration of speech comes from studies contrasting audiovisual conditions with different levels of acoustic noise (Callan et al., 2003; Sekiyama et al., 2003) and degraded visual input (Callan et al., 2004). These studies (Callan et al., 2003; Sekiyama et al., 2003). As with the pSTS/STG regions, the speech motor areas seem to follow the principle of inverse effectiveness, where the enhancement of activity during multisensory stimulation is greatest when the unimodal acoustic stimulus is the least effective, displaying inverse effectiveness. The role of motor regions of speech production in audiovisual speech perception is further supported by evidence from non-human primates indicating that the ventral premotor cortex (the monkey homologue of Broca's area) has multisensory properties (Kohler et al., 2002).

1.4.2 Paralinguistic information processing

Alongside conveying linguistic information, faces and voices are both rich in information on a person's biological characteristics, including unique identity and gender, as well as communicating their affective state. We receive and process this paralinguistic information from different sources, but also integrate it into a unified percept: for example, when we see someone smiling and hear them laughing, our conclusion is not 'They look happy' or 'They sound happy', but rather simply 'They *are* happy'. Our ability to do this is a crucial part of social interaction, allowing us to identify our counterparts, as well as inferring their intentions. This further forms a basis for initiating action (e.g. approaching the person if you recognise them; avoiding the person if they appear angry). However, despite our natural, bimodal perception of paralinguistic information, the overwhelming amount of literature in this field has separated visual and auditory processing, preferring to concentrate on unimodal paradigms – particularly, face perception. Below, I briefly describe general behavioural and neural models of unimodal face and voice processing before moving on to discuss face-voice integration of paralinguistic information.

1.4.2.1 Unimodal face processing

It has long been understood that faces are special. From birth we are drawn to faces, and recognising and responding to the information contained within them is something we are especially good at. *Why* this is the case remains debated: for example, the ***domain-specificity*** hypothesis (e.g., Kanwisher, 2000; McKone and Kanwisher, 2005; Yin, 1969) suggests that the 'special' processing used for faces occurs only for faces, emphasising that it has an innate component (de Haan et al., 2002; Morton and Johnson, 1991) and/or that it is necessary to obtain appropriate face experience at a particular time in development (i.e., a 'critical period' during infancy; Le Grand et al., 2001, 2003); in contrast, the ***expertise hypothesis*** (e.g., Diamond and Carey, 1986; Carey, 1992) suggests that 'special' processing for faces is a potentially generic ability that arises for faces because of substantial experience in individual level discrimination and predicts that the special processing can also arise for any other object class through the same mechanism (e.g., bird watchers being able to distinguish between different types of bird). Within the expertise hypothesis the period of life when this experience is obtained is irrelevant: object expertise can be developed entirely as an adult, and the predictor of processing style is merely the amount of practice (Diamond and Carey, 1986; Carey, 1992).

Combined evidence for the 'specialness' of faces has come from three different experimental sources: cognitive psychology, clinical neuroscience, and neuroimaging. Cognitive psychology experiments reveal phenomena such as the face-inversion effect (where presenting a face upside-down dramatically affects its recognition), or the face-composite effect (where judgments about the top halves of two faces are influenced by irrelevant differences in the bottom halves of the faces), that are unique to (or more marked for) faces as compared to other objects (e.g. Yin, 1969). Clinical neuroimaging studies have described patients with selective impairments in the identification of faces (i.e.,

prosopagnosia, a deficit of familiar face recognition (e.g. Behrmann and Moscovitch, 2001)); and neuroimaging techniques including fMRI, event-related potentials (ERPs), magnetoencephalography (MEG) and also single cell recordings and fMRI in primates, have highlighted regions of the visual cortex with high selectivity for faces, some consisting of mostly face-selective neurons.

In particular, there exists a well-documented region in the lateral fusiform gyrus where the activity in response to faces is consistently greater than that evoked by the perception of nonsense (control) stimuli or by non-face objects, which has been named the 'fusiform face area' (FFA; Kanwisher et al., 1997). Support for the nature of this region has come from observations of prosopagnosic patients with lesions encompassing either the right hemisphere or bilateral FFA, a result not seen with object agnosic patients (De Renzi et al., 1994). However, it should be noted that Gauthier and colleagues challenged the notion of the face specificity of the FFA by pointing out that earlier studies failed to equate the level of experience subjects had with non-face objects with the level of experience they had with faces (Gauthier et al., 1997; Gauthier et al., 1999) and also showed that the FFA was activated when car and bird experts were shown pictures of the animals in their area of expertise (Gauthier et al., 2000). Competing evidence regarding the exact nature of the FFA is provided in Grill-Spector et al. (2004) and Rhodes et al. (2004).

Regardless of this controversy, the complexity of the processes involved in face perception is well represented by the cognitive model proposed by Bruce and Young (1986) (Figure 1.4). This model assumes the existence of separate face processing pathways, with one designed to identify the person, and others acting in parallel processing the age, race, gender and emotional expression of the same face. A 'view centred description' is derived from the perceptual input. Simple physical aspects of the face are used to work out age,

gender or basic facial expressions. The route labelled 'directed visual processing' is involved in the direction of attention to a particular face or facial feature. That initial information is used to create a structural model of the face, which allows it to be compared to other faces in memory, and across views. The structurally encoded representation is transferred to notional 'face recognition units' that are used with 'person identity nodes' ('PINs') to identify a person through information from semantic memory. The idea of separate routes for the recognition of facial identity and expression has been supported by studies in cognitive psychology, cognitive neuropsychology, single-cell recording in nonhuman primates and functional imaging.

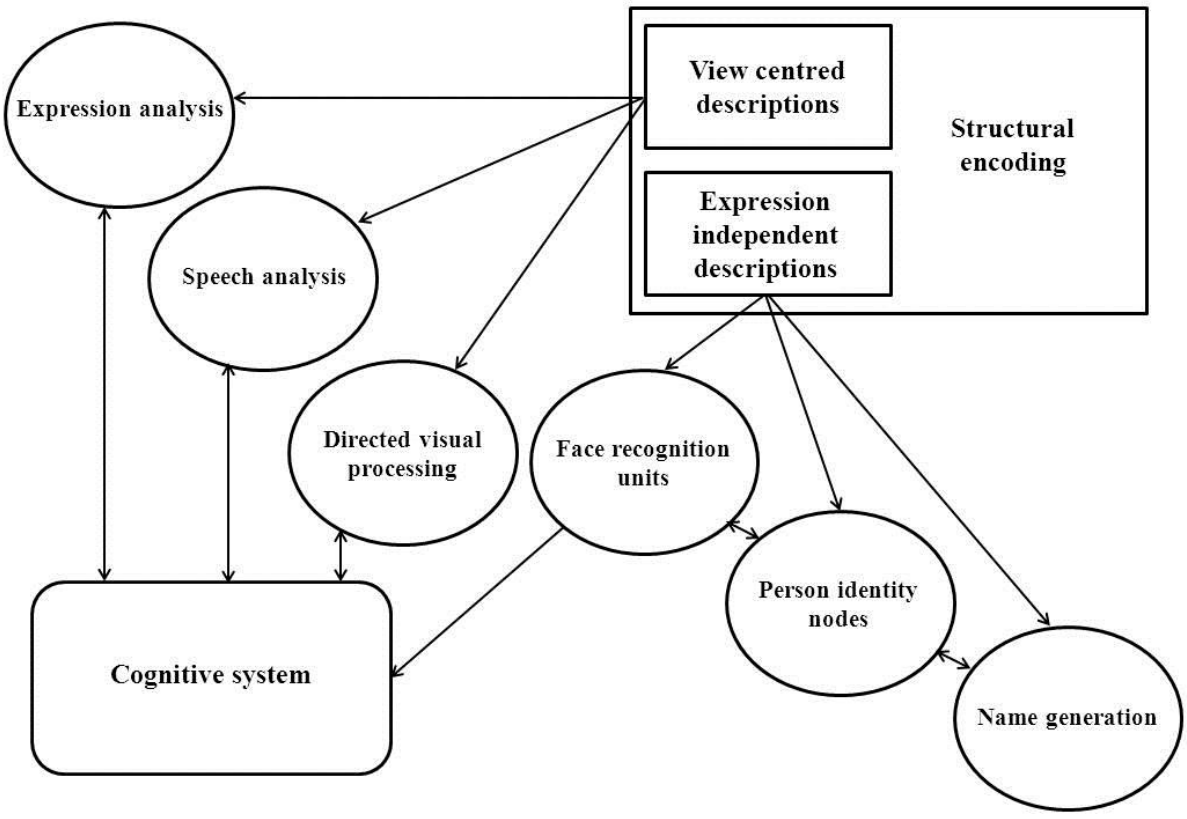


Figure 1.4. A copy of Bruce and Young's 1986 model of face perception.

A neural model corresponding to that of Bruce and Young (1986) was proposed by Haxby et al. (2000) based on a review of single unit and fMRI studies in both humans and

monkeys (e.g., Hoffman and Haxby, 2000; Puce et al., 1996). The model (illustrated in Figure 1.5) assumed that the changeable aspects of faces, such as facial expression, eye-gaze and mouth movement, are processed in STS, whereas the invariant properties of faces such as facial identity were processed in the FFA. The occipital face area (OFA), according to Haxby et al.'s (2000) scheme, receives input from early visual stages and feeds the output to both the FFA and STS. The organisation of this system allows a distinction to be made between the representation of the invariant aspects of faces – which allows us to perceive an individual's identity – and the perception of changeable aspects, such as eye-gaze and speech related movements. This relative independence ensures that a change in expression or a speech-related movement is not interpreted as a change in identity. This model is further supported by temporal and anatomical segregation with evidence from MEG and electroencephalography (EEG) studies in which the early and late signatures in the time course face processing were located in inferior occipital and temporal cortices, respectively (Liu et al. 2002; Smith et al., 2009; Sugase et al., 1999). Electrophysiology in monkeys also suggests a separation in which neurons in STS are tuned to expression and orientation, whereas those in the inferior temporal gyrus (ITG) are tuned to identity (Eifuku et al., 2004; Hasselmo et al., 1989).

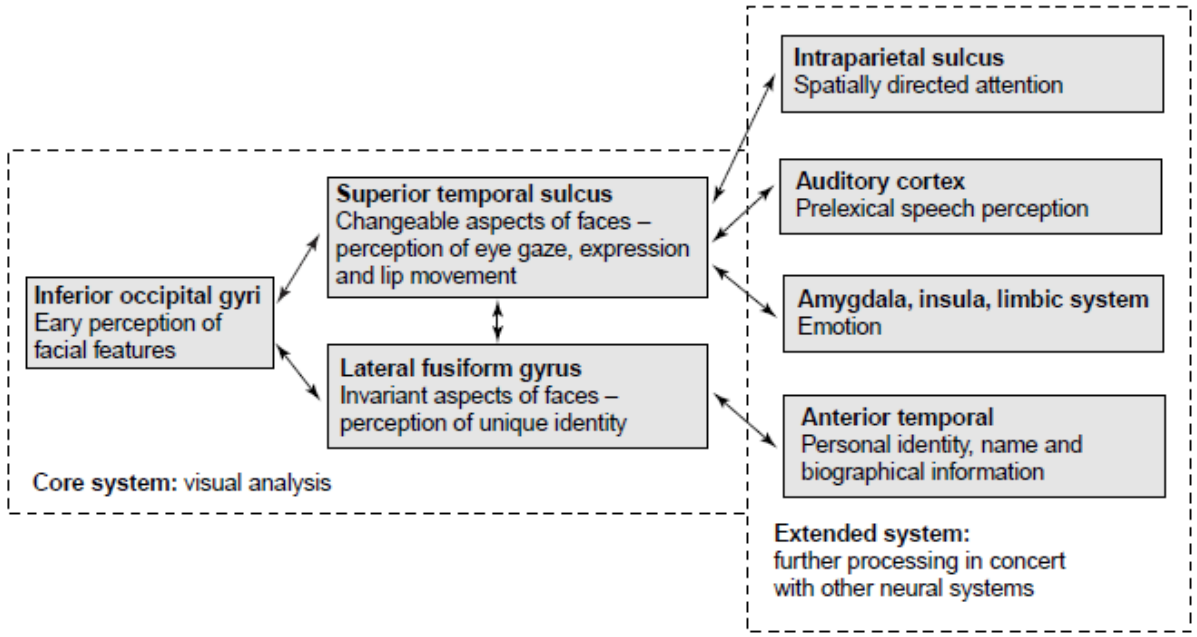


Figure 1.5. A model of the distributed human neural system for face perception. Figure taken from Haxby et al. (2000).

Electrophysiological evidence for face-specific brain processes has been obtained through intracranial recordings (Allison et al., 1999), as well as in many studies using ERPs. These ERP studies have uncovered several components that are linked to different stages in face perception, face recognition, and the processing of emotional facial expression (e.g. Eimer, 2000; Eimer and Holmes, 2007). The earliest, most prominent and by far the most widely studied face-sensitive ERP component is the N170: specifically, when compared to different categories of non-face objects, human faces consistently elicit a larger negative-going ERP component at occipitotemporal electrodes. The presence of an N170 component in response to faces was demonstrated in two early ERP investigations of human face perception (Bentin et al., 1996; Bötzel et al., 1995), and the N170 has since featured prominently in face perception research. There are currently more than 200 published studies that have used this component to investigate different aspects of face processing in the human brain (Calder et al., 2007) although it should be noted that the N170 has also attracted some controversy (Thierry et al., 2007; see also Bentin et al., 2007). More

recently, an 'M170' component with response properties that are very similar but perhaps not identical to the N170 has been identified in experiments that used MEG measures to study face processing (e.g. Halgren et al., 2000; Harris and Nakayama, 2008).

1.4.2.2 Unimodal voice processing

Far less study has been conducted into the perception of vocal information: however, research on voice-specific auditory processing has increased significantly in the past decade. Voices are often referred to 'auditory faces' (Belin et al., 2000; Belin et al., 2011), due to the similarity of the information carried by faces and voices. Although it is not yet as strong and convincing as for faces, similar evidence for voices is accumulating. Evidence for cognitive phenomena specific to voice processing is still elusive, but converging clinical and neuroimaging evidence suggests there are indeed voice-selective cerebral processes.

fMRI studies conducted by my lab group (the Voice Neurocognition Laboratory; *http://vnl.psy.gla.ac.uk*) and several others have demonstrated the existence of voice-selective neuronal populations (e.g., Belin et al., 2000; Ethofer et al., 2009; Grandjean et al., 2005; Linden et al., 2011): these voice-selective regions of cortex (the 'temporal voice areas' (TVA)) are located bilaterally along the mid and anterior parts of superior temporal gyrus (STG)/STS (Figure 1.6). They show greater blood oxygenation (BOLD signal) in response to vocal sounds than to non-vocal sounds from natural sources, or acoustical controls such as amplitude-modulated noise or scrambled voices. Although it is particularly strong for speech sounds, the voice-selective response has also been observed for non-speech sounds (Belin et al., 2002; Charest et al., 2009), showing that the TVA, particularly in the right hemisphere, are not just interested in processing the linguistic content of a voice. The importance of this region has been supported by work with autistic

individuals: these individuals - who show impairments in social interaction and atypical social information processing - failed to activate STS voice-selective regions in response to vocal sounds, whereas they showed a normal activation pattern in response to non-vocal sounds (Gervais et al., 2004)

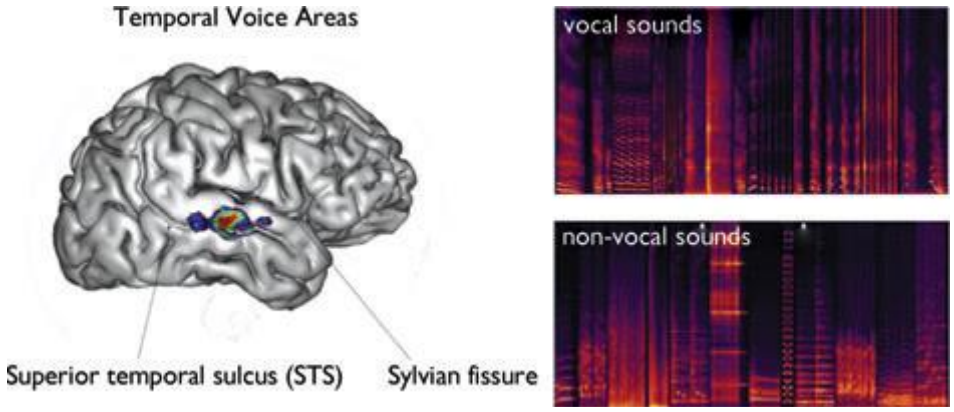


Figure 1.6. Voice-selective cerebral activity. The contrast of cerebral activity measured in the adult brain by functional magnetic resonance imaging (fMRI) in response to auditory stimulation with vocal versus non-vocal sounds highlights voice selective TVA with greater activity in response to the vocal sounds. Figure taken from Belin et al. (2011).

In an influential model of voice perception proposed by Belin et al. (2004) (see also Belin et al., 2011), Bruce and Young's model of cerebral face processing was extended to propose a similar functional architecture for voice processing (Figure 1.7). In summary, this model proposes that after a stage of voice structural encoding restricted to vocal sounds, three partially dissociable functional pathways are proposed to process the three main types of vocal information: speech, identity, and affect.

Referring to the 'auditory face' model of voice processing, an initial low-level analysis occurs in sub-cortical nuclei and core regions of auditory cortex, after which voices are processed in a voice-specific stage of 'structural encoding'. The 'structural' encoding stage

is viewed as being accessed only by vocal stimuli (as are faces in Bruce and Young's (1986) model of face processing). It is at this stage of the functional architecture that a vocal sound would be identified as such; in other words, that it has been produced by a human vocal apparatus.

From that stage onwards, irrespective of the exact nature of the information being the attention's focus, voice stimuli are proposed to recruit processes not activated by other non-vocal sounds. In other words, voices are 'special' for the brain. After structural encoding, the three main types of vocal information are then extracted and further processed in three interacting, but partially dissociable functional pathways: (1) a pathway for analysis of speech information, involving the anterior and pSTS as well as inferior prefrontal regions and the PMC predominantly in the left hemisphere; (2) a pathway for analysis of vocal affective information, involving temporo-medial regions, the anterior insula, and amygdala and inferior prefrontal regions predominantly in the right hemisphere; and (3) a pathway for analysis of vocal identity, involving 'voice recognition units' – probably instantiated in regions of the right anterior STS – each activated by one of the voices known to the person. Integrating all this voice-relevant information would then lead to person recognition (computed within PINs).

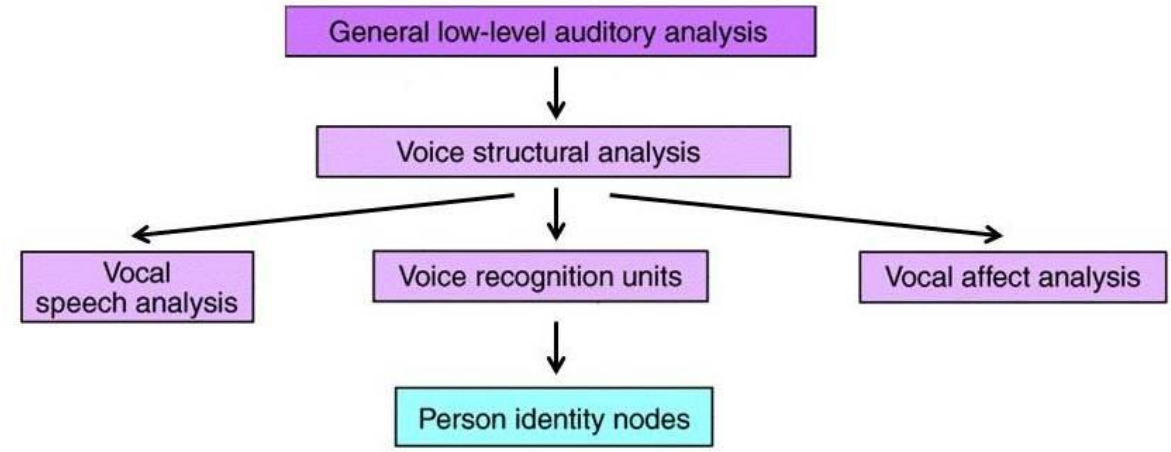


Figure 1.7. A model of voice perception. Figure modified from Belin et al. (2004).

Few ERP studies have compared voices to other stimuli: however, two papers report a positive deflection 320ms after stimulus onset that is larger for voices than for musical instruments (Levy et al., 2001; Levy et al., 2003). This response has been labelled the 'Voice Selective Response'. A recent study which compared voices to a number of non-vocal sounds suggests that vocal discrimination could possibly occur earlier, in the range of the auditory P200 (160-240ms; Charest et al., 2009), at a stage more compatible with face processing.

1.4.2.3 Face-voice integration of paralinguistic information

Until very recently, studies examining the combination of facial and vocal non-linguistic information were scarce. However, over the past few years especially, there has been a surge of studies within this area – particularly those using neuroimaging techniques such as fMRI. This thesis contributes to this new wave of research, by developing upon the pioneering work already accomplished in this young and exciting area. The experimental work conducted is described further in the thesis rationale (to be found at the end of this General Introduction).

As the main focus of this thesis is the integration of paralinguistic information, the rest of the introduction will be devoted to introducing different aspects of that process. Onwards, I provide an overview of recent studies of integration of non-linguistic information from the face and the voice, with a focus on identity, gender and emotion recognition. Generally, work conducted on face-voice identity integration has focussed on person recognition/familiarity. Although this thesis does not directly investigate this aspect of identity perception, I believe it is of value to provide a summary of work conducted in this area: firstly, a description will provide a fuller picture of paralinguistic face-voice perception; secondly, a number of studies conducted in this area have inspired the design and questions of my own research, and so I feel it is worthy to acknowledge these where appropriate.

Integration of face-voice identity information

Clear evidence suggests that healthy individuals are able to combine facial and vocal information in order to decide upon the identity of a person. Identity information from one modality has been found to aid recognition of the same individual presented in the other modality, indicating a cross-modal facilitation effect comparable to that of audiovisual speech integration. In one of the first studies in this area, Ellis et al. (1997) showed that over short time-intervals, crossmodal priming occurred. They demonstrated that the presentation of a familiar voice-prime significantly improved the recognition of a face of corresponding identity presented immediately afterwards. Similar results were also demonstrated for face primes in relation to voice test stimuli. In another study (Schweinberger et al., 1997) participants were presented with samples of famous and unfamiliar voices and were asked to decide whether or not the samples were spoken by a famous person. In different conditions, participants were cued with either a second voice

sample, the occupation, or the initials of the celebrity. The authors found that initials were most effective in eliciting the name.

The link between auditory and visual modalities in identity recognition can again be traced to the mechanics of vocal production, in that a voice conveys physical properties of the speaker's unique facial structure. For instance, overall body size and gender affect both the size of the face and that of the vocal tract, in turn affecting formant frequencies, a salient aspect of voice timbre. The fact that the face comprises the outer surface of the vocal tract means that there is a close connection between vocal and facial information, and it appears that as well as allowing us to 'speech read', this connection can also specify identity across a change in modality.

A study by Kamachi et al. (2003) demonstrated this above effect by showing that participants could match an unfamiliar face to an unfamiliar voice, and vice versa. In their experiment, a face or a voice was presented in a 'first phase', which was followed by two voices (or two faces) in a 'second phase'. The participants' task was to choose which of the stimuli in the second phase corresponded to that presented in the first phase. Stimuli were controlled for gender, ethnicity and age, and different sentences were used across the two phases in order to remove the effect of speech. Participants performed above chance in both the Face – Voice and Voice – Face matching conditions, indicating that common information across modality can be used to match identity. The effect of varying sentence content across modalities was small, showing that identity-specific information was not limited to certain utterances. In a second experiment stimuli were played backwards: voices played backwards are known to be unintelligible, but speaker identity can still be recognised; and similarly, while playing a video backwards affects motion-based recognition, faces can still be recognised from movement-independent image properties.

For both modalities, local properties were left unchanged. Despite identity being untouched in this experiment, when stimuli were played backwards performance dropped to chance levels, indicating that nonlocal auditory and visual spatiotemporal patterns were crucial for this task. The authors also describe a similar experiment that was run with static images, in which performance was also not above chance. The fact that crossmodal matching occurred only when stimuli were played forward, and were moving highlights a real importance of time varying information in integration of identity information.

Similarly, Rosenblum et al. (2006) reported above-chance face to voice matching when only dynamic facial information was presented. Using a point-light technique, where illuminated spots were visible on a face in complete darkness, they were able to isolate facial speech movements. They compared the normal and idiosyncratic speech movements with conditions in which the movements were distorted. They found that face-voice matching was significantly better for the conditions in which the normal facial movements were presented. Rosenblum et al. (2006) highlight particularly clearly the importance of isolated facial movements to the relationship between a speaker's face and voice. Furthermore, voices that have been transformed in the temporal domain can be viewed as an auditory analogue to point-light facial movement displays (Lachs and Pisoni, 2004a; Lachs and Pisoni, 2004b). In these two studies, it was found that face-voice matching is still achievable even when the modalities are significantly degraded.

A study by Sheffert and Olson (2004) provided yet further evidence for the strong links between face and voice identity, this time within a facilitative context. The authors investigated the effects of voice and face information on the perceptual learning of talkers and on long-term memory for spoken words. In the first phase, listeners were trained over a number of days to identify voices from words presented auditorily or audiovisually. The

training data showed that visual information about speakers enhanced voice learning with training performance improving considerably more quickly in the AV condition than in the A condition: participants required a fewer number of training sessions in the AV condition, thus revealing cross-modal connections in talker processing akin to those observed in speech processing. The authors suggest that the additional visual information about the speaker's idiosyncratic speaking style is compatible with the speaker's auditory attributes, and may therefore lead to better encoding of voice identity. Additionally, after the talker training task, the participants completed a generalisation test to determine the extent to which their talker-specific knowledge would transfer to a novel set of words (rather than being tied to the particular training words). Generalisation performance was higher after AV training than after A training, showing that the knowledge acquired from the AV displays generalised to different words and to a different test modality. Furthermore, they found that word recognition was better when words were spoken by familiar speakers compared to words spoken by unfamiliar speakers, which might suggest that speaker and linguistic perception are intertwined.

In the case of familiar people, it is conceivable that multimodal representations of a familiar person's identity may be encoded in long term memory. Schweinberger et al. (2007) provided the first direct evidence that audiovisual integration occurs in the recognition of familiar voices. In their experiment, participants judged whether a standardised sentence was spoken by a personally familiar or unfamiliar voice. They presented the voices either on their own, or in several audiovisual conditions. Specifically, each voice could be combined with a face of either corresponding or non-corresponding identity, and a face could either be dynamic time synchronised or static. The authors found that the recognition of familiar voices was enhanced (i.e., more accurate responses and faster reaction times) when they were combined with a corresponding articulating face. In

contrast, they observed significant performance costs (as measured by both an increase in reaction time and decrease in recognition accuracy) when a familiar voice was matched with a non-corresponding face. Additionally, while these effects were pronounced for familiar voices, they were far smaller or non-existent for unfamiliar voices. The authors suggest that this indicates that the observed audiovisual effects depend on the existence of a previously learned multimodal representation of an individual's identity. Importantly, as in the aforementioned study by Kamachi et al. (2003), effects were significantly larger for dynamic and time-synchronised stimuli, as compared to static stimuli. This again suggests that articulatory movements of the face have much to offer in the way of person identification, in that their effects did not simply reflect participants using the facial identity as a 'cue' for voice recognition. Results from this study are shown below in Figure 1.8.

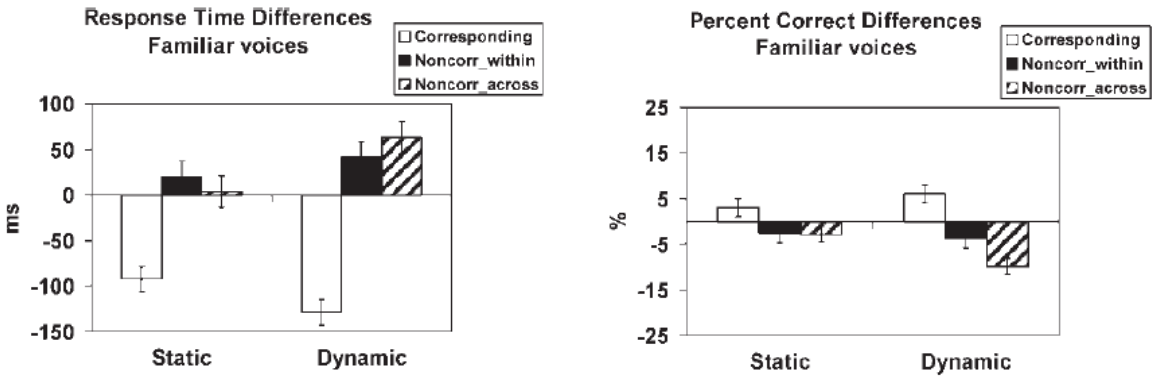


Figure 1.8. Behavioural results from Schweinberger et al. (2007). Reaction time (RT) differences and percentage of correct responses (for which positive values indicate benefits in accuracy, negative values indicate costs) relative to the voice-only baseline condition (in ms) for voice recognition responses when familiar voices were combined with corresponding faces. Figure reproduced from Schweinberger et al. (2007).

With regards to the neural basis of face-voice identity integration, the questions that have been asked of multisensory integration in general have also been applied to this specific field: that is, does the association of information from the two modalities depend upon a relay of information through functionally distinct, supramodal regions, or is it mediated by direct cross-talk between 'unimodal' visual and auditory processing neural systems?

The conventional model assumes that faces and voices are processed separately until the person is identified at a supramodal level of person recognition. Ellis et al. (1997) proposed that there is a separate processing hierarchy for vocal and facial features: voices and faces are first separately encoded on a basic level and thereafter examined for familiarity in voice and face recognition units. These recognition units are assumed to project to PINs (i.e., supramodal nodes) – a semantic representation of information about the identity of a particular person which provide biographical information (e.g., a name) and can be assessed from either facial or vocal information or both. Prosopagnosic and (although less commonly reported) phonagnosic patients, who are impaired in recognising faces and voices respectively, but have spared recognition in the other modality have substantiated the assumption of independent processing hierarchies for faces and voices.

The proposed independent processing streams for faces and voices bears some similarity with initial views of multisensory processing described earlier, which assumed that integration of different sensory inputs takes place solely in multisensory convergence zones (such as the pSTS, the MTG, perirhinal cortex, and intraparietal sulcus) which are known to receive input from sensory specific processing streams (Beauchamp et al., 2004; Calvert et al., 2001; reviewed in Kayser and Logothetis, 2007).

The concept of supramodal convergence zone - or PINs - mediating identity integration was supported by a study by Shah et al. (2001) which investigated the neural correlates of person familiarity with fMRI. Participants were scanned while they viewed personally familiar and unfamiliar faces, and listened to familiar and unfamiliar voices. Changes in neural activity associated with stimulus modality – but irrespective of familiarity – were observed in the fusiform gyrus (FG) and STG: regions which have been described as face- and voice-selective, respectively. The authors then performed the reverse contrast - changes in activity associated with familiarity – but irrespective of modality. Familiarity with either face or voice was associated with an increase in activity in a single region of the posterior cingulate cortex, including the retrosplenial cortex – a region which has been implicated in episodic memory and emotional salience, which they suggest could act as a possible cortical locus for a supramodal PIN.

Joassin et al. (2004) extended neuroimaging work to electrophysiology, and performed an event-related potential (ERP) study aimed at examining the electrophysiological correlates of the cross-modal audiovisual interactions in an identification task, in the first attempt to directly define the neural correlates of person recognition from faces and voice combined. Participants either were presented with previously learned faces and voices together (AV condition), or faces and voices alone, and performed an identification task whilst ERPs were measured. The comparison of the responses evoked during the audiovisual condition as compared to the two unimodal conditions (AV – (A + V)) gave prominence to three separate cerebral activities: 1) a central positive/ posterior wave ~ 110 ms, explained by a pair of dipoles localised in the associative visual cortex; 2) a central negative/posterior positive wave ~ 170 ms, due to of a pair of dipoles localised in the associative auditory cortex and 3) a central positive wave ~ 270 ms, thought to reflect a network of cortical regions, including not only the FG and the associative auditory cortex, but also the superior

frontal gyrus (SFG) and SC – two proposed multimodal convergence regions. This study therefore provides direct evidence for cerebral processes at different latencies when combining face and voice identity information. In support to Ellis et al.'s (1997) proposition, they propose that the third central positive wave observed could correspond to a supramodal stage of integration, possibly reflecting the use of PINs. However, they also propose that the dipoles observed in the associative visual and auditory cortices could relate to initial integrative responses in the 'unimodal' sensory cortices.

Other studies have developed this work by aiming to test whether attribute specific modules (i.e., face and voice modules) can be directly and reciprocally connected, without the need for a supramodal node as an obligatory interface. Here, reciprocal interactions between different senses would be relayed through associative cortices, and would not necessarily require a feedback from a supramodal node. Such a direct integration of information would in theory prove advantageous for optimising person recognition under natural conditions (e.g., providing useful constraints to resolve ambiguity in noisy environments, or under less than optimal viewing or hearing conditions).

Von Kriegstein et al. (2005) used fMRI to test a cross-modal effect in the context of recognition of persons through voices – analogous to that they had previously shown in response to semantically meaningful stimuli in speech perception (Von Kriegstein et al., 2003) – and further investigated the underlying functional connectivity: that is, how involved brain regions were coupled to mediate any cross-modal effects. Participants were presented with sentences spoken by either personally familiar or unfamiliar speakers, and performed a recognition task, with the focus either being on the voice of the speaker, or the verbal content of the target sentence. The authors observed cross-modal responses to voices of familiar (but not unfamiliar) people in the FFA (see Figure 1.9) : however, this

response was only seen in the task that emphasised speaker recognition over recognition of verbal content. The functional connectivity analysis showed that the FFA was also coupled with the STS voice-selective region during familiar speaker recognition, but not with any other cortical regions usually active in person recognition. However, all these regions showed a familiarity-dependent correlation with the medial parietal cortex and the voice-responsive STS. Thus, this evidence suggests that the voice-responsive regions were functionally involved in two distinct interactions – one with the FFA and the other with an identity retrieval network.

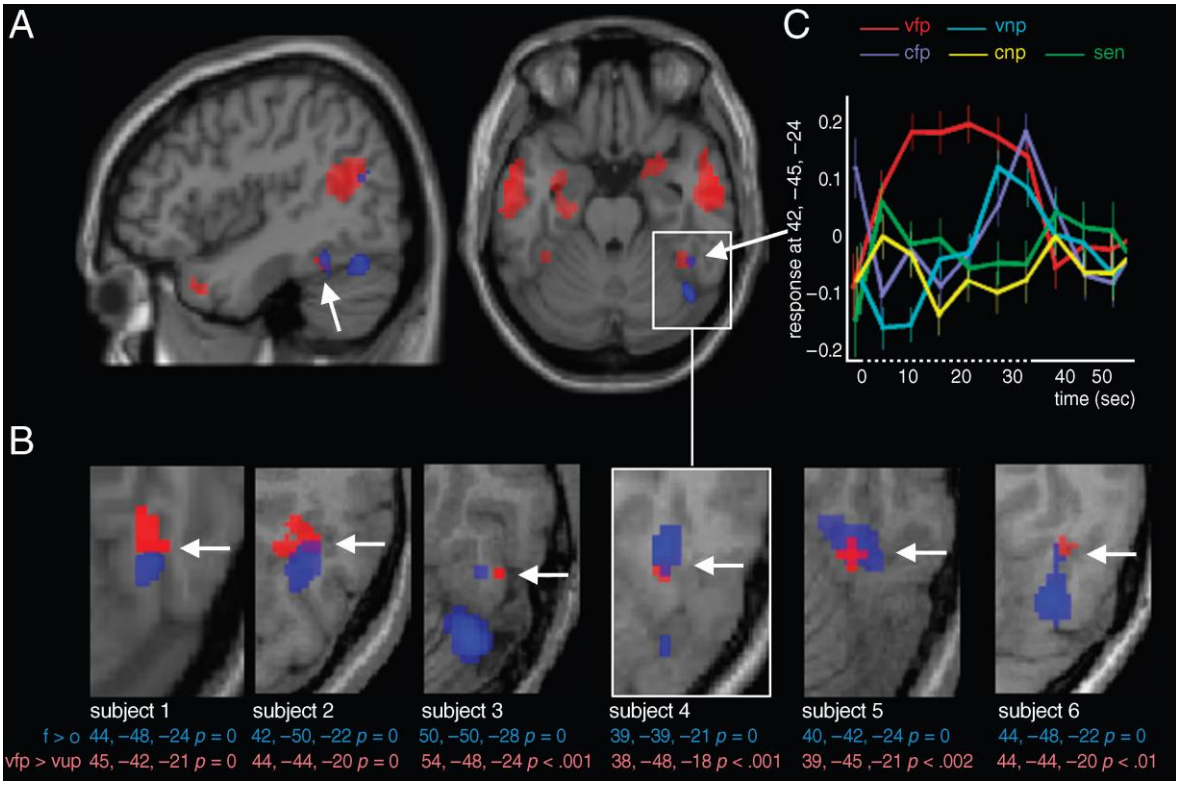


Figure 1.9. fMRI results from von Kriegstein et al. (2005). Activation of fusiform regions in the auditory experiment and the face area localizer study. (A) Group analysis. Contrast of familiar speaker versus non-familiar speaker (red), contrast of non-familiar faces versus objects (blue). (B) Single-subject analyses. Contrast of recognition of familiar speakers' voices versus non-familiar speakers' voices (red) and faces versus objects (blue), results from six of the subjects. (C) Time course of fMRI signal in the right fusiform region in response to the experimental conditions. Red = voice task (familiar); purple = verbal content task (familiar); cyan = voice task (non-familiar); yellow = verbal content task (non-familiar); green = noise task (speech envelope noises).

In a follow-up study, the same authors addressed whether, even under conditions of unimodal sensory input, crossmodal neural circuits that have been shaped by previous associative learning are activated, and could possibly underpin a performance benefit (Von Kriegstein and Giraud, 2006). Brain activity of participants was measured with fMRI either before, while or after they had learned to associate voices and faces, or other multimodal combinations (e.g. voices and written names, ring tones and cell phones). In the latter stage, participants performed a voice recognition task. After learning, participants were

better at recognising voices that had been paired with faces than those that had been paired with written names (measured by a significantly higher percentage of correctly recognised stimuli) and the association of voices with faces resulted in an increased functional coupling between face and voice areas – but only after face-voice learning, which confirmed its functional role in retrieving vocal identity. Furthermore, no such effects were observed for the other arbitrary multimodal combinations, which highlights that ecologically valid voice-face pairings induce specific multimodal associations. Overall, this study demonstrated that the optimisation of functional connectivity between cortical sensory modules specific to voices and faces entails a behavioural benefit for voice recognition by granting access to early distributed multisensory representations.

A recent study by Focker et al. (2011) investigated the time course of audiovisual interactions during person recognition using ERPs. In unimodal trials, two successive voices of the same or different speakers were presented. In the crossmodal condition, the first speaker consisted of the face of the same or a different person with respect to the following voice stimulus. Participants had to decide whether the voice probe was from an elderly or a young person. Reaction times to the second speaker were shorter when these stimuli were person-congruent, both in the uni- and crossmodal conditions. ERPs recorded to the person-incongruent as compared to the person-congruent trials were enhanced at both early (100-140 ms) and later processing stages (270-530 ms) in the crossmodal condition. A similar later negative ERP effect (270-530 ms) was found in the unimodal condition as well. These results suggest that not only is identity information conveyed by a face is capable to modulate the sensory processing of voice stimuli, but that it can do so at the level of perceptual encoding (<200 ms).

Yet more evidence for early interactions in audiovisual person recognition was provided by Blank et al. (2011). This study tested for evidence of direct structural connections between face- and voice-recognition areas by combining functional and diffusion magnetic tensor imaging (DTI). In individual participants, the authors localised three voice-sensitive areas in the anterior, middle and pSTS; and face-sensitive areas in the FFA. Using probabilistic tractography, they found evidence that the FFA was structurally connected with the voice-sensitive areas in the STS – particularly, to middle and anterior regions. Additionally, they provide evidence that the three different voice-sensitive regions within the STS were all connected with each other. Their results suggest that the assessment of person-specific information does not necessarily have to be mediated by supramodal cortical structures (like so-called modality-free PINs (Bruce and Young, 1986; Burton et al., 1990; Ellis et al., 1997), but could also result from direct cross-modal interactions between voice- and face-sensitive regions (von Kriegstein et al., 2005; von Kriegstein and Giraud, 2006). They propose that direct reciprocal interactions between auditory and visual sensory-processing steps serve to exchange predictive (i.e., constraining) information about the person's characteristics. These predictive signals could be used to constrain possible interpretation of unisensory, noisy, or ambiguous sensory input and thereby optimise recognition (Ernst and Banks, 2002).

Finally, in a recent fMRI study, Joassin et al. (2011) aimed to investigate the cerebral correlates of voice-face interactions in a recognition task. During the scanning session, three different conditions were presented: previously learned faces, voices, or face-voice associations, and participants categorised each trial according to its identity (i.e., its name). Behaviourally, voices were classified slower than both faces and face-voice associations (with no significant difference between the latter two conditions). Participants were also less accurate at classifying voices than both faces and face-voice associations (again, there

being no significant difference between the latter two conditions). With regards to neural activity, the authors found that voice-face associations (calculated using a subtraction method between bimodal and unimodal conditions) activated both unimodal auditory and visual areas, in addition to multimodal regions in the left angular gyrus and right hippocampus (Figure 1.10). Furthermore, a functional connectivity analysis confirmed the connectivity of the right hippocampus with both of the unimodal face- and voice-areas. These results appear to suggest that cross-modal person recognition relies on the activation of a distributed cerebral network, including both unimodal and other multimodal regions, which may contribute to processes such as cross-modal attention (the left angular gyrus) and audiovisual representations of people in memory (the hippocampus).

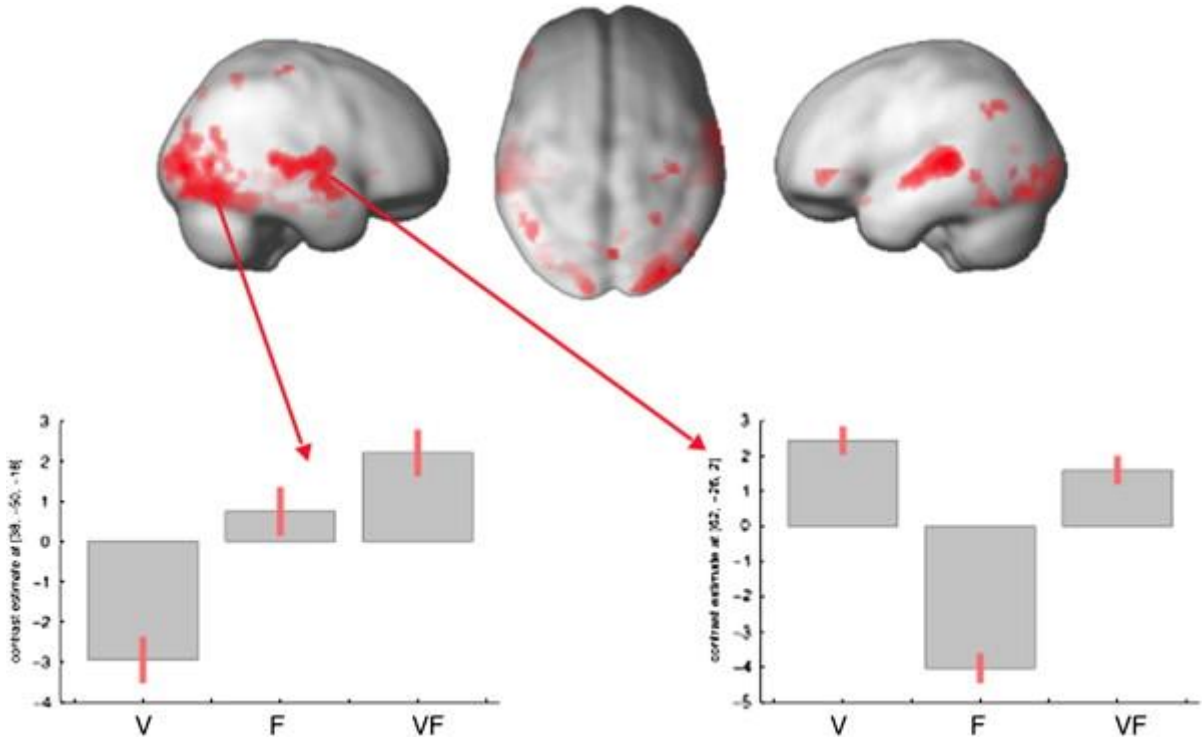


Figure 1.10. fMRI results from Joassin et al. (2011). Brain regions activated in the contrast [VF − (V + F)] (V = voices, F = faces, VF = face/voice associations). Top: Statistical parametric maps superimposed on MRI surface renders (left, top and right views); Bottom left: activation changes for each condition in the right middle fusiform gyrus; Bottom right: activation changes for each condition in the right MTG. Figure taken from Joassin et al. (2011)

Overall, findings from the aforementioned studies show that audiovisual information can crucially affect person recognition: congruent information across modalities can facilitate recognition of those familiar people around us, and bimodal identity information provokes unique neural responses. Results from neuroimaging studies also imply that conventional models of person recognition need to be modified to take a direct exchange of information between auditory and visual person-recognition areas into account. These results also integrate well with recent developments in multisensory research showing that information from different modalities interact earlier and on lower processing levels than traditionally thought (Cappe et al., 2010; Kayser et al., 2010; Klinge et al., 2010; Beer et al., 2011; reviewed in Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008). Indeed, direct connections between FFA and voice-sensitive cortices may be especially relevant in the context of person identification, in comparison to other aspects of face-to-face communication, such as speech or emotion recognition, where other connections might be more relevant. For example, speech recognition may benefit from the integration of fast-varying dynamic visual and auditory information (Sumby and Pollack, 1954) and therefore direct connections between visual movement areas and auditory cortices might be used (Ghazanfar et al., 2008; von Kriegstein et al., 2008; Arnal et al., 2009). Ultimately, recent results lend themselves to a more dynamic view of cross-modal interactions, in which heteromodal or multimodal regions are not simply engaged at a final stage of a hierarchical unimodal-to-multimodal processing model, but instead may work in parallel with unimodal processes, showing a reciprocal influence on one another.

Integration of face-voice gender information

The ability to match the gender of faces and voices is apparent from a young age. Walker-Andrews et al. (1991) investigated integration of gender information by showing infants aged 4 and 6 months videos of a male and a female face speaking side-by-side, paired with

a single soundtrack that corresponded to the gender of one visual display, but was synchronised with the articulations of both displays. They found that the ability to match dynamic faces and voices based on gender cues emerged around the 6 month period of infancy. In another study (Patterson and Werker, 1994), six separate experiments tested the sensitivity of young infants to vowel and gender information in dynamic faces and voices. Infants were presented with simultaneous displays of two faces articulating vowels. The heard voice matched the gender of one face in some of the conditions, and the vowel of one face in others. In the remaining conditions, vowel and gender were incongruent. Results showed that infants aged 4.5 months showed no evidence of matching face and voice on the basis of gender, but were able to match on the basis of the vowel. It was not until 8 months of age that the infants matched on the basis of gender. This suggests that gender matching is a later occurring ability than phonetic matching; however, both sets of studies demonstrate that infants combine faces and voices in gender recognition before 1 year of age. This early ability to integrate these sources of information is perhaps testament to its significance.

Despite its importance in social interaction and person identification, very few studies have investigated the integration of facial and vocal gender information in adults. In the first related behavioural study using adult participants, Smith et al. (2007) demonstrated auditory-visual integration in the perception of face gender by testing the perception of androgynous faces paired with pure tones where the fundamental-speaking frequency range was altered between male and female. Observers indicated the gender of each face when the face was accompanied by one specific tone (either low or high). When a face was accompanied by the other tone, observers performed a foil task of indicating the race (Asian or Caucasian) of the face. Observers were thus instructed to use the tones as task indicators. When an androgynous face was presented together with pure tones in the male

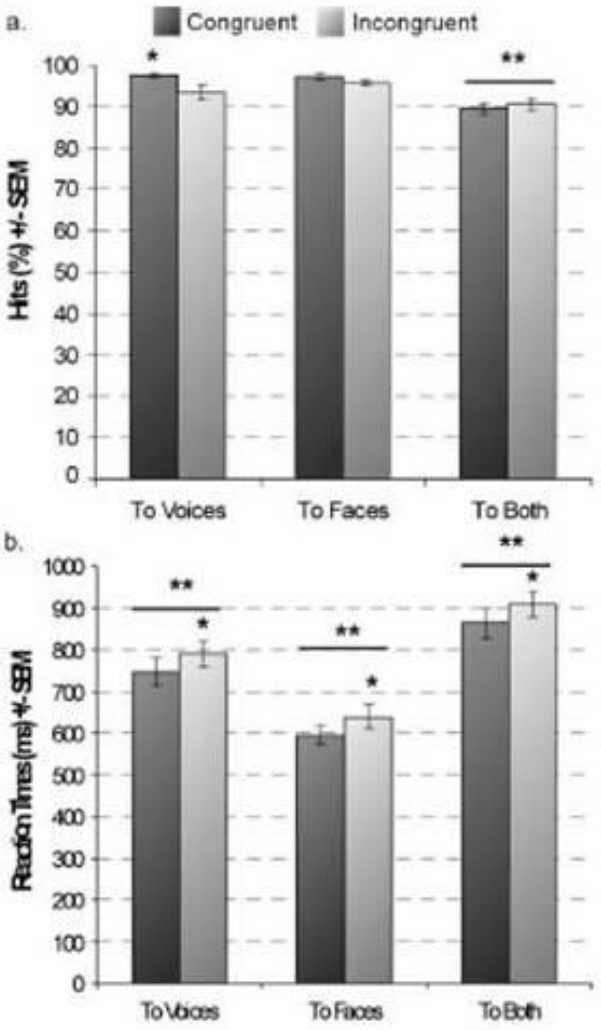
fundamental range, faces were more likely to be judged as male, whereas when faces were presented with pure tones in the female fundamental-speaking-frequency range, they were more likely to be judged as female.

Importantly, the authors were able to show that the crossmodal-integration effect they observed was primarily due to sensory integration as opposed to some manner of semantic interaction: in other words, the visual face processing was fundamentally dependent on the concurrent processing of gender-consistent auditory-frequency signals. As the authors note, pure tones do not sound like human vocalisation. By presenting pure tones (single-frequency tones) with frequencies in the male and female fundamental-speaking-frequency ranges, the authors were able to present gender-specific auditory information without the spectral components that allow conscious recognition of the signal as a human voice. Furthermore, the nature of the task meant that none of the participants were aware of any association between the tone frequency and face gender. This was confirmed by postexperiment interviews. Finally, there was a perceptual dissociation of the crossmodal effect of the tones (based on absolute frequency) from the explicit perception of the tones (based on relative frequency). The authors suggest that the overlap between the frequency tuning of the crossmodal-integration effect and the male and female fundamental-speaking-frequency ranges implies that the underlying auditory-visual integration develops because of concurrent neural processing of visual gender and gender-associated auditory frequencies. The fact this crossmodal effect on gender perception was achieved using relatively impoverished set of stimuli illustrates the strength of our ability to integrate this type of information.

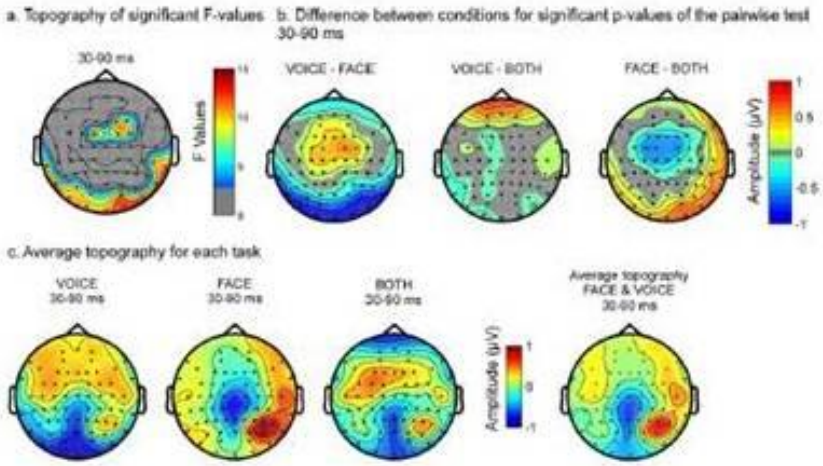
In another recent study, Latinus et al. (2010) investigated audiovisual and crossmodal interactions in gender categorisation whilst ERPs were recorded. Subjects performed three

gender judgement tasks: in the first, participants judged if the gender of the face and voice were congruent or not; and in the second and third, subjects categorised the bimodal stimuli by gender, in one case attending only to voices and in the other only to faces. The directed attention aspects of the task allowed the authors to determine the influence of top-down modulation on multimodal processing (i.e., effects due only to the task), whereas the use of congruent and incongruent stimuli provided information on bottom-up stimulus-dependent processing. They found that an incongruent face disrupted the processing of voice gender (indicated by significantly lower categorisation 'hits') while an incongruent voice had a lesser, non-significant on the perception of face gender, suggesting that in their experiment, vision dominated over audition in terms of overall gender categorisation. However, reaction times were longer for incongruent stimuli regardless of the direction of attention; thus, the unattended modality affected processing in the attended modality, revealing the automatic processing of bimodal information. The authors also showed that bottom-up processing of bimodal stimuli (congruency judgement) arose later (~190 ms) than when attention was directed to either one of the two modalities; the latter two tasks modulated early ERPs (~30-100 ms) over unisensory cortices, with respect to the attended modality. However, this influence on early ERPs depended on the preferential modality for the task, providing evidence for a visual bias in the case of face/voice gender categorisation. The fact that congruency judgement had a later influence seems to suggest that bottom-up multimodal interactions for gender processing are relatively late. Behavioural and electrophysiological results from this study are shown overleaf in Figure 1.11.

Behavioural measures



Attention modulated early brain activity (30-90ms)



Task and stimulus effects (150-250ms)

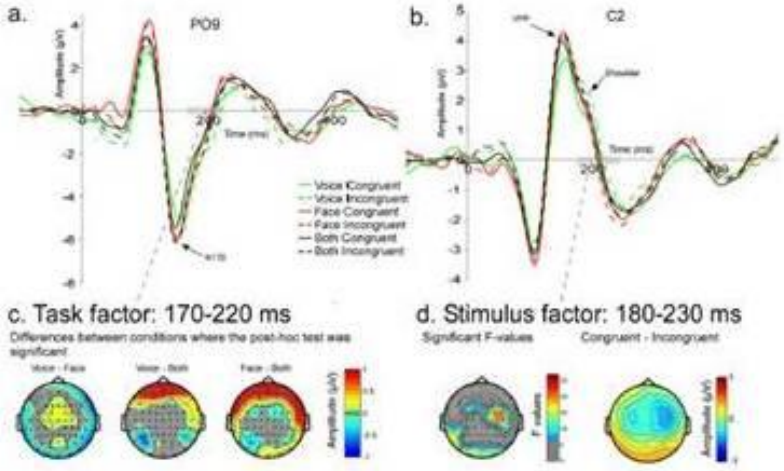


Figure 1.11 (previous page). Behavioural and electrophysiological results from Latinus et al. (2010). ***Left: Behavioural measures.*** **(a) Accuracy for the different tasks; (b) Reaction times. Congruent stimuli = dark grey; incongruent stimuli= light grey.**
Top right: Attention modulated early brain activity (30-90 ms). **(a) Topography of the average F-values in this time range. Non-significant F-values are in grey. (b) Topography of the absolute differences between the two tasks where the p-values of the post-hoc test were significant. Non-significant data are represented in grey. (c) Average topographic maps for each task between 30 and 90 ms. Left to right: VOICE, FACE, BOTH and the average between FACE and VOICE, shown as a comparison.**
Bottom right: Task and Stimulus effects between 150 and 250 ms. **N170 (a) at PO9 and VPP (b) at C2 for the 6 conditions. In green: VOICE task, in red: FACE task, in black: BOTH task. Solid lines: congruent stimuli; dashed lines: incongruent stimuli. Bottom: The maps represent the absolute differences between two conditions where post-hoc tests were significant. Non-significant data are represented in grey; c) Effects of task between 170 and 220 ms. d) Modulation of brain activity due to the stimuli between 180 ms**
and 230 ms for congruent and incongruent stimuli. Left map shows the significant F-values between 180 ms and 230 ms for the factor "stimulus" (non-significant F-values are represented in grey) and the right map shows the difference between topography to congruent and incongruent stimuli (scale: -1 1). Figure modified from Latinus et al. (2010).

The brain regions involved in audiovisual gender perception are still under question; however, one study (Joassin et al., 2011) has recently attempted to address this for the first time. Within their experiment, the authors asked participants to categorise faces, voices or both, according to their gender. There were four block designed acquisition runs, with each run compromising of six experimental blocks (with three conditions (A, V or AV) each repeated twice), interleaved with fixation periods. They found that the crossmodal processing of gender (calculated by the subtraction between the bimodal condition and the sum of the unimodal ones (i.e., super-additive criterion)) was associated with increased activation of several cortical and subcortical regions, including the unimodal face and voice areas and supramodal structures such as the striatum, the left superior parietal gyrus

and the right inferior frontal gyrus (Figure 1.12). Moreover, psychophysiological interaction analyses (PPI) revealed that both unimodal regions were inter-connected and connected to the prefrontal gyrus and the putamen, and that the left parietal gyrus had an enhanced connectivity with a parieto-premotor circuit, known to be involved in the crossmodal control of attention. This study provides early evidence that, similar to the brain's integration of face-voice identity information and indeed multisensory integration in general, the integration of face and voice gender is sustained by a network of cerebral regions including both early processing regions (i.e., visual and auditory cortices) and 'supramodal' regions. However, this evidence will need to be supported with results from other studies. Furthermore, the specificity of these regions of activation still needs to be clarified: for example, is this network activated independent of the task performed, or are there particular areas that integrate specifically gender information?

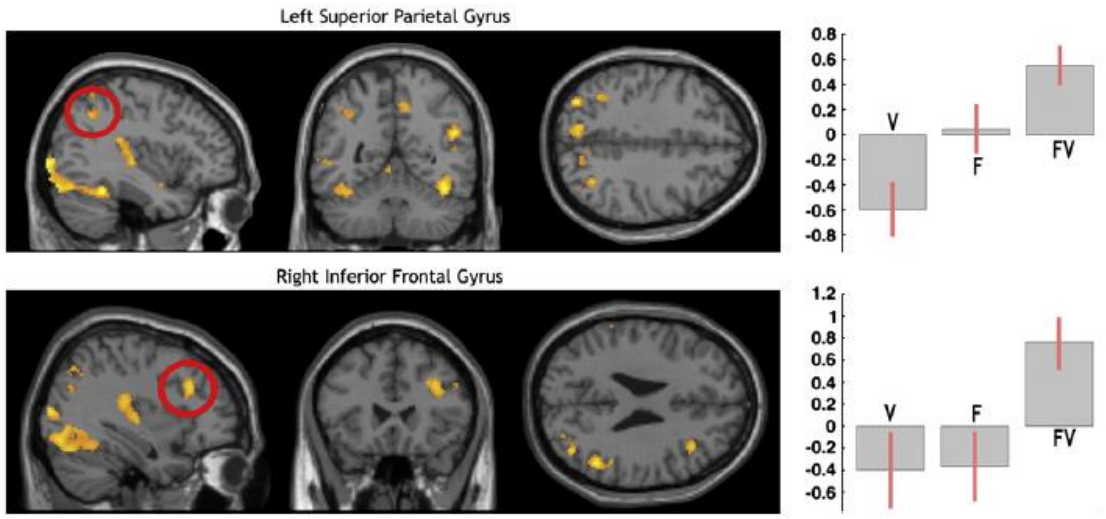


Figure 1.12. fMRI results from Joassin et al. (2011). Left side: brain sections of the contrast [FV−(V+F)] centred on the left superior parietal gyrus (top) and the right inferior frontal gyrus (bottom). Right side: activation changes for each condition in the left superior parietal gyrus (top) and the right inferior frontal gyrus (bottom). V=voices, F=faces, FV=face/voice associations. Figure taken from Joassin et al. (2011)

These pioneering studies have formed the foundation for further study in this area: with such little work conducted in this field the scope for related work is large. **Chapter 4** of this thesis attempts to add to the understanding of face-voice gender integration, with a behavioural experiment that is detailed further in the thesis rationale.

Integration of face-voice affective information

Early research on face-voice emotion perception concentrated on commonalities between the two modes, looking for common processing resources and overlapping brain structures for face and voice expressions (e.g. Royet et al., 2000). For example, researchers looked to brain damaged patients in order to examine whether a deficit in the perception of facial expression had a parallel in impairment of vocal expression, and whether a deficit in one mode could exist without a deficit in the other mode (see van Lancker, 1997). In one instance, Scott et al. (1997) reported that an amygdalectomy patient who was impaired in recognition of facial expression also showed a deficit in processing of vocal emotion. However, as noted by Pourtois et al. (2005), how the brain combines multiple sources of information is not something that can be elucidated by simply juxtaposing results obtained in studies that have investigated facial and vocal emotion processing separately. Indeed, the issue of multisensory or audiovisual integration is far more complex that that concerning common processing resources, or so-called 'amodal' representations.

As in the case of gender recognition, research shows that we combine facial and vocal affective information from a young age. An early series of behavioural experiments was conducted to examine infants' recognition of emotional expressions (Soken and Pick, 1992; Walker, 1982; Walker-Andrews, 1986), in which infants had to detect the correspondence between emotional information provided by the face and voice. In these experiments infants were presented simultaneously with two different dynamic facial

expressions accompanied by a single vocal expression that affectively matched one of the facial displays. Walker (1982) showed that infants looked longer at the facial display that affectively matched the voice. In another experiment (Walker-Andrews, 1986), the mouth was occluded so that synchrony between lip movements and vocal expressions could not account for infants' differential looking behaviour. The results revealed that 7 month olds, but not 5 month olds, looked longer at the facial display that was congruent to the vocal affect. These behavioural findings suggest that 7 month old infants can detect common affect across audiovisual expressions of emotion, and can do so even in the absence of temporal synchrony between face and voice.

Regarding adult subjects, the most frequently used approach to investigate both psychologyical and physiological aspects of audiovisual emotion perception has been to use bimodal perception paradigms in order to demonstrate that the response to affective information expressed in the face *and* voice differs to that when it is expressed in only one of these modalities. Behaviourally, this is commonly indicated by a facilitated or impaired categorisation judgement for congruent and incongruent information, respectively; or simply that the percept of emotion is somehow altered when different types of facial and vocal affective information is combined within an audiovisual stimulus. With regards to cerebral activity, researchers seek to demonstrate that an audiovisual stimulus provokes a unique pattern of activity. Although a number of studies have addressed the merging of emotional information using a variety of methodological approaches and experimental paradigms, the behavioural and neuroimaging results seem to be relatively consistent, as will be detailed further in this section.

Arguably, Massaro and Egan (1996) sparked the growth of research in this field. These authors presented their participants with a single word, spoken in one of three tones

(neutral, happy, angry), and showed them a computer generated face showing one of the three expressions. The task of the participants was to classify the emotion as either happy or angry. The authors found that the frequency of either response was dependent on the emotions expressed in both the face and the voice.

In a further set of pioneering experiments, de Gelder and Vroomen (2000) also showed that identification of emotion in the face can be biased towards that in a simultaneously presented voice. The study extended the work by Massaro and Egan (1996) by using photographs taken from a *morphed* continuum, extending between sadness and happiness; and using a whole sentence instead of a monosyllabic word. In this study, participants were presented with either faces alone, voices alone, or faces and voices together and were asked to indicate whether they thought the presented individual was happy or sad. The authors found that when presented with a face and a voice expression, participants appeared to combine both sources of information, with categorisation of each face (apart from those congruent) shifted in the direction of the simultaneously presented voice (see Figure 1.13 below).

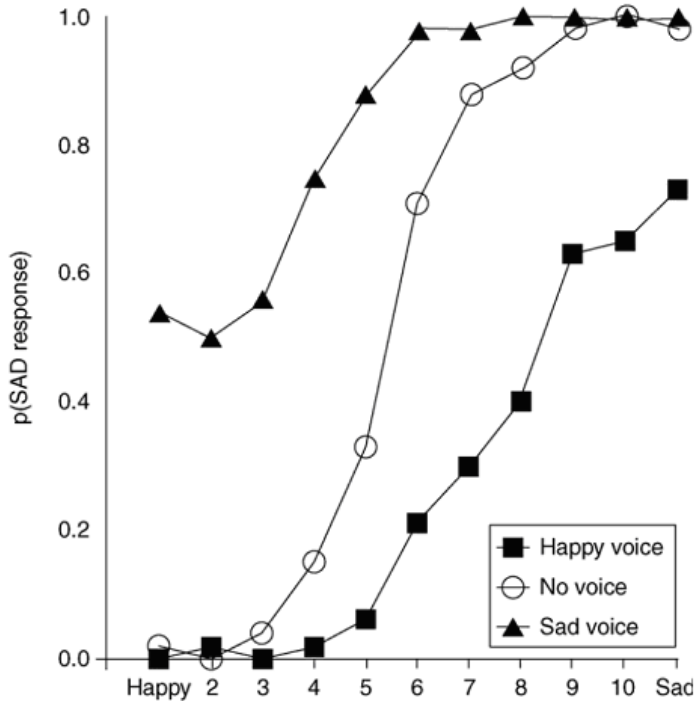


Figure 1.13. Behavioural results from de Gelder and Vroomen (2000). Percentages of 'sad' responses as a function of the face continuum when combined with the happy, sad and no voice. Figure taken from de Gelder and Vroomen (2000).

Ethofer et al. (2006) used a similar perception paradigm to de Gelder and colleagues, presenting participants with either faces alone, voices alone or faces and voices together after which participants gave an emotional valence rating for each stimulus. Visual stimuli were from a morphed continuum of still photographs, ranging from either happiness to neutral, or neutral to fear. Auditory stimuli were sentences spoken in either happy or fearful prosody. The authors found that the participants rated fearful and neutral facial expressions as being more fearful when presented in the presence of a fearfully spoken sentence as compared to that in the no voice condition. However, the presence of a sentence spoken in a happy tone did not significantly alter valence ratings. These results differ somewhat to that of de Gelder and Vroomen (2000), whose interaction effects were attributable to incongruity of audiovisual emotion information of *both* positive and negative valence. The authors propose that effect of only fearful prosody was perhaps due

to the higher biological relevance of this emotion, as compared to more positive emotions such as happiness, which although of high value in social situations are of less immediate survival value. These results already suggest some differences in integrative mechanisms within the category of affect recognition, with the nature of integration somewhat dependent on emotion category.

In addition to altered categorisation of emotion, researchers have observed *faster* categorisation of emotion in bimodal, as opposed to unimodal conditions (e.g. Giard and Peronnet, 1999; de Gelder and Vroomen, 2000; Massaro and Egan, 1996; Ethofer et al., 2006; see Figure 1.14 overleaf for an example)). This result can be linked to the 'redundant target effect' (RTE), which states that people typically respond faster to double targets (two targets presented simultaneously) than to either of the targets presented alone. The difference in latency is termed the redundancy gain (RG). There are two approaches to explaining the RTE. The simpler approach models are called race models (Raab, 1962), which propose that the signals on different channels (from different senses) cause separate activations and the response is caused by the first one of these processes to finish. Hence the response to a redundant signal is fast because it is produced by the faster of the two separate signals. Coactivation models, on the other hand, allow activation from different channels to be combined somewhere in the processing system. According to coactivation models, the response to a redundant signal is fast because two signals interact to initiate a response (Miller, 1982, 1986). However, in order to see a gain of audiovisual presentation, the two signals need to be congruent: as can be observed in Figure 1.14, when these two signals are incongruent it leads to a lengthening of reaction times, both compared to unimodal and congruent audiovisual information,

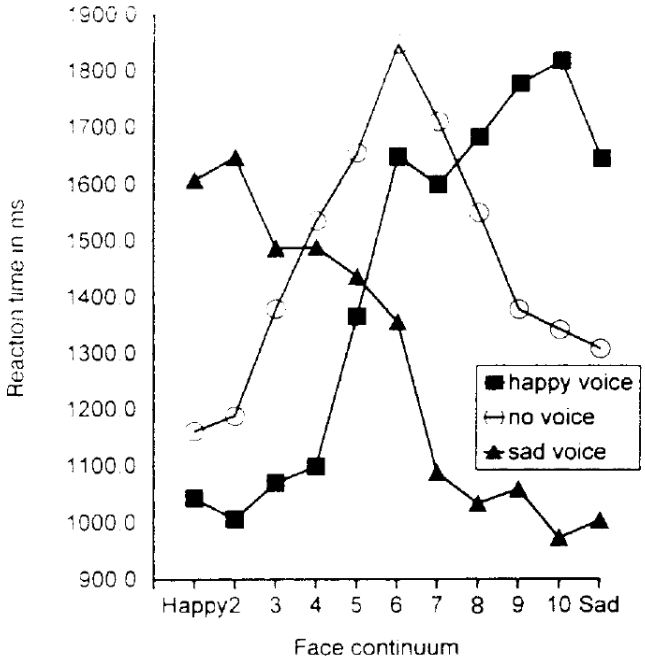


Figure 1.14. Behavioural results from de Gelder and Vroomen (2000). Mean reaction times of the identification responses as a function of the face continuum when combined with the happy, sad and no voice. Figure taken from de Gelder and Vroomen (2000).

Collignon et al. (2008) combined both the speed and accuracy of emotion categorisation to test the 'inverse effectiveness' principle – which, as previously described, states that the result of multisensory integration is inversely proportional to the effectiveness of the relevant stimuli (Stein and Meredith, 1993) - by presenting stimuli with the addition of noise in one sensory channel (i.e., vision), in order to decrease the reliability of the sensory information presenting. Stimuli with no noise were also included as a control. Participants were required to discriminate between 'fear' and 'disgust' affective expressions presented auditorily, visually, or audiovisually, in a congruent or incongruent way. The authors observed improved performance (calculated using 'inverse efficiency scores', which takes into account both speed and accuracy – 'corrected reaction times') in the congruent condition, compared to any unimodal condition. This effect was greatest in the noisy condition, as would be predicted by the inverse effectiveness principle (Figure 1.15).

Furthermore, when incongruent pairs were presented in the noiseless condition, participants orientated their responses more often towards the visual modality. However, when they were presented with audiovisual stimuli composed of noisy visual stimuli, the participants categorised more often the affect expressed in the auditory modality. These results suggest that visual dominance in affect perception follows flexible, situation-dependent rules, as opposed to a more rigid manner, that allow information to be combined with maximal efficacy (Ernst and Bulthoff, 2004).

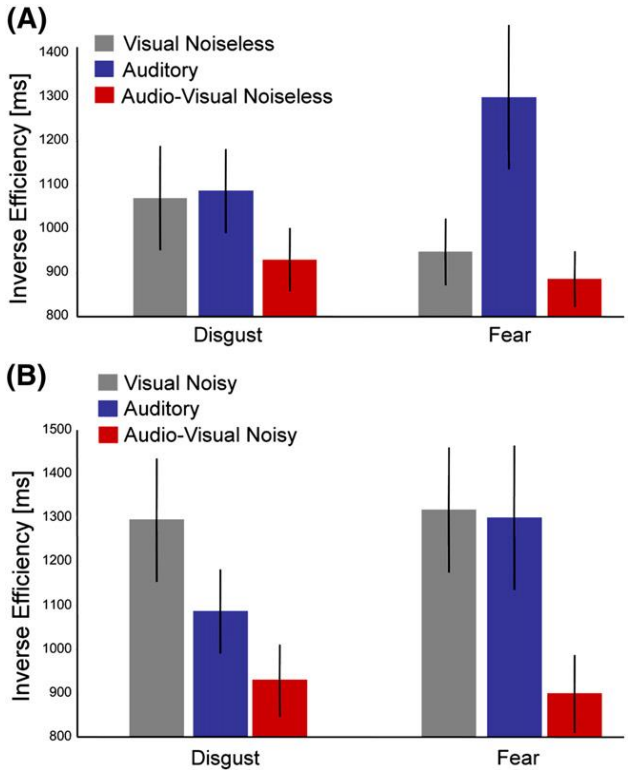


Figure 1.15. Behavioural results from Collignon et al. (2008). Mean IE scores and standard errors obtained for unimodal stimuli and congruent audio-visual stimuli for both emotion expressions. The figure displays the results obtained with noiseless visual stimuli (panel A) and for noisy visual stimuli (panel B). Figure taken from Collignon et al. (2008).

Regarding attention, de Gelder and Vroomen (2000) propose that affective information integration emerges in a mandatory fashion, without the necessity to attend to the

respective unimodal stimuli. In a second and third experiment within their study, they complemented their initial results by using a 'top-down' approach. Participants were explicitly instructed to attend to only one of two modalities, and to ignore the information in the other modality even although it was still being presented. They found that even when participants were instructed to ignore information in a particular modality, and only take account of the information presented in the other domain, responses still indicated an unconscious integration of the bimodal inputs, with significant categorisation shifts in both experiments.

In their study Collignon et al. (2008) also supplemented their results described above by explicitly requesting the participants to attend to only one sensory modality at a time, while completely disregarding the irrelevant sensory modality. Results clearly demonstrated a performance increase when the non-target modality was congruent and a decrease when it was incongruent, attesting to the automaticity of multisensory interactions in the perception of emotion expressions. The authors propose this situation could be related to an 'emotional stroop'. The influence of the irrelevant modality was especially strong when delivered with noisy sensory targets, again in accord with the inverse effectiveness theory – when the attended modality was less reliable, participants automatically attributed more weight to the irrelevant sensory modality in their processing of bimodal emotional expressions.

Vroomen et al. (2001) used manipulation of perceptual load to show that the integration effect they observed was unconstrained by the allocation of attentional resources. Participants judged whether a voice expressed happiness or fear, whilst trying to ignore a concurrently presented static facial expression. As an additional task, the subjects had to add two numbers together rapidly, count the occurrences of a target digit in a rapid serial

visual presentation, or judge the pitch of a tone as high or low. The visible face had an impact on judgments of the emotion of the heard voice in all the experiments, showing that integration was independent of attention. This suggests that integration of visual and auditory information about emotions may be a mandatory process.

Altogether, the findings of de Gelder and Vroomen (2000), Collignon et al. (2008) and Vroomen et al. (2001) support the automatic nature of the processes which serve to integrate facial and vocal affective information. Furthermore, evidence which suggests that emotional stimuli attract attention even if they are task-irrelevant (Mack and Rock, 1998) (an 'attentional capture effect') also attests to the highly autonomous nature of emotional evaluation. This would infer that affective information in the face and voice might combine at an early stage of processing, and results from a range of neuroanatomical, electrophysiological and neuroimaging studies – described in further detail below – substantiate this argument.

Recordings of electric brain responses across the human scalp (ERPs) have been used to investigate the time course of crossmodal binding. In a pioneering EEG study, de Gelder et al. (1999) presented congruent and incongruent emotional faces and prosody. They found that a facial expression paired with an incongruent affective voice provoked a mismatch negativity response around 180ms after presentation – even though the participants were explicity instructed to ignore the auditory stimulus - indicating not only that auditory processing is modulated by concurrent visual information, but this integration happens even when the participant does not pay attention to one of the modalities.

In a similar vein, Pourtois et al. (2000) provided additional support for fast emotion-specific neural patterns in sensory cortices. These authors demonstrated that the auditory

N1 component - which occurs around 110 ms after the presentation of an affective voice – was significantly enhanced by an emotionally congruent (upright) facial expression, but not by incongruent or inverted congruent faces. In a second ERP study, Pourtois et al. (2002) used happy and fearful stimuli, and found that emotionally congruent and incongruent face–voice pairs elicited a positive ERP component (named the 'P2b component') which peaked earlier in congruent than in incongruent pairs; incongruent emotions in the face and voice delayed an auditory deflection around 220-260 ms post stimulus. The source generating this effect was localised in the anterior cingulate cortex, which has previously been implicated in error monitoring. Overall, the perceptual integration of incongruent audiovisual stimuli seemed to be deaccelerated compared to congruent ones, indicating a higher neural processing effect.

Finally, Jessen and Kotz (2011) presented participants with complex audiovisual emotion displays including voices, faces and bodies. They observed an amplitude reduction of the early auditory N1 component, followed by an enlarged P2 potential in udiovisual compared to unimodal conditions. Furthermore, they also showed an emotion effect on the auditory N1, expressed by a shorter latency for fearful than for neutral audiovisual sets. Fearful displays also induced larger late positive components than all other emotional conditions. These results indicate some preference for the neural representation of emotionally relevant stimuli during early auditory processing stages. Together, results from the above electrophysiological studies suggest that congruent audio-visual emotional information enhances sensory-specific processing, and that incongruent emotional information delays the timing of the ongoing processes. Furthermore, these studies suggest that multisensory integration can occur at an early stage of processing (i.e. 110-250ms post-stimulus), at the perceptual rather than the later decisional stages.

ERP studies have provided valuable information on the timing of affective integrative processes. However, such a technique does not allow inference as to which brain regions are involved when combining an affective face with a voice. This has been typically been achieved using neuroimaging techniques (mostly fMRI). Although studies in this field are limited (in comparison to those investigating the unimodal processing of emotional information), a number have identified a network of different regions responding to the audiovisual presentation of affective information, including the well-documented STS, thalamus and affective processing structures such as the amygdala, in addition to the early face- and voice-selective regions. The involvement of the latter regions in particular would further support integration at the perceptual stages of stimulus processing, in turn strengthening the argument that affective face-voice integration could potentially be automatic or independent of attention.

In one of the first audiovisual studies in this area using fMRI, participants were scanned whilst categorising a static facial expression as either 'fear' or 'happiness', and ignoring a concurrently presented emotional voice (Dolan et al., 2001). The aim of the authors was to define the neuronal mechanisms for a perceptual bias in processing simultaneously presented emotional voices and faces; specifically, whether and how the bimodal presentation of an affective voice could facilitate the recognition of that emotion expressed in the face. The authors found that perceptual facilitation during face fear processing was expressed through modulation of neuronal responses in the amygdala and the fusiform cortex. There was an enhanced response in the left amygdala to congruent fearful faces (fearful voice + fearful face) compared with incongruent (happy voice + fearful face). Additionally, in the fear- congruent condition effects of crossmodal affective integration were also observed at earlier, unimodal levels of face processing, with stronger activation of the right FFA during judgement of facial expressions. The modulation they observed

was context-specific in that it was expressed *exclusively* during presentation of congruent fearful face–voice combinations, not congruent happy face-voice expressions. The authors take this data to suggest that the amygdala is important for emotional crossmodal sensory convergence, specifically during fear processing, with this convergence being mediated by task-related modulation of face-processing regions of the fusiform cortex.

Pourtois et al. (2005) used PET to investigate the brain regions that were activated during the perception of happiness and fear, in the face, voice and combined audiovisual pairs. Their work extended upon Dolan et al.'s (2001) study by including not only bimodal conditions, but also single modality conditions in order to investigate the difference between each single modality separately and their combination. In addition, they used an indirect processing task (a gender decision task) in which participants were not consciously attending to the emotional meaning of the stimuli. Their analysis highlighted the left MTG as a region activated more by audiovisual pairs as compared to unimodal stimuli. Their results also revealed convergence areas in the left hemisphere for happy face-voice pairings, and in the right for fear face-voice pairings, indicating that there might exist separate neuro-anatomical substrates for integration of positive and negative emotions. The results also confirmed the involvement of the bilateral pSTS regions in affective integration by showing that their cerebral activity was linearly related to the behavioural gain in classification accuracy (for the bimodal versus unimodal condition). Moreover, functional connectivity between audiovisual integration areas and associative auditory and visual areas was increased during the bimodal condition. This evidence suggests that the multisensory perception of emotion from the face and voice converges in heteromodal regions of the brain, but can also interact at an earlier processing stage.

Later functional imaging studies have particularly emphasised the integrative role of the STG/MTG as well as the pSTS. The pSTG and pSTS are well-known areas supporting the multisensory integration of audiovisual stimuli. As mentioned earlier in this introduction, studies in non-human primates suggest that the pSTS is involved in forming multisensory representations of observed actions (Barraclough et al., 2005). Furthermore, multisensory areas of the pSTS are situated at the interface of auditory and visual association cortices, and they receive multiple converging projections from the respective primary sensory areas (Seltzer and Pandya, 1978); thus, they are well suited to subserve the combination of facial and vocal stimuli.

For example, in an fMRI study, Kreifelts et al. (2007) found that audiovisual presentation of non-verbal emotional information resulted in a significant increase in correctly classified stimuli when compared with visual and auditory stimulation, a gain which was paralleled by enhanced activation in the bilateral pSTG and right thalamus, when contrasting audiovisual to auditory and visual conditions (with the exception of happiness, although there was an enhanced sensitivity of the integration sites to stimuli with emotional non-verbal content (for all emotion categories) as compared to neutral stimuli). Furthermore, a characteristic of these brain regions was a linear relationship between the gain in classification accuracy and the strength of the BOLD response during the bimodal condition (for an illustration of these results, refer to Figure 1.16). Finally, enhanced effective connectivity between audiovisual integration areas and associative auditory and visual cortices was observed during audiovisual stimulation, offering further insight into the neural process accomplishing multimodal integration. However, it should be noted that in this study super-additive integration emerged for neutral stimuli as well. Thus, it must be considered that increased responses to emotional stimuli could have just been an unspecific

influence of attention due to their higher salience, rather than an attribute specific to the supramodal representation of emotional information.

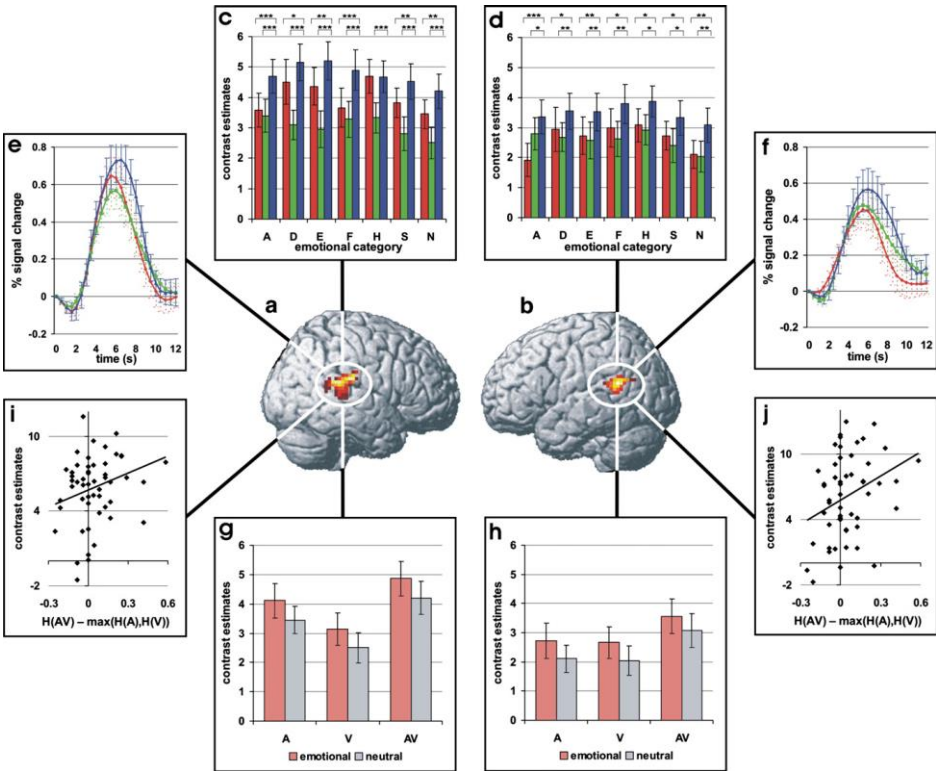


Figure 1.16. Behavioural and fMRI results from Kreifelts et al. (2007). Face–voice integration of dynamic affective information: cerebral effects. Increased activation during audiovisual (AV) stimulation compared with either auditory (A) or visual (V) stimulation within (a) right and (b) left posterior STG (pSTG). (c,d) Contrast estimates for auditory (red), visual (green) and audiovisual (blue) stimulation show a significant integration effect within bilateral pSTG for all emotional categories with the exception of happiness in right pSTG. Asterisks mark significant differences. Event-related responses for A (red), V (green) and AV (blue) stimulation show a stronger and slightly prolonged activation for bimodal stimulation in (e) right and (f) left pSTG. (g,h) Both regions exhibit stronger responses to emotional than to neutral stimuli under every experimental condition (A, V, AV). A positive correlation between contrast estimates during the AV condition and behavioural gain, estimated as the difference between classification hit rate during the bimodal condition and the maximum of hit rates during the unimodal conditions, was significant over subjects in (i) left pSTG and showed a tendency versus significance in (j) right pSTG. Figure taken from Kreifelts et al. (2007).

Kreifelts et al. (2009) developed this work by examining exclusively the role of the STS in audiovisual integration of non-verbal emotion signals. Participants were run in three separate experiments: two of these experiments tested sensitivity to faces and voices (i.e., 'localiser scans'), whilst another tested for sites that showed an increased response to audiovisual affective information, as compared to both affective faces and voices alone. Regardless of the individuals' spatial variability of the STS, the authors demonstrated that all three functional characteristics under investigation were represented in the STS: maximum voice sensitivity was located in the mid STS and maximum face sensitivity in the pSTS (specifically, the posterior terminal ascending branch), and audiovisual integration of affective signals peaked in the anterior pSTS, at an overlap of face- and voice-sensitive regions. Thus, these results suggest some manner of functional subdivision of the STS into modules subserving the processing of different aspects of social communication, including integration of affective information, and furthermore imply a possible interaction of the underlying voice- and face-sensitive neuronal populations during the formation of the audiovisual percept.

Finally, the same group extended this work in 2010, by investigating how trait emotional intelligence (EI) – a behavioural measure – was associated with audiovisual emotional fMRI activation patterns. A general conjunction analysis of audiovisual emotional integration revealed integration in not only the bilateral pSTS and thalamus, but also in regions with well known sensitivity for social signals carried in emotional expressions, such as the amygdala and FG. However, trait EI was linked only to haemodynamic responses in the right pSTS - the area observed in their earlier study; and furthermore, it was only this region which exhibited both a face and voice selective response. The authors suggest that this combined sensitivity to these 'social information' sources may be an essential characteristic of the neural structures subserving the audiovisual integration of

human social communicative signals. Within all other regions shown to subserve integration of affective information in this experiment (.e. thalamus, amygdale), no linked responses to behaviour were seen, a reason for which the authors do not speculate on, although they suggest factors such as emotional salience may play a role. Overall, the findings of this study point to the right pSTS as playing a unique and pivotal role in the processing of human social signals. Furthermore, the specific correlation with a behavioural measure of emotion intelligence in this region provides good evidence to refute the claim that increased activation to audiovisual emotion in the pSTS is just an unspecific influence of attention.

Recently, MEG has become a more widely neuroimaging technique. MEG offers a very direct indication of neural electrical activity, measuring the magnetic fields as opposed to the signal dependent on the blood oxygen level, and thus has high temporal resolution, unlike fMRI. Additionally, it measures the magnetic fields produced by neural activity, which are likely to be less distorted by surrounding tissue (particularly the skull and scalp) compared to the electric fields measured by EEG (electroencephalography). In the first study to use MEG to study affective integrative processes, Hagan et al. (2009) examined the role of the STS in audiovisual emotion perception. They measured the neural responses of participants as they viewed and heard fearful voices and static faces (audio only, visual only and audiovisual). Static faces were used to minimise responses in the pSTS, which is known to already play a role in processing of transient facial changes made during emotional expression. Additionally, the authors presented neutral faces with minimally congruent neutral nonverbal vocal signals (i.e., polite coughs) to determine the extent to which integration mechanisms are engaged for facial–vocal pairings irrespective of congruence, because it has been suggested that facial and vocal integration is a mandatory process. The authors observed a significant super-additive response in the right pSTS

within the first 250ms for emotionally congruent AV stimuli. Furthermore, these authors also compared the time course of the fear super-additive response with the time courses of the responses observed during the individual unimodal conditions. The response to the unimodal auditory stimulus occurred within the first 150 ms, whereas the response to the unimodal visual stimulus occurred within the first 300ms. The authors propose that the super-additive response in the pSTS could arise through interactions with the auditory cortex, a result which has been previously shown in monkeys (Ghazanfar et al., 2008).

Contributions of the pSTS region to multisensory emotional face-voice pairs were also described by Robins et al. (2009). Participants in the present studies viewed short (dynamic) movies blocked by modality (audio, video, audio-video) and/or emotion (angry, fearful, happy, neutral), as well as unimodally presented facial and auditory emotional cues while undergoing fMRI scanning. Activation or enhancement of activation to the AV emotional stimuli was contrasted with activation during unimodal conditions; additionally, specific effects of emotion were also investigated. In this study, the perception of bimodal emotional stimuli increased activation in the bilateral STS/STG relative to unimodal emotional conditions. Interestingly however, the effects of emotion were distinct from the effects of AV integration: in addition to the pSTS, effects of emotion were consistently demonstrated in the anterior STG (aSTG) bilaterally. The authors suggest that a role for the aSTG in emotion perception makes sense, since rostral regions of the STG have projections to multiple nuclei in the amygdala. Overall, these findings support the role of the pSTS in integration of affective signals, but alsoprovide evidence that areas of the STG which are traditionally considered parts of the auditory cortex can also be modulated in a multimodal fashion.

However, two recent studies have somewhat challenged the specific role of the pSTS in audiovisual affective processing. In the first, Muller et al. (2011) investigated incongruence effects in crossmodal emotional integration. Behavioural data confirmed an audiovisual integration effect: subjects rated fearful and neutral faces as being more fearful when accompanied by screams as compared to yawns. Additionally, the imaging data revealed that incongruence of emotional valence between faces and sounds led to increased activation in the middle cingulate cortex, right superior frontal cortex, right supplementary motor area as well as the right temporoparietal junction; many regions which have been previously implicated in cognitive conflict and attentional control. However, there was no effect of (in-) congruency in the pSTS, as might have been expected. Neither did they find an effect of congruency in the amygdala, as previously reported by Dolan et al. (2001). However, the authors suggest that a neutral stimulus in one sensory channel (as was the case in this experiment) could have potentially counteracted the emotional saliency provided by the stimulus in the other modality by attenuating amygdala responses, as demonstrated by a significant effect in the left amygdala when incongruence effects with respect to the presence and absence of emotional stimuli was tested.

In one of the most recent studies investigating emotion integration, Klasen et al. (2011) attempted to determine whether perceptual integration of facial and vocal emotions takes place in primary sensory areas, multimodal cortices, or in affective structures. They combined emotional faces and voices in congruent and incongruent ways and assessed functional brain data during an emotional classification task. Both congruent and incongruent audiovisual stimuli evoked larger responses in thalamus and superior temporal regions compared with unimodal conditions. However, whilst incongruent emotions (compared to congruent) activated a frontoparietal network and bilateral caudate nucleus, congruent emotions (in the reverse contrast) were characterised by activation in amygdala,

insula, ventral posterior cingulate (vPCC), temporo-occipital, and auditory cortices. Notably, the STS and thalamus were absent in this contrast. The authors point to the fact that audiovisual integration studies have reported enhanced activation in superior temporal regions and thalamus to emotionally neutral audiovisual stimuli, suggesting a more general role of these structures in multimodal perception. They suggest that because the thalamus and superior temporal cortex respond similar to emotionally congruent and to incongruent bimodal stimuli, activity in these areas may not necessarily reflect a semantic integration of bimodal emotions because the latter can only be expected in congruent conditions. Shared with the putative AV network, happy emotions yielded higher activity in the left amygdala, but only the vPCC responded to congruent facial and vocal expressions in all three emotion categories compared with incongruent stimuli.

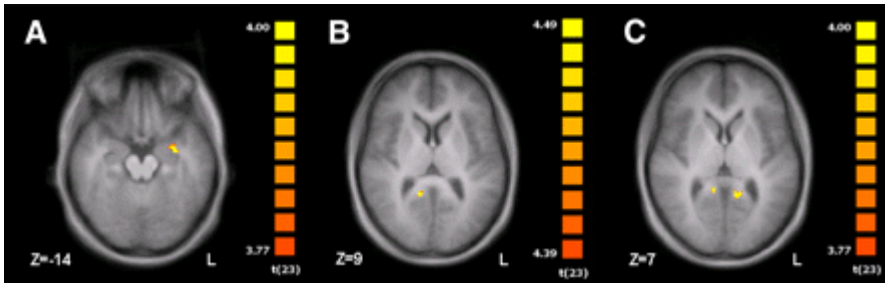


Figure 1.17. fMRI results from Klasen et al. (2011). Conjunction analyses on multimodal integration. The left amygdala yielded stronger responses to bimodal compared with unimodal and stronger responses to congruent compared with incongruent trials (*A*). Right (*B*) and left (*C*) vPCC integrated affective facial and vocal emotion independent from emotional category as confirmed by the conjunction (Congruent Neutral > Incongruent) ∩ (Congruent Angry > Incongruent) ∩ (Congruent Happy > Incongruent). Figure taken from Klasen et al. (2011).

The proposed role of the amygdala as a potential convergence region for audiovisual emotion information has been supported by studies previously described in this section (i.e., Kriefelts et al., 2010; Dolan et al., 2001; Pourtois et al., 2005) and is well in line with

the relevance of this structure for emotion processing (e.g. Fusar-Poli et al., 2009). The amygdala's part in *implicit* emotional integration was emphasised also by Ethofer et al. (2006) who reported a positive correlation of amygdala activity and the influence of fearful prosody on a facial emotion judgment task. It appears that the amygdala may process affectively relevant information without awareness (Vuilleumier et al., 2001; Dolan and Vuilleumier, 2003), and indeed, the proposal that the amygdala may be involved in conscious as well as unconscious emotion processing was supported by results from Ethofer et al. (2006) who showed that unattended fearful prosody enhanced functional connectivity between amygdala and FFA, further indicating an attentional modification in face-encoding areas by prosodic cues.

However, the posterior cingulate is a far less documented region with regards to audiovisual emotion processing. The vPCC is involved in the processing of self-relevant emotional and non-emotional information as well as in self-reflection (Vogt et al., 2006). Klasen et al. (2011) propose that via reciprocal connections of the vPCC with the ACC, the emotional information can gain access to the cingulate emotion sub-regions, helping to establish the personal relevance of sensory information coming into the cingulate gyrus. This would make it a suitable candidate for supramodal representation of emotion information from different modalities independent from low-level sensory features.

Taking the results from all these studies into account, behavioural work shows that hearing and seeing emotional expressions can support and influence each other in a similar way to that of gender and identity perception, a notion which is supported by investigations on the underlying neurobiology. Behavioural advantages arising from multimodal perception are paralleled by specific integration patterns on the neural level. Although the nature of these regions – along with proposed integrative mechanisms - have varied somewhat from study

to study, overall the combined evidence seems to suggest that integration can occur at the encoding stage in early sensory cortices, right through to late cognitive evaluation in higher association areas. Emotional face–voice integration appears to be a complex process that cannot be related to a single neural event taking place in a single brain region, but rather engages an interactive, dynamic network with activity distributed in time and space. Further work can only serve to clarify the exact mechanisms of audiovisual emotion perception, and how these might be dependent on particular emotional expressions or context. Indeed, **Chapter 5** of this thesis attempts to supplement and develop this work on face-voice emotion integration, with an fMRI experiment that is detailed further in the thesis rationale.

1.4.3 Dynamic vs. static stimuli in face-voice integration studies

Significantly, much of the observed integration effects – whether this be integration of identity, gender or affective information - have been observed using relatively impoverished, unecological stimuli (i.e., static faces paired with voices, often obtained independently). Although this impressively illustrates the robustness of the ability to combine two sources of information, it is not representative of what we perceive in the environment, where faces are dynamic and synchronised with vocalisation. Thus, it is more appropriate in research involving perception of faces to use dynamic stimuli, as these are encountered in real life. Furthermore, it is proposed that integration effects would be stronger when dynamic faces were used (Campanella and Belin, 2007; Schweinberger et al., 2007; Sugihara et al., 2006), and that dynamic faces are processed differently to static faces.

For example, with regards to affect, neuroimaging studies known to be implicated in the processing of facial emotion (e.g., the pSTS, amygdala and insula) respond more to

dynamic than static facial expressions (e.g. Haxby et al., 2000; Kilts et al., 2003); and there have also been cases where neurologically affected individuals that were incapable of recognising static facial expressions could recognise the same expressions expressed dynamically (e.g. Adolphs et al., 2003). Furthermore, the aforementioned study by Schweinberger et al. (2007) found that the presentation of time-synchronised articulating faces influenced more strongly the identification of familiar voices than when accompanied by static faces.

Kreifelts and colleagues (2007, 2009, 2010) were one of the first groups to use dynamic stimuli in their studies. In these experiments, the authors video captured actors expressing words spoken in either neutral or one of six emotional intonations with a congruent emotional facial expression, enabling them to create a set of dynamic stimuli. This approach works well if one is only looking to compare congruent audiovisual to unimodal information: however, if a researcher is aiming to examine incongruence effects, this makes the problem of dynamic stimuli more complex. This requires a pairing of different information (e.g., expression) in the face and voice, which would be virtually impossible to capture in a simple video recording of one actor.

This was exactly the dilemma faced by Schweinberger et al. (2007), whose study aimed not only to investigate how recognition of familiar voices was affected when they were combined with a face of corresponding or non-corresponding speaker identity, respectively, but also to examine if static and dynamic face elicited different effects on participants' recognition responses. In order to create audiovisual stimuli which were not only congruent but also incongruent (e.g., familiar voice paired with an unfamiliar face) the authors used a novel approach which is described fully in their paper, but discussed briefly here. Eight people (four familiar, four unfamiliar) were video-recording saying a

sentence with standardised timing, and these recordings were then adjusted to a uniform duration of 1700ms. Videos were edited (and timing adapted where necessary) such that video and audio tracks could be recombined both within and across speakers, preserving synchronisation. Editing consisted of inserting or deleting periods of relative silence in the audio tracks and of inserting or deleting video frames during relatively motionless periods. The authors also noted that in natural utterances starting with stop consonants (as did their sentence), articulatory movements of the face precede the onset of the audio speech signal by a few tens of milliseconds. Thus, the onset of visual articulation was identified as the first frame of speech motion and using a fixed delay of 80 ms to account for this visual lead in the present stimuli, the initial consonantal burst in the audio file was defined as the acoustic onset and was aligned with the onset of visual articulation. In this way, the authors managed to create incongruence within an audiovisual stimulus, and also preserve its ecological validity.

Collignon et al. (2008) also used dynamic face-voice stimuli, within the context of affect perception. Similar to Schweinberger et al. (2007), the authors also wished to create a set of incongruent stimuli, but used a different approach. They obtained emotive dynamic visual stimuli and nonverbal vocal clips from two separate validated databases (Simon et al., 2008 and Belin et al., 2008 respectively) and combined these to create a range of congruent and incongruent audiovisual stimuli. Although this certainly develops upon the stimuli used in previous studies, it should be noted that multisensory stimulus integration relies on spatial and temporal coincidence (King and Palmer, 1985; Stein and Wallace, 1996) and thus respective paradigms therefore call for precisely matched dynamic stimuli. Thus, simply matching visual and auditory stimuli obtained from independent databases is unlikely to provide the ideal temporal synchrony required as the stimuli by their nature have been gathered to be part of unique unimodal, not audiovisual, sets.

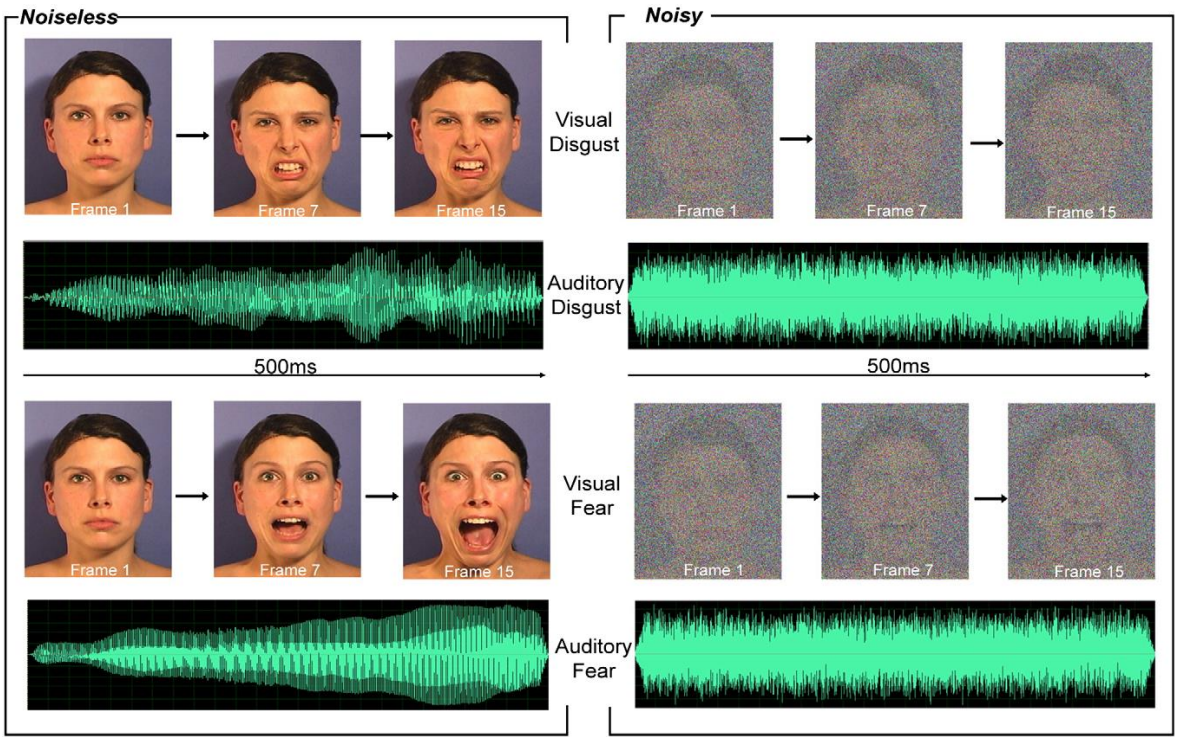


Figure 1.18. Stimuli used in Collignon et al. (2008). Stimuli consisted in video (from Simon et al., 2008) and non-linguistic vocal clips (from Belin et al., 2008). Depending on the task, the clips were either displayed in noiseless condition or were presented with the addition of noise in order to decrease the reliability of the sensory information. These stimuli were either displayed alone and in bimodal congruent (the same expression in both modalities) or bimodal incongruent (different expressions in both modalities) combinations. Figure taken from Collignon et al. (2008).

Klasen et al. (2011) employed a novel approach by using dynamic virtual characters (avatars) exhibiting angry, neutral, and happy facial emotions and combining them with pseudowords with angry, neutral, and happy prosody in congruent and incongruent trials. Although avatars do not provide a 'real' facial image, the effectiveness of virtual characters for emotion recognition tasks has been successfully validated in patient and control populations (Dyck et al., 2008, 2010; Wallraven et al., 2008). The authors used a lip synchronization tool that allows for a precise matching of speech and lip movements. The authors were therefore able to combine face and voice information in such a way that allowed the study to be the first to investigate the supramodal representation of emotional

information with dynamic stimuli expressing facial and vocal emotions congruently and incongruently.

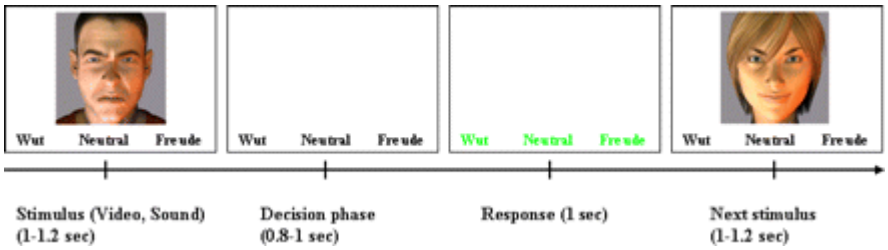


Figure 1.19. Stimuli used in Klasen et al. (2011). The authors used dynamic virtual characters (avatars) exhibiting angry, neutral, and happy facial emotions and combined them with pseudowords with angry, neutral, and happy prosody in congruent and incongruent trials, assuring standardized facial expressions and perfect lip–speech synchronization. Figure taken from Klasen et al. (2011).

In the experimental work presented in this thesis, we took care to provide our participants with an experience as close to real life as possible by using dynamic stimuli, where vocalisations were time-synchronised with facial articulation. An overview of the experiments in this thesis, and basis behind them, is provided below in the thesis rationale.

1.5 Thesis rationale

As previously described in a review in *Trends in Cognitive Sciences*, Belin et al. (2004) suggested a model of voice perception similar to Bruce and Young's model of face perception (Bruce and Young, 1986). This model is also further described in a recent and updated review in the *British Journal of Psychology* (Belin et al., 2011). Relevant to this thesis, the Belin et al. (2004) model suggests that the pathways for voice processing are analogous to, and *interacting* with equivalent functional pathways involved in facial processing during audiovisual integration (see also Campanella and Belin, 2007). Notably,

these include not only the speech-processing pathways, but also those responsible for interpreting non-linguistic information. Within this model of voice processing the authors propose a supramodal stage of information processing, where interactions between the different stages of voice and face would lead to the recognition of the person's identity or emotion. It is important to note that this model does not propose that all aspects of face and voice processing are exactly similar: for example, it has been suggested that whereas sex and identity information appear to be processed independently for faces, their processing might not be independent for voices (Burton, and Bonner, 2004). Nonetheless, the model provides a good framework for envisaging the interactions that may occur between face and voice processing.

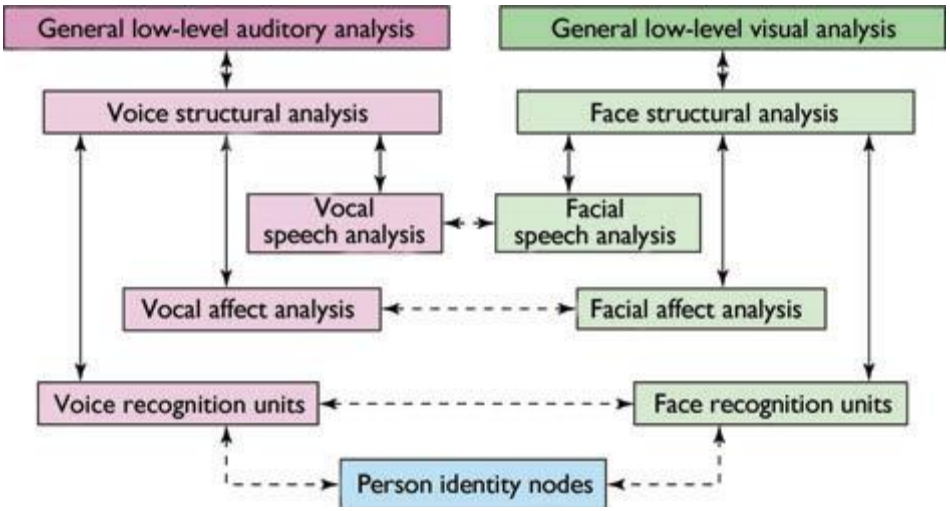


Figure 1.20. The Belin et al. (2004) model of voice perception. Dotted arrows indicate interactions between face and voice processing pathways. Figure taken from Belin et al. (2004).

For the purpose of this thesis, I concentrate on the interactions proposed as part of their suggested model with experimental **Chapters 3, 4** and **5** designed to explore factors related to the integration of information from the face and the voice, with a main focus on paralinguistic processing.

Firstly, **Chapter 3** provides a more general investigation of face-voice integration. The described experiment uses fMRI to explore the neural correlates of audiovisual integration under passive conditions, with no specific focus on speech information or one type of paralinguistic information. Specifically, here the aim was to define regions dedicated to selectively integrating face-voice information, as compared to information from non-face and non-voice information (i.e., objects). We also investigated convergence of unimodal sources without an emphasis on integration, specifically identifying regions which may be 'heteromodal'. The focus was on the STS region in particular, due to its much documented involvement in many aspects of audiovisual integration, as well as unimodal face and voice processing.

The latter two chapters target two integration of two different types of paralinguistic information: gender, and emotion. **Chapter 4** of this thesis outlines a psychophysical experiment exploring integration of gender information from the face and the voice, which also examines how attentional demands can affect the integration process. This has further allowed for the inference regarding the automaticity of audiovisual gender perception, along with modality dominance. **Chapter 5** follows on from this and uses fMRI to investigate the bimodal perception of emotion. The experiment detailed in this chapter made use of stimulus adaptation or the 'repetition effect' in order to move beyond the spatial limitations of fMRI, and infer the properties of single neurons. Specifically, the hypothesis was that 'true' multisensory neurons – those single neurons that integrated information from two modalities – would show crossmodal adaptation effects.

Additionally, in both of the latter chapters we made use of the possibilities offered by the recent developments in both facial and auditory morphing techniques in order to create a range of novel, up-to-the-minute audiovisual stimuli that were parametrically morphed in

both modalities. To my knowledge, this is the first time that information has been morphed in both modalities, either paralinguistic or otherwise. Additionally, our face-voice stimuli were dynamic with time-synchronised vocalisations, in order to provide an ecological experience that has rarely been seen in previous experiments. In **Chapter 4** we also directly compare the behavioural response to dynamic vs. static information, in an attempt to quantify any behavioural gains (as previously implied in audiovisual speech and identity research) of presenting articulating faces.

Before moving onto the experimental chapters, **Chapter 2** provides an overview of methods used within this thesis – specifically, the theory behind and practical application of MRI and fMRI; statistical criteria used in audiovisual fMRI experiments; and a discussion on neural adaptation or priming. I will also familiarise the reader with the experimental design used in **Chapter 5** - the so-called 'continuous carry-over' design (Aguirre, 2007). Although both **Chapter 4** and **5** employ morphing techniques, the stimulus preparation in each utilises slightly different procedures and software – **Chapter 4** uses 'regular' two-dimensional video recording and 'Psychomorph' software (Tiddeman and Perrett, 2001) for face morphing, whilst **Chapter 5** uses Did3 video capture software, in conjunction with Matlab, in order to morph emotional facial expressions. Although both chapters use the same voice morphing algorithm (STRAIGHT; Kawahara, 2003, 2006) the full description of stimulus preparation is detailed within the respective experimental chapters, for ease of read.

2. Thesis methods

2.1 Magnetic Resonance Imaging

This section will introduce the concept of magnetic resonance imaging (MRI), and the fundamental processes when acquiring an MRI image. This will then be followed by a description of a specific MRI technique, functional magnetic resonance imaging (fMRI). fMRI is now used widely as a clinical and neuropsychological tool to better understand and characterise brain function, and is the neuroimaging technique employed in this thesis.

2.1.1 NMR theory

Medical magnetic resonance imaging (MRI) is a technique used for obtaining high-resolution images of organs within the human body. MRI relies on the nuclear magnetic resonance (NMR) of hydrogen nuclei, which are found in lipid molecules and the water of human tissue. Hydrogen nuclei consist of a single proton possessing a 'nuclear spin'. Spin is a fundamental property of nature, akin to electrical charge or mass. When placed in a uniform external magnetic field of strength F0, a particle with a net spin precesses around F0 with an angular or resonance frequency ω, which depends on the gyromagnetic ratio (γ) of the particle. This frequency is known as the Larmor frequency, and can be defined as:

$$\omega = \gamma.F0$$

Ordinarily, hydrogen nuclei are orientated randomly; however, the spin of the proton in a magnetic field has a magnetic moment vector, causing it to create micro-magnetic fields around itself. In the magnetic field, the proton behaves like a tiny bar magnet, with the north and south poles along the axis of spin, and will align itself with (parallel) or against

(anti-parallel) the applied field, corresponding to low and high energy states respectively. The proportion of magnetised nuclei aligned in either direction depends on both the strength of the magnetic field and thermal agitation. At thermal equilibrium, the number of spins in the lower energy level, N↑, slightly outnumbers the number in the upper level, N↓, forming the bulk magnetisation vector. However, transitions between the two states can be induced by applying electromagnetic energy at the Larmor frequency. Applying an RF pulse (F1) perpendicular (90 degree impulse) to F0 at the Larmor frequency will rotate the net magnetisation out of the F0 alignment into the transverse plane. The relative angle of this rotation is determined by the RF magnitude and duration, and is known as the flip angle (FA).

Following application of an RF pulse, a small voltage (or signal) is induced in the receiver coil due to the oscillating transverse component of the magnetisation. The NMR signal decays in the absence of F1 as a result of NMR relaxation processes, and is known as the Free Induction Decay (FID) signal.

The relaxation process in resonance is controlled by the biological parameters T1 and T2. Both are tissue dependent and provide a means of differentiating among different tissues. T1 is the time constant characterising the rate at which excited nuclei dissipate excess energy to the environment (lattice), referred to as the spin-lattice relaxation time (or longitudinal relaxation time). T2 is the time constant characterising the rate at which excited nuclei exchange energy, and is referred to as the spin-spin (or transverse) relaxation time because it is the loss of transverse magnetisation that determines the T2 relaxation time. However, in our actual environment, the NMR signal decays faster than T2 would predict. The assumption when characterising pure T2 decay is that the main external F0 field is completely homogenous. In reality, there are a number of factors

creating imperfections in the homogeneity of a magnetic field (e.g., in an MRI experiment, manufacturing flaws in the main magnet). The resulting inhomogeneity in the field causes adjacent protons to precess at slightly different frequencies. Every tissue has a different magnetic susceptibility that distorts the field at tissue borders, and the sum total of all these effects is called T2* ('real world' T2*).

2.1.2 The MRI experiment

In an MRI experiment, three different gradients which can spatially localise where the collected signal originated are used to build an accurate representation of the organ being scanned. These are the slice selection and thickness, phase encoding, and frequency encoding gradients. Slice selection and excitation is achieved by the use of an RF pulse applied in the presence of a magnetic field gradient. The RF pulse is modulated by a frequency envelope such as a sinc or Gaussian waveform, and has a narrow frequency bandwidth. Only protons with resonant frequencies within this range will be excited, and contribute to the resultant MR signal. The slice thickness can be controlled either by changing the gradient strength or by altering the bandwidth of the RF pulse, with the envelope of the RF pulse controlling the slice profile.

Having defined a slice, it will have been localised in one direction by selective excitation. The frequency and phase encoding gradients work to spatially localise the spins within a slice, and are applied after excitation. The excited spins then precess at a particular frequency and phase angle depending on their location, and this allows for the individual signals to be distinguished from one another.

The MRI signal rapidly decreases as the individual pixels signals get out of phase with each other, and this can cause the signal to disappear before it can be measured. To avoid

this, a dephasing gradient is applied before acquiring the data. During readout, the MRI signals are rephased giving a maximum signal when the gradient areas are equal - a gradient echo. Once the MRI signal has been collected, the frequency information is extracted via a Fourier transform (FT), which gives the amplitude at each frequency.

In order to achieve high spatial localisation, the pulse sequence is repeated several times with the size of the phase gradient changed each time (or 'stepped'). By repeating the pulse many times (e.g. 256 times for a 256x256 matrix image), a 2D data set can be built up. The application of all the gradients selects an individual slice, with frequency encoding along one axis of the slice, and phase encoding along the other axis. The system can now locate an individual signal within the image by measuring the number of times the magnetic moments cross the receiver coil (frequency) and their position around the precessional path (phase). When data of each signal position is collected, the information is written in K Space. K Space is the complement to the image space and the image formed is the Fourier Transform of the K Space data.

2.1.3 Functional magnetic resonance imaging (fMRI)

fMRI is an advanced MRI technique for measuring brain activity. It works by detecting the changes in blood oxygenation level and blood flow that occur in response to neural activity, and can be used to produce activation maps showing which part of the brain are involved in a particular mental process.

When a brain area is more active the neurons within it consumes more energy and therefore more oxygen. This increased oxygen consumption in the active neurons causes increased blood flow to and blood volume in the relevant neural tissue. Oxygen is delivered to neurons by haemoglobin in capillary red blood cells. Haemoglobin is

diamagnetic when oxygenated, but paramagnetic when deoxygenated. This difference in magnetic properties leads to small differences in the MR signal of blood depending on the degree of oxygenation, and these differences can be used to detect brain activity. One point of note is the direction of oxygenation change with increased activity: the haemodynamic response follows a more complex function than the simple decrease in blood oxygenation that would perhaps be expected. There is a momentary decrease in blood oxygenation immediately after neural activity increases, known as the 'initial dip'. This is followed by a period where the blood flow increases, but not just to a level where oxygen demand is met - it overcompensates for the increased demand, meaning that blood oxygenation actually increases following neural activation. The blood flow peaks after around 6-7 seconds and then falls back to baseline, often accompanied by a 'post-stimulus undershoot'. To exploit and record the susceptibility change due to increased oxygen fMRI uses a pulse sequence where images result from the blood-oxygen-level-dependent (BOLD) contrast. The pulse sequence used to assess the BOLD contrast is called Echo Planar Imaging (EPI).

In an fMRI experimental paradigm, the subject first undergoes a safety screening. This is used to ensure that the participant has no magnetic material on (e.g. jewellery) or within (e.g. pacemakers) them that could either align with the magnetic field or stop working, potentially causing harm to the participant. In addition, this screening can also raise counter-indications (e.g. pregnancy, claustrophobia) for participant scanning. After the subject has entered the scanner a short anatomical localiser localises the subject's head in the magnetic field before the functional run is performed. Following the T2* functional runs, a T1 anatomical scan is performed. This serves as an anatomical reference for the functional runs, which observed brain activity can be mapped upon. Finally, the BOLD signal is estimated using a haemodynamic response function (HRF), which is assumed to represent the overall response of the brain to stimulus presentation.

The type of response in an fMRI experiment means that the measured signal is an indirect and delayed reflection of neural activity: nevertheless, by comparing the response of brain regions in different conditions, inferences can be drawn about the conditions under which a brain region becomes more active. By systematically comparing conditions differing in a number of ways, we can learn more about a normal brain at work.

2.2 Statistical criteria in audiovisual fMRI experiments

This next section briefly discusses the different statistical criteria used to define brain regions as 'integrative', or 'audiovisual'. In **Chapter 3** of this thesis, the intention was to define regions which integrated audio and visual information (comparing the response to audiovisual and unimodal stimuli), and therefore it was necessary for us to select one of the following criteria to use. Consequently, I believe it is useful to provide an overview of these measures and give reason for the selection made.

In audiovisual integration imaging studies, along with the inferences drawn from the statistical criteria used to define different brain regions as 'active' or 'inactive' within the experimental manipulation, we have to seek for means to objectively define integrative fMRI responses. The definition of the appropriate analysis in audiovisual studies, in order to assign the active brain regions to different functional roles, is not straightforward: in a typical audiovisual fMRI experiment activity is typically found in many brain regions.

There has been much discussion around the pros and cons of the statistical criteria used to classify audiovisual integration when comparing bimodal to unimodal conditions using fMRI (e.g. Beauchamp, 2005; Calvert, 2001; Goebel and van Atteveldt, 2009; Laurienti et al., 2005; Stein et al., 2009; Love et al., 2011). Integrative effects can be modelled in a

number of ways, and a number of different statistical criteria have been proposed ranging from stringent to liberal: namely, the criterion of super-additivity, the 'max criterion' and the 'mean criterion', respectively. Regardless of the criterion used however, integration is typically defined by a positive outcome (enhancement); in this case the unimodal stimuli are assumed to 'bind together'. A negative outcome is typically interpreted as inhibited processing (suppression), which can be viewed as another type/direction of integration, for stimuli that are assumed to 'not bind together'. No difference between audiovisual and unisensory responses (additivity, no interaction) is interpreted as no integration - in this case, two inputs do not influence each other's processing in that voxel or region. Super-additivity, the max criterion and the mean criterion have been reviewed elsewhere (e.g. Beauchamp, 2005; Laurienti et al., 2005; Ethofer, 2006; Goebel and van Atteveldt, 2009; Love et al., 2011; James and Stevenson, 2012) but are described briefly below. An illustration of the different integrative responses is also provided in Figure 2.1.

2.2.1 'Super-additivity'

Calvert (2001) argued that the electrophysiological criteria for multimodal integration could be applied to the BOLD effect. Here, cells which subserve multimodal integration show responses to congruent information that exceed the sum of the responses to the unimodal stimuli, known as super-additivity (Bimodal (Congruent) > Unimodal 1 + Unimodal 2). In contrast, conflicting multimodal information results in a response depression in which the response to incongruent bimodal information is smaller than the stronger of the two unimodal responses (Bimodal (Incongruent) < Maximum (Unimodal 1, Unimodal 2).

Under the super-additive criterion, portions of the temporal, occipital, parietal and frontal lobes have all been proposed as part of a face-voice integration network. Two recent fMRI

studies, for example, report responses in sub-regions of all these lobes to be higher for audiovisual stimuli than the sum of both unimodal responses (Joassin et al., 2011a, 2011b). Similarly, Calvert et al. (1999) reported enhanced activity in regions of the temporal and occipital lobes for audiovisual speech perception relative to perceiving each cue in isolation. In a follow up study, the group also reported super-additive responses in the temporal, occipital, parietal and frontal lobes, whilst focussing their discussion on the left posterior superior temporal sulcus (pSTS) (Calvert et al., 2000).

At a theoretical level, super-additivity is attractive because it proposes using the same criterion that has been applied in recording studies of multisensory neurons. It is tempting to consider that neuroimaging measurements, like BOLD activation measured with fMRI, are directly comparable with findings from single-unit recordings. However, there remains a fundamental difference between BOLD activation and single-unit activity: BOLD activation is measured from the vasculature supplying a heterogeneous population of neurons, whereas single-unit measures are taken from individual neurons. The ramifications of this difference are not inconsequential because the principles of multisensory phenomena established using single-unit recording may not apply to population-based neuroimaging data.

Laurienti et al. (2005) point to a number of reasons why this might be the case: firstly, the proportion of AV neurons is small compared to unisensory neurons; secondly, of those multisensory neurons, only a small proportion are actually super-additive; and thirdly, super-additive neurons have low impulse counts relative to other neurons. To exceed the additive criterion, the average impulse count of the pool of bimodal neurons must be significantly super-additive for population-based measurements to exceed the additive criterion. However, the presence of super-additive neurons in the pool is not enough by

itself because those super-additive responses are averaged with other sub-additive, unisensory, and even suppressive, responses. Therefore, it seems super-additivity will be unlikely to be observed because the heterogeneity of these response types that may cancel each other out at the voxel level (Beauchamp 2005b; Laurienti et al. 2005). Indeed, although some early studies successfully identified brain regions that met the super-additive criterion (Calvert et al. 2000, 2001), subsequent studies did not find evidence for super-additivity even in known multisensory brain regions (Beauchamp 2005; Beauchamp et al. 2004a, 2004b; Laurienti et al. 2005; Stevenson et al. 2007). In summary, even though the super-additive criterion is appropriate because it represents the correct null hypothesis, the statistical distribution of cell and impulse counts in multisensory brain regions may make it particularly inflexible as a criterion. Consequently, it may be overly strict and introduce type II (false-negative) errors (Beauchamp, 2005).

However, conversely it can actually also result in false positives due to a negative response in one of the modalities. Super-additivity can often be biased towards classifying a multisensory response as integrative in sensory-specific brain regions. For example, this response was seen in a study by Love et al. (2011), where a significant super-additive effect in the bilateral occipital gyrus was driven by the audiovisual condition being contrasted to the sum of a positive visual response and a large negative auditory response. Joassin et al. (2011a, 2011b) also highlight that their super-additive effects in the occipital and temporal cortex were the result of the bimodal response being compared to the sum of a positive and a negative unimodal response. However, the interpretation of this situation is complicated and it remains an open question whether we can really infer integration from this type of response profile (Calvert et al., 2001; Goebel and van Atteveldt, 2009). It is particularly problematic because many recent studies support involvement of low-level sensory-specific brain regions in multisensory integration (see Schroeder and Foxe, 2005;

Ghazanfar and Schroeder, 2006 for reviews). The super-additive criterion is often described as the strictest of the multisensory integration criteria, but this is actually only true when the implementation of it is restricted to brain regions showing increased activity for both unimodal conditions relative to baseline. Otherwise, 'sensory-specific' cortices, which deactivate to stimulation of other senses, are likely to be categorised as super-additive and multisensory in nature (Goebel and van Atteveldt, 2009). In order to avoid this, one option could be to initially apply a heteromodal contrast (A> baseline V > baseline) which ensures significant activation to unisensory activation in both modalities.

2.2.2 The 'max-criterion'/'conjunction' analysis

At the single neuron level, the max criterion states that the multisensory fMRI response should be stronger than the most effective unimodal response (AV > Maximum(A,V); Stein and Meredith, 1993). Although this approach can also be used for fMRI data, more common has been to use a conjunction analysis to investigate brain areas that show a significantly stronger response to bimodal stimuli than to unimodal stimuli of *both* modalities AV > A ∩ AV > V. The classification based on the max criterion seems most robust to different unisensory response profiles. Such an approach has been used to identify, for example, the superior colliculus (SC) – a well-recognised multisensory structure – and the bilateral superior temporal cortex (STC) to be loci of face-voice integration (e.g. Wright et al., 2003; Kreifelts et al., 2010; Szycik et al., 2008). Qualitatively, this criterion is less strict than super-additivity (granted that there is not deactivation in one modality). Still, one disadvantage is it can induce a slight loss in sensitivity. Such a loss in sensitivity becomes critical when two different contrasts that are expected to yield small effects are submitted to such an analysis. Two ways to increase the sensitivity of these conjunctions is by correcting the search volume to small anatomical regions (regions of interest (ROI)), or to define separate conjunction analyses for specific

emotions, for example (AV happy – A happy) ^ (AV happy – V happy) (as in Pourtois et al. (2005); see also Ethofer et al. (2006)). In **Chapter 3**, this is the criterion we chose to define our audiovisual region, due to the level of its stringency and minimal disadvantages (certainly compared to other defining criteria).

2.2.3 The 'mean criterion'

The mean criterion can be expressed as AV > (A + V)/2. In other words, it requires that the response to an audiovisual stimulus is bigger than the average of the two unimodal responses. This contrast can provide a useful index of the degree of multisensory integration in an area, and because it reflects the contribution of both unisensory responses it can test the null hypothesis that the response across all conditions is similar. Because it is more liberal than both super-additivity and conjunction analyses, it is relatively able to identify presumed multisensory regions, including the STS. However, a disadvantage is that it may be too liberal, especially when one of the unisensory responses is weak or negative. This will reduce the mean in such a way that a multisensory response exceeds the mean even when weaker than the largest unisensory response. Therefore, the mean criterion can be misleading: like super-additivity, without an initial criterion which requires unisensory activation in each modality to be above 0, the mean criterion can classify purely unimodal regions as multisensory. Although for non-speech stimuli it has been used to classify areas of the superior temporal cortex (STC) as multisensory (e.g. Beauchamp et al., 2004), there have been no fMRI studies that have used the mean criterion to implicate brain regions as sites of integration for face and voice (Love et al., 2011). In an investigative paper exploring this topic, Love et al. (2011) found that using the mean criterion, the occipital and temporal regions were implicated as integrative regions. However, examination of response profiles from these regions showed almost no

difference between the response to the combined face–voice and the ‘sensory-specific’ unimodal response of the region.

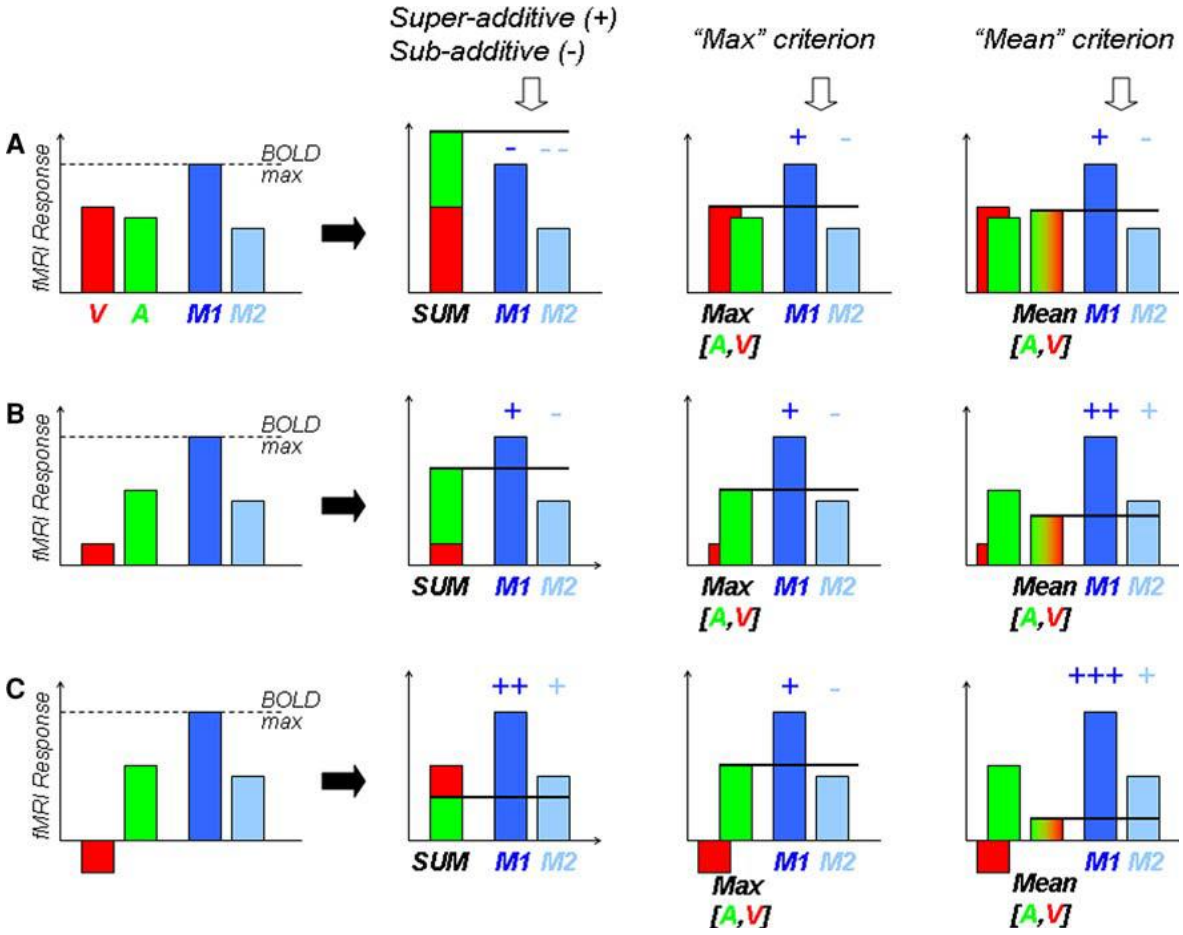


Figure 2.1. The use of different statistical criteria for hypothetical brain regions with different unisensory (fMRI) response profiles (a–c). a. Heteromodal response: a significant positive response to both unisensory stimulation modalities (auditory and visual). b. A positive auditory response and a weak, positive visual response. c. A positive auditory response and a negative visual response. *Bars* indicate the fMRI activation level for different unisensory and multisensory stimulation conditions: visual (V *red*), auditory (A *green*), and two different audiovisual/multisensory conditions (M1 *dark blue*; M2 *light blue*). The *dotted line* in the first column ('BOLD max') represents the maximal fMRI. The *solid lines* in columns 2–4 represents the degree of BOLD activation required for the different criteria: summed unisensory activation level (A + V) for the super-additivity criterion, maximal unisensory activation level ([A, V]max) for the 'max' criterion, and mean unisensory activation level (A + V)/2 for the 'mean' criterion. *Plus* and *minus symbols* indicate whether this BOLD activation meets this level (i.e., + = enhancement, - =suppression). The number of plus or minus signs indicates the strength of this enhancement/suppression. Figure taken from Goebel and van Atteveldt (2009).

Ethofer (2006) notes, however, that perhaps *none* of these approaches can be considered as the optimal method to clarify as to which brain structures participate in multisensory integration. Rather, each of these analyses highlights different aspects of the interplay of brain regions in integrative processes, thus providing complementing information. Indeed, this was also the conclusion drawn by Love et al. (2011). They suggest that an overemphasis on super-additivity as being the litmus test for multisensory integration and that a failure to explore other criteria could have a detrimental effect on our understanding of integration mechanisms (see also Stanford and Stein, 2007). Thus, they argue that multisensory research using fMRI would benefit from exploring several integration criteria within the same experiment. Goebel and van Atteveldt (2009) also propose that because all the main criteria for comparing multisensory to unisensory responses have limitations, an alternative would be to manipulate the congruency of the different inputs. In this type of analysis, two bimodal conditions are contrasted with each other (congruent vs. incongruent), which eliminates the unimodal component from the metric. This comparison follows the assumption that only in a congruent condition the unimodal inputs are integrated successfully, and therefore the contrast of congruent vs. incongruent can be used as a supplemental criterion for multisensory integration. Indeed, this was one of the approaches used later in this thesis, in **Chapter 5**.

2.3 'Continuous carry-over' designs

In **Chapter 5** of this thesis, I also use a specific type of event related design, called a 'continuous carry-over' design (Aguirre, 2007). A 'carry-over' effect is the modulation of the neural response to the current stimulus by the previously presented stimulus – a type of neural adaptation. Below I provide more background on neuronal adaptation and further discuss the continuous carry-over design in more detail.

2.3.1 Adaptation or repetition suppression

Adaptation – or priming - studies are often used to support inferences regarding neural populations within voxels. The typical procedure is to adapt a neuronal population by repeating the presentation of the same stimulus in a control condition (leading to a reduction in fMRI signal), and to vary one stimulus property and further assess recovery from adaptation. In theory, the first presentation of a stimulus would probe neurons sharing common functional characteristics (i.e. coding for the same stimulus dimensions, such as the same colour or shape), or responding to the same object of stimulation, increasing processing efficiency of the repeated presentation (Grill-Spector et al., 2006) and further leading to a decrease in measured brain signal (Henson and Rugg, 2003).

In cognitive neurosciences, adaptation has also been termed 'repetition suppression' (Desimone, 1996, Grill-Spector et al., 2006). Repetition suppression persists even when unrelated stimuli are presented between the repetition (Miller and Desimone, 1994) and increases with increasing number of repetitions of the stimulus (Li et al., 1993). This neural repetition effect has been reported at multiple spatial scales, from the level of individual cortical neurons in monkeys (Li et al. 1993; Miller and Desimone 1994; Sobotka and Ringo 1996) to the level of hemodynamic changes measuring the pooled activation of millions of neurons in humans using fMRI (Demb et al. 1995; Stern et al. 1996; Grill-Spector et al. 1999; Jiang et al. 2000; Naccache and Dehaene 2001).

The exact neural mechanisms behind repetition suppression are still not fully understood. However, a number of theoretical models have been suggested to try better explain the adaptation effect. Firstly, the model of *fatigue* suggests that if a neuron initially responds to the stimulus, a proportional decrease in firing rates is observed with repetition of the stimulus. Secondly, the model of *sharpening* proposes that if in a given region a neuron

processes features that are irrelevant for repetition suppression it stops firing. This would lead to fewer responsive neurons within this specific region/voxel, and consequently a smaller brain signal. Finally, within the model of *facilitation* an initial then repeated presentation of a stimulus would lead to faster processing, as indexed by a decrease in the duration of neural firing (Henson and Rugg, 2003; Grill-Spector et al., 2006). The three models share one common property in that the repeated presentation of the object of stimulation increases processing efficiency (Henson and Rugg, 2003; Grill-Spector et al., 2006).

Some potential underlying neural mechanisms have also been suggested to play a role in inducing repetition supression. The first one is the firing-rate adaptation. Here, the increase in potassium ion currents, further leading to an increase in conductance, would reduce the importance of synaptic input and reduce the probability of neural firing. Another mechanism proposed is synaptic depression, which is characterised by a temporary reduction in synaptic efficacy. This would reflect a reduction in the release of neurotransmitters before the synaptic moment. Finally, a mechanism referred to as long-term depression would involve plasticity changes at multiple processing stages leading to the reduction of synaptic efficacy (Grill-Spector et al., 2006).

An application of the repetition suppression phenomenon has been suggested as an experimental design for functional brain imaging studies. This design was termed “fMR-adaptation" (fMR-A; Grill-Spector et al., 1999; Grill-Spector and Malach, 2001; Grill-Spector et al., 2006). In this framework, neuronal populations are adapted by repeated presentation of a single stimulus. In a typical fMR-A experiment, pairs or blocks of stimuli that are the same or different along a dimension of theoretical interest are presented. If adaptation remains (fMRI signal stays low), this is taken to mean the adapted neurons

respond invariantly to the manipulated property, whereas a recovered fMRI signal indicates sensitivity to that property, i.e., that a different set of neurons is responding within the voxel. Since (presumably) only the targeted neural population adapts, its functional properties can be investigated without being mixed with responses of other neural populations within the same voxel. In the visual system for example, heterogeneous clusters of feature-selective neurons (e.g., for different object orientations) within voxels were revealed using fMR-A (Grill-Spector et al., 1999), whereas in a more standard stimulation design, the averaged voxel response was not different for the different features since all of them activated a neural population within that voxel.

Overall, the fMR-A method can tag specific neuronal populations within an area and investigate their functional properties non-invasively, and thus, can provide a powerful tool for assessing the functional properties of cortical neurons beyond the spatial resolution of several mm imposed by conventional fMRI (where one voxel contains several hundred thousand - potentially highly selective – neurons whose activity is averaged out by the fMRI signal).

Stimulus adaptation works in contrast to a ‘direct effect’ – the average BOLD response to multiple presentations of a given stimulus, used to determine the direct result of stimulus variation upon the amplitude of neural response. The direct effect provides a stimulus response function that relates modulation of a stimulus to the average response across a population of neurons within a voxel. These different modes of fMRI inference yield complementary information regarding the neural representation of stimuli, and have traditionally been used as part of separate experiments. However, Aguirre (2007) proposes that it may be advantageous to measure these effects simultaneously in an fMRI experiment both for the sake of efficiency, as well as for the opportunity to examine the

relative contribution of different forms of neural coding to the representation of stimulus variation.

2.3.2 The continuous carry-over experiment

The continuous carry-over design is an experimental paradigm that makes use of the repetition phenomenon in a well-controlled fashion. This design is decribed in full in Aguirre (2007) but is briefly summarised here. In studies using this type of design, stimuli are presented in an unbroken, sequential manner, allowing the experimenter to not only measure the mean difference in neural activity between stimuli (the 'direct effect'), but also the effect of one stimulus upon another (the 'carry-over', or adaptation effect). With this approach, the adapting effects of stimuli may be studied in a continuous sequence, as opposed to within isolated blocks. These studies are ideally conducted with serially balanced sequences, in which every stimulus precedes and follows every other stimulus (i.e., the 'Type 1 Index 1' sequence), which allow for efficient and unbiased estimation of both the direct and carry-over effects, and accounts for stimulus counterbalancing.

Carry-over designs are particularly useful when there is more complex variation in a set of stimuli (e.g. a set of cuboids differing in length, height and breadth; musical notes varying in tone, duration and amplitude), whose differences can be expressed as changes along a number of different axes. In **Chapter 5**, I use stimuli in which affective information in both the face and voice is parametrically varied: thus, a carry-over design is suited with regards to these set of stimuli. In line with the fMR-A framework (Grill-Spector et al., 1999; Grill-Spector and Malach, 2001, Henson and Rugg, 2003), a reduction of BOLD signal magnitude should be observed with the repetition of two stimuli, proportionally to the amount of shared physical properties. The continuous carry-over design also allows us to investigate categorical main effects. Because of the sequential ordering of the stimulus

presentation, every stimulus can be studied on its own and compared to each other. For example, one might be interested in investigating whether different brain regions respond to happy and angry voices. This could be addressed by directly comparing the brain signal following each presentation of the 90% angry voice with the brain signal following every presentation of the 90% happy voice. Similarly, a comparison of incongruent vs. congruent information would simply involve a contrast of the response to the appropriate stimuli. In summary, the opportunity to observe both direct and adaptation effects provides the experimenter with a valuable opportunity to address a number of important questions within the same design, as opposed to carrying out a number of different experiments.

3. People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus

3.1 Abstract

The functional role of the superior temporal sulcus (STS) has been implicated in a number of studies, including those investigating face perception, voice perception, and face-voice integration. However, the nature of the STS preference for these 'social stimuli' remains unclear, as does the location within the STS for specific types of information processing. The aim of this study was to directly examine properties of the STS in terms of selective response to social stimuli. We used functional magnetic resonance imaging (fMRI) to scan participants whilst they were presented with auditory, visual, or audiovisual stimuli of people or objects, with the intention of localising areas preferring both faces *and* voices(i.e., 'people-selective' regions) and audiovisual regions designed to specifically integrate person-related information. Results highlighted a 'people-selective, heteromodal' region in the trunk of the right STS which was activated by both faces and voices, and a restricted portion of the right pSTS with an integrative preference for information from people, as compared to objects. These results point towards the dedicated role of the STS as a 'social-information processing' centre.

3.2 Introduction

In the last decade, the human superior temporal sulcus (STS) and surrounding areas have been widely studied (see Hein & Knight, 2008 for a review). The STS is a major sulcal landmark in the temporal lobe, lying between cortices on the surface of the superior

temporal gyrus (STG) and middle temporal gyrus (MTG). An extensive region, it can be divided into three distinct sections: the anterior, mid, and posterior STS (aSTS, mid-STS, pSTS). Furthermore, in most individuals, the pSTS divides into two spatially separable terminal ascending branches - the so-called anterior and posterior terminal ascending branches. Thus, the STS can also be anatomically separated into the branch, bifurcation (equivalent to pSTS) and trunk parts (equivalent to mid-STS, aSTS) (Ochai et al., 2004). There is now a large body of evidence which suggests the STS is a major player in social perception – particularly, the pSTS region. This evidence has been provided from two separate camps of research; the first which has investigated unimodal face and voice processing, and the second which has pointed to the role of the pSTS in multisensory integration of social signals (Allison et al. 2000).

We rely greatly on information gathered from both facial and vocal information when engaging in social interaction. Along with the inferior occipital gyri (IOG) and lateral fusiform gyrus (specifically, the fusiform face area (FFA) (Kanwisher et al., 1997) the pSTS has been highlighted as a key component of the human neural system for face perception (Haxby et al., 2000). It appears to be particularly involved in processing the more changeable aspects of faces: when attending to these aspects the magnitude of the response to faces in the FFA is reduced and the response in the pSTS increases (Hoffman and Haxby, 2000). Although perhaps not as strong as for faces, evidence for voice-selective regions, particularly in the STS, is accumulating. Several fMRI studies (e.g. Belin et al., 2000; Ethofer et al., 2009; Grandjean et al., 2005; Linden et al., 2011) have demonstrated the existence of voice-selective neuronal populations: these voice-selective regions of cortex (‘temporal voice areas’(TVA)) are organized in several clusters distributed antero-posteriorly along the superior temporal gyrus (STG) and STS bilaterally, generally with a right-hemispheric preponderance (Belin et al., 2000; Kreifelts et al.,

2009). The aSTS and pSTS in particular appear to play an important role in the paralinguistic processing of voices, such as voice identity (Belin et al. 2003; Latinus et al. 2011; Andics et al. 2010). Thus parts of the pSTS appears to show greater response to social signals compared to non-social control stimuli in both the visual and auditory modalities, although the relative location of face- and voice-sensitive regions in pSTS remains unclear.

Turning away from unimodal face and voice processing, another vital skill for effective social communication is the ability to combine information we receive from multiple sensory modalities into one percept. Converging results point to the role of the pSTS in multisensory integration, particularly in audiovisual processing. The logic of fMRI experiments on audiovisual integration has been to search for brain regions which are significantly involved in the processing of unimodal visual and auditory stimuli, but show an even stronger activation if these inputs are presented together—the so-called 'supra-additive response', where the response to the bimodal stimuli is larger than the sum of the unimodal responses. Integration of speech (Calvert et al., 2000; Wright et al., 2003; Love et al., 2011), affective (Ethofer et al., 2006; Pourtois et al., 2005; Kreifelts et al., 2009), and identity (Blank et al., 2011) information from faces and voices have all been found in the pSTS. However, it should also be noted that integration of 'non-social' information – such as tools and their corresponding sounds (Beauchamp et al., 2004) and letters and speech sounds (van Atteveldt et al., 2004) – has also been observed in the pSTS, and thus it is unclear whether this region performs a more 'general' integrative role, or shows preferences for particular stimulus categories.

Here we brought together these distinct lines of research by examining properties of the STS in terms of selective response to social stimuli. Normal adult volunteers participated

in an ‘audiovisual localiser’ scan during which they were stimulated with auditory, visual, or audiovisual stimuli of people or objects. We proposed, given that face-selective, voice-selective and integrative regions are all located within the STS, that in addition to areas preferring both faces *and* voices(i.e., ‘people-selective’ regions) there could also be audiovisual regions that are more sensitive to social stimuli, as compared to information from non-social categories, such as objects.

We found that a restricted portion of the right pSTS was characterised by a conjunction of (1) an ‘integrative’ response, i.e. stronger response to audiovisual stimuli compared to visual and compared to auditory stimuli and (2) ‘people-selectivity’, i.e. preference for social stimuli irrespective of the modality (voice > objects; face > objects). Furthermore, a large region further extending down the trunk of the right STS was observed to be heteromodal: that is, this region was activated by both faces and voices, but did not necessarily show integrative properties.

3.3 Materials and Methods

3.3.1 Participants

Forty English-speaking participants (15 males and 25 females; mean age: 25 years ± 5 years) took part in the scan. All had self-reported normal or corrected vision and hearing. The ethical committee from the University of Glasgow approved the study. All volunteers provided informed written consent before, and received payment for, participation.

3.3.2 Stimuli

24 people (12 males and 12 females) were video-recorded producing a variety of vocal expressions, both speech and non-speech (e.g. saying the word ‘had’, humming, yawning).

Recordings took place in the television studio at the Learning and Teaching Centre, Glasgow University, and participants were paid at the rate of £6 per hour. The participants were filmed under standard studio lighting conditions (standard tungsten light), and sat directly facing the camera, at a distance so that the whole face was in frame. Videos were recorded with 25 frames per second (40ms per frame) using a Panasonic DVC Pro AJD 610 camera, fitted with a Fujiform A17 x 7.8 BERM-M28 lens, and transferred and edited using Adobe Premier Elements. Within the video recording, vocalisations were recorded with 16-bit resolution at a sampling frequency of 44100 Hz. Under the same conditions, 24 moving objects producing sound were also filmed (e.g. a moving toy car, a ball bouncing, a violin being played). The objects were filmed with the intention of recording the canonical view. Videos were edited so that every production of a vocal sound by a participant formed a separate clip, with the clips lasting two seconds each. The videos of the objects were edited to form separate clips of two seconds each also.

Stimulus clips were combined together in Adobe Premier Elements to form 18 different 16 second blocks. Thus, each block contained eight different stimuli. These blocks were broadly categorised as:

1) Faces paired with their corresponding vocal sounds (AV-P)
2) Objects (visual and audio) (AV-O)
3) Voices alone (A-P)
4) Objects (audio only) (A-O)
5) Faces alone (V-P)
6) Objects (visual only) (V-O)

Thus, categories 1 and 2 were audiovisual; 3 and 4 were audio only; and 5 and 6 were visual only. There were three different stimulus blocks within each type, each containing different visual/auditory/ audio-visual stimuli. A 16-second null event block comprising silence and a grey screen was also created. Each of the 18 blocks was repeated twice, and the blocks were presented pseudo-randomly: each block was always preceded and followed by a block from a different category (e.g. a block from the 'Faces alone' category could never be preceded/followed by any other block from the 'Faces alone' category). The null event block was repeated six times, and interspersed randomly within the presentations of the stimulus blocks.

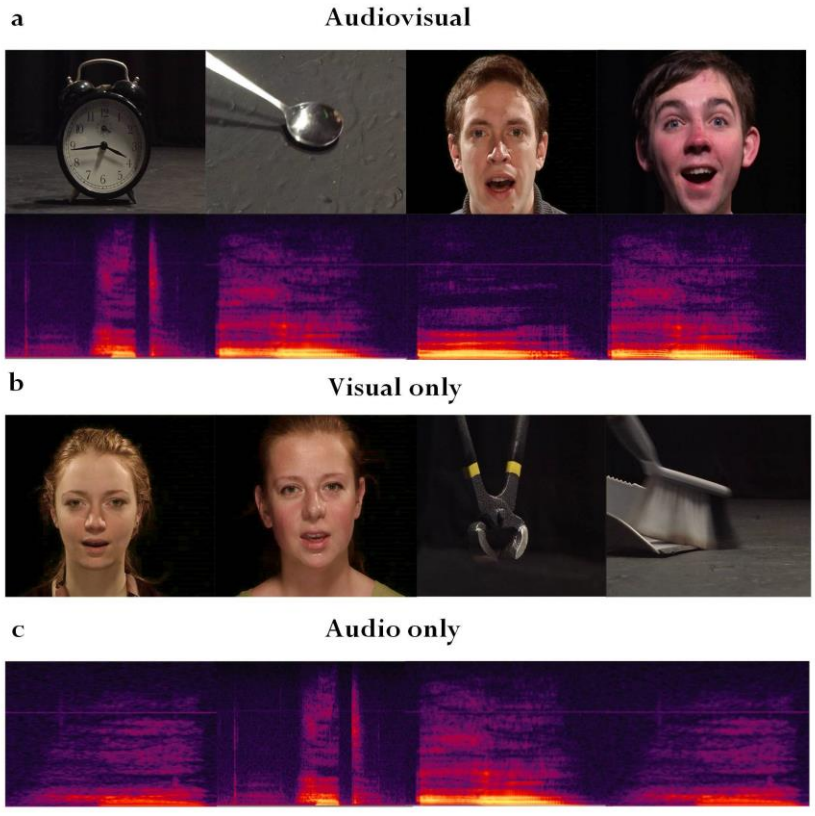


Figure 3.1. Examples of a) audiovisual, b) visual and c) auditory stimuli. Stimuli for the audiovisual localiser are available at *http://vnl.psy.gla.ac.uk/resources*

3.3.3 Design and Procedure

Stimuli were presented using the Psychtoolbox in Matlab, via electrostatic headphones (NordicNeuroLab, Norway) at a sound pressure level of 80 dB as measured using a Lutron Sl-4010 sound level meter. Before they were scanned, subjects were presented with sound samples to verify that the sound pressure level was comfortable and loud enough considering the scanner noise. Stimuli were presented in one scanning run while blood oxygenation-level dependent (BOLD) signal was measured in the fMRI scanner. Participants were not required to perform an active task; however, they were instructed to pay close attention to the stimuli.

3.3.4 Imaging parameters

Functional images covering the whole brain (slices=32, field of view=210x210 mm, voxel size=3x3x3 mm) were acquired on a 3T Tim Trio Scanner (Siemens) with a 12-channel head coil, using an echoplanar imaging (EPI) sequence (interleaved, TR=2 s, TE=30 ms, Flip Angle=80 degrees). We acquired 336 EPI volumes for the experiment. The first 4 s of the functional run consisted of 'dummy' gradient and radio frequency pulses to allow for steady state magnetisation during which no stimuli were presented and no fMRI data collected. MRI was performed at the Centre for Cognitive Neuroimaging (CCNi) in Glasgow, UK.

At the end of each fMRI session, high-resolution T1-weighted structural images were collected in 192 axial slices and isotropic voxels (1 mm^3; field of view: 256x256 mm, TR=1900 ms, TE = 2.92 ms, time to inversion = 900 ms, FA = 9 degrees).

3.3.5 Imaging analysis

SPM8 software (Wellcome Department of Imaging Neuroscience, London, UK; *http://www.fil.ion.ucl.ac.uk/spm*) was used to pre-process and analyse the imaging data. First the anatomical scan was AC-PC centred, and this correction applied to all EPI volumes.

Functional data were motion corrected using a spatial transformation which realigned all functional volumes to the first volume of the run and subsequently realigned the volumes to the mean volume. The anatomical scan was co-registered to the mean volume and segmented. The anatomical and functional images were then normalised to the Montreal Neurological Institute (MNI) template using the parameters issued from the segmentation keeping the voxel resolution of the original scans (1x1x1 and 3x3x3 respectively). Functional images were then smoothed with a Gaussian function (8x8x8 mm).

EPI time series were analysed using the general linear model as implemented in SPM8. Functional data were analysed in one two-level random-effects design. The first-level, fixed-effects individual participant analysis involved a design matrix containing a separate regressor for each block category (1-6). These regressors contained boxcar functions representing the onset and offset of stimulation blocks convolved with a canonical hemodynamic response function (HRF). To account for residual motion artefacts the realignment parameters were also added as nuisance covariates to the design matrix. Using the modified general linear model parameter estimates for each condition at each voxel were calculated and then used to create contrast images for each category relative to baseline: AV-P > baseline, AV-O > baseline, A-P > baseline, A-O > baseline, V-P > baseline, V-O > baseline . These six contrast images, from each participant, were taken forward into the second-level two factor (modality and category) ANOVA. The order of

conditions was: Audiovisual (Person); Audiovisual (Object); Audio only (Person); Audio only (Object); Visual only (Person); Visual only (Object).

Stimulus condition effects were tested with A(P+O) > baseline for sounds, V(P+O) > baseline for images and AV(P+O) > baseline for cross-modal sound-image. These contrasts were thresholded at $p<0.05$ (FWE peak-voxel corrected) with a minimum cluster size of 5 contiguous voxels.

The inclusion of non-face and non-vocal stimuli also allowed us to examine selectivity for faces and voices. We identified face-selective and voice selective regions, firstly with inclusion of audiovisual conditions (i.e., AV-P+V-P>AV-O+V-O for face selective, AV-P+A-P>AV-O+A-O for voice-selective), and then with only unimodal conditions included. These contrasts were thresholded at $p<0.05$ (FWE correction for cluster size) in conjunction with a peak-voxel threshold of $p<0.0001$ (uncorrected). In addition, we imposed a minimum cluster size of 10 contiguous voxels.

We then identified 'people-selective' regions as those who showed a 'person-preferring' response, regardless of the condition, whether this was audiovisual, audio only, or visual only (i.e., AV-P+A-P+V-P>AV-O+A-O+V-O). This contrast was thresholded at $p<0.05$ (FWE peak-voxel corrected) with a minimum cluster size of 10 contiguous voxels.

Conjunction analyses

We further performed a series of conjunction analyses in SPM8 in order to identify regions meeting a number of functional criteria:

a) Audio-visual Integration

We tested for general audiovisual, integrative regions with the conjunction analysis AV(P+O) > V(P+O) ∩ AV(P+O) > A(P+O) (i.e., the 'max rule' (Beauchamp, 2005)). This localised regions which showed a higher response to audiovisual stimuli as compared to both visual-only and audio-only stimuli.

We then tested for audiovisual regions which were also people selective (AV(P+O)>V(P+O) ∩ AV(P+O)>A(P+O) ∩ (AV-P+A-P+V-P>AV-O+A-O+V-O)).

b) Heteromodal response

We tested for regions that responded to both auditory and visual information (irrespective or their response to audiovisual stimuli) with the conjunction analysis A(P+O) ∩ V(P+O). It is important to note that alongside identifying heteromodal regions, integrative regions could also emerge from this criterion, as there was no criteria/requirement regarding the strength of the AV response.

We then tested for heteromodal regions that were also 'people selective' with the conjunction A(P+O) ∩ V(P+O) ∩ (AV-P+A-P+V-P>AV-O+A-O+V-O).

For all conjunction analyses, results were thresholded at $p<0.05$ (FWE peak-voxel corrected) with a cluster extent threshold of $k>5$.

3.4 Results

Regions activating more to auditory information (voices and object sounds) than the baseline condition were bilateral auditory cortex, right inferior frontal gyrus, and bilateral middle frontal gyrus (Table 3.1a). Regions activating more to visual information (silent faces and objects) than the baseline condition were the broad visual cortex, bilateral STG, left medial frontal gyrus, bilateral inferior frontal gyrus, right superior frontal gyrus, the posterior cingulate and the precuneus. (Table 3.1b). Regions activating more to

audiovisual persons and objects than baseline were bilateral visual and auditory cortex, bilateral inferior frontal gyrus and right medial frontal gyrus (Table 3.1c).

Brain regions	***Coordinates (mm)***			***k***	***t-statistic***
	x	***y***	***z***		
a) A > baseline					
Superior temporal gyrus (STG)	-48	-25	7	1846	20.76
STG	51	-22	4	2062	20.14
Inferior frontal gyrus (IFG)	39	17	25	112	6.22
Middle frontal gyrus (MFG)	-42	17	25	136	6.11
b) V > baseline					
Middle occipital gyrus (MOG)	45	-70	1	6135	24.21
IFG	42	11	28	650	9.30
Superior parietal lobule	30	-55	49	145	7.74
IFG	-39	11	22	272	7.74
IFG	30	32	-14	47	6.29
Superior frontal gyrus (SFG)	3	59	34	20	5.52
Medial frontal gyrus	-3	53	-14	27	5.50
Posterior cingulate gyrus	0	-52	16	22	5.43
Precuneus	-27	-55	49	15	4.96
c) AV > baseline					
MOG	45	-70	1	8670	22.65
IFG	42	14	25	608	10.38
IFG	-39	11	22	123	7.34
Precentral gyrus	-48	-1	49	48	5.82
Medial frontal gyrus	6	59	4	11	5.55
IFG	27	32	-11	19	5.35
IFG	-39	29	1	13	5.22
Superior parietal lobule	30	-55	49	11	5.03

Table 3.1 (previous page). Stimulus condition effects. Results of independently contrasting unimodal (a and b) and audiovisual (c) conditions against baseline.
Contrasts were height thresholded ($t = 4.51$) to display voxels reaching a significance level of $p < 0.05$ with FWE correction and an additional minimum cluster size of greater than 5 contiguous voxels. MNI coordinates and *t*-scores are from the peak voxel of a cluster.

Face-selective regions were found in the right STG and left MTG, the right middle frontal gyrus, precuneus and caudate. At a more liberal threshold, the right inferior frontal gyrus and right fusiform face area (FFA) emerged as face-selective regions (see Table 3.2a,b). Voice-selective regions were found in the bilateral STG/MTG, precuneus and right middle frontal gyrus (Table 3.2c,d).

Brain regions	*Coordinates (mm)*			*k*	*t-statistic*
	x	*y*	*z*		
a) Face-selective regions (including AV information)					
STG/	51	-34	1	867	13.98
Superior temporal sulcus (STS)					
MFG	51	2	46	735	9.05
Middle temporal gyrus (MTG)	-60	-16	-5	405	8.12
Precuneus	3	-58	31	249	7.72
Inferior occipital gyrus (IOG)	27	-97	-5	45	5.79*
b) Face-selective regions (excluding AV information)					
STG/STS	51	-37	4	820	10.51
MFG	51	-1	46	856	8.86
Precuneus	3	-58	28	197	5.62
STG/STS	-57	-40	7	171	4.88
Caudate	18	-4	16	184	4.56
IOG	42	-82	-11	72	5.38*
Fusiform gyrus (FG)	42	-46	-17	13	4.20*
c) Voice-selective regions (including AV information)					
STG/STS	51	-34	1	521	12.08
MTG	-60	-10	-8	295	9.25
Precuneus	3	-58	28	99	7.12
MFG	45	20	25	45	5.56
d) Voice-selective regions (excluding AV information)					
STG/STS	57	-19	-5	247	5.03
STG	-60	-10	-8	105	4.12
Precuneus	3	-58	28	33	3.69

Table 3.2 (previous page). Face and voice selective regions. Results of independently contrasting faces and voices against object images and non-vocal sounds (a,b and c,d respectively).
Contrasts were height thresholded ($t = 3.13$) to display voxels reaching a significance level of $p < 0.0001$ combined with an FWE correction of p<0.05 for cluster size. MNI coordinates and *t*-scores are from the peak voxel of a cluster. Starred contrasts were significant at a peak voxel threshold of p<0.0001 (uncorrected), with no cluster thresholding.

Regions which showed a greater response to people-specific information as compared to object-specific information (regardless of the modality) included the bilateral STG, bilateral inferior frontal gyrus, the right precuneus, and right hippocampus (Table 3.3a/Figure 3.2a).

Brain regions	*Coordinates (mm)*			*k*	*t-statistic*
	x	*y*	*z*		
a) 'People-selective' regions					
STG/STS	51	-34	1	710	15.01
STG	-60	-16	-5	324	9.25
IFG	42	20	25	406	8.85
Precuneus	3	-58	28	187	8.83
Hippocampus	21	-7	-14	25	6.39
IFG	-39	14	22	11	4.96

Table 3.3. People selective regions. Results of independently contrasting people-related information against object related information, regardless of condition.
Contrasts were height thresholded (t = 4.51) to display voxels reaching a significance level of p < 0.05 (FWE corrected for multiple comparisons). MNI coordinates and t-scores are from the peak voxel of a cluster.

3.4.1 Conjunction Analyses

a) Audiovisual, integrative regions

Audiovisual integrative regions (regardless of stimulus category), i.e., following the 'max rule' (AV(P+O)>A(P+O) ∩ AV(P+O)>V(P+O)) were found in the bilateral thalamus and bilateral STG/STS (Table 3.4a/Figure 3.2b). An integrative, people-selective region, i.e., following both the max rule and showing a greater response to people than object in both audition (voice>object) and vision (face>object) was observed in a localizer cluster of the right STG/pSTS (Table 3.4b/Figure 3.2c). This region can also be seen at the level of individual participants in Figure 3.3.

b) Heteromodal regions

Regions which responded to both visual and auditory information, as compared to baseline, consisted of the bilateral STG, and bilateral inferior frontal gyri (Table 3.4c/Figure 3.1d). Note that whereas the 'heteromodality' criterion does not make any assumption on what should be the response to the AV condition, a large part of the right pSTS also followed the 'max rule'. People-selective heteromodal regions, i.e., regions that responded significantly to both auditory and visual stimuli and that preferred social stimuli in both modalities, extended anteriorly to a large part of the STG/STS, and also activated the bilateral inferior frontal gyrus (Table 3.4d/Figure 3.2e). These regions can also be seen at the level of individual participants in Figure 3.3.

Brain regions	*Coordinates (mm)*			*k*	*t-statistic*
	x	*y*	*z*		
a) Integrative regions (max rule: AV>A ∩ AV>V)					
Thalamus	-15	-25	-5	21	7.04
STG/STS	60	-37	16	108	6.18
Thalamus	15	-25	-5	10	5.83
STG	-51	-46	13	14	5.36
b) People selective integrative regions					
STG/STS	51	-40	13	52	5.97
c) Heteromodal regions (A ∩ V)					
STG/STS	54	-40	13	575	11.10
STG/STS	-54	-46	13	183	8.51
IFG	39	17	25	109	6.15
IFG	-42	14	25	95	6.08
STG	36	2	-20	16	5.56
d) People selective heteromodal regions					
STG/STS	51	-40	10	325	10.50
IFG	39	17	25	108	6.22
IFG	-39	14	22	11	4.96

Table 3.4 (previous page). Results of conjunction analyses: a. Integrative audiovisual regions (AV > A ∩ AV > V); b. Integrative, people-selective regions; c. Heteromodal regions (Auditory > Baseline ∩ Visual > Baseline); d. Heteromodal, people-selective regions.

Contrasts were height thresholded ($t = 4.52$) to display voxels reaching a significance level of $p < 0.05$ with FWE correction and an additional minimum cluster size of greater than 5 contiguous voxels. MNI coordinates and *t*-scores are from the peak voxel of a cluster.

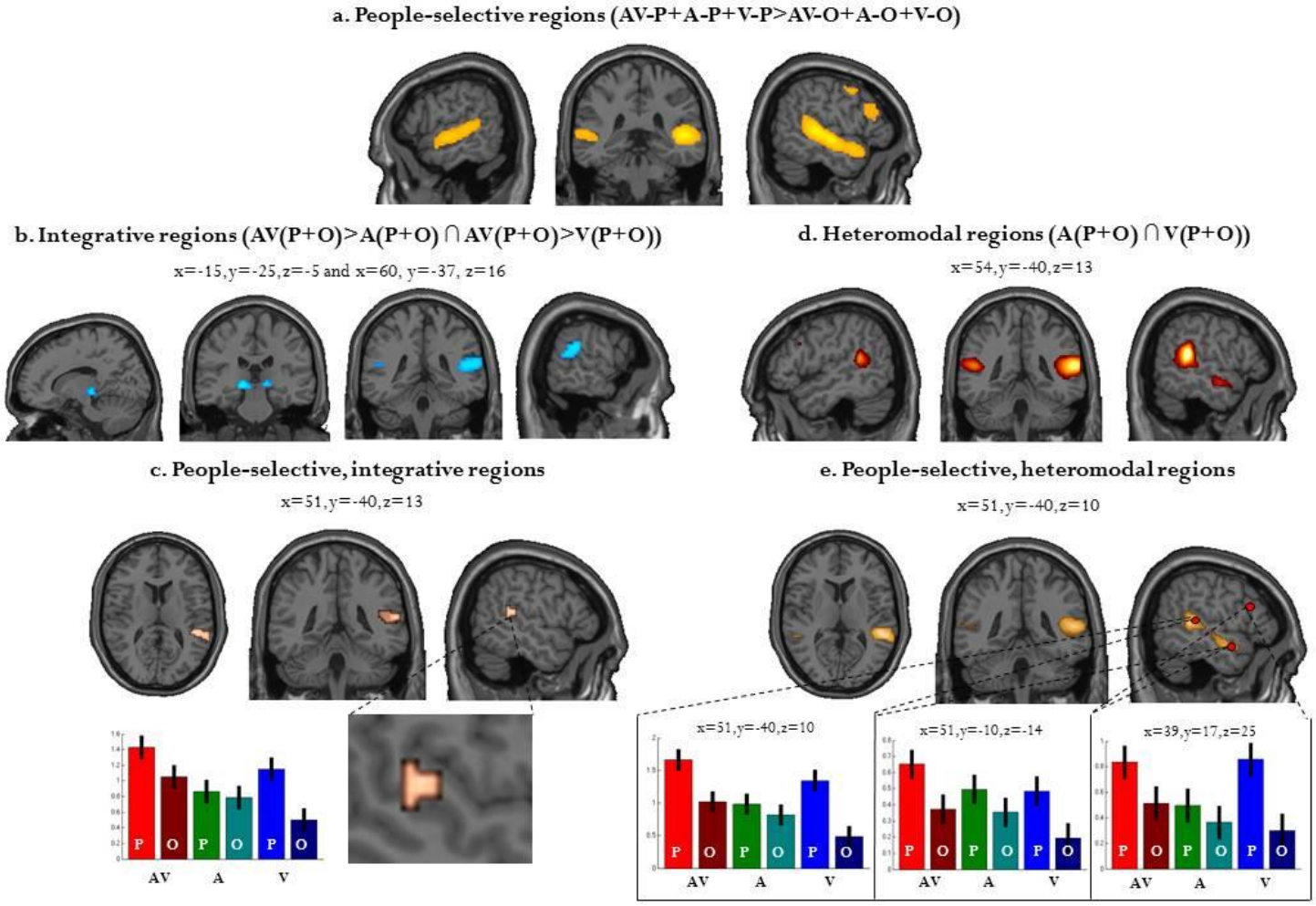

a. People-selective regions (AV-P+A-P+V-P>AV-O+A-O+V-O)
b. Integrative regions (AV(P+O)>A(P+O) ∩ AV(P+O)>V(P+O))
x=-15,y=-25,z=-5 and x=60, y=-37, z=16
d. Heteromodal regions (A(P+O) ∩ V(P+O))
x=54,y=-40,z=13
c. People-selective, integrative regions
x=51,y=-40,z=13
e. People-selective, heteromodal regions
x=51,y=-40,z=10
P O P O P O
AV A V
x=51,y=-40,z=10
x=51,y=-10,z=-14
x=39,y=17,z=25

Figure 3.2 (previous page). People-selectivity, audiovisual integration and heteromodality: a. 'People-selective' regions; b. Integrative audiovisual regions; c Conjunction of a and b: Integrative, people-selective regions; d. Heteromodal regions; e. Conjunction of a and d: Heteromodal, people-selective regions.

Contrasts were height thresholded ($t = 4.52$) to display voxels reaching a significance level of $p < 0.05$ with FWE correction and an additional minimum cluster size of greater than 5 contiguous voxels. MNI coordinates and *t*-scores are from the peak voxel of a cluster.

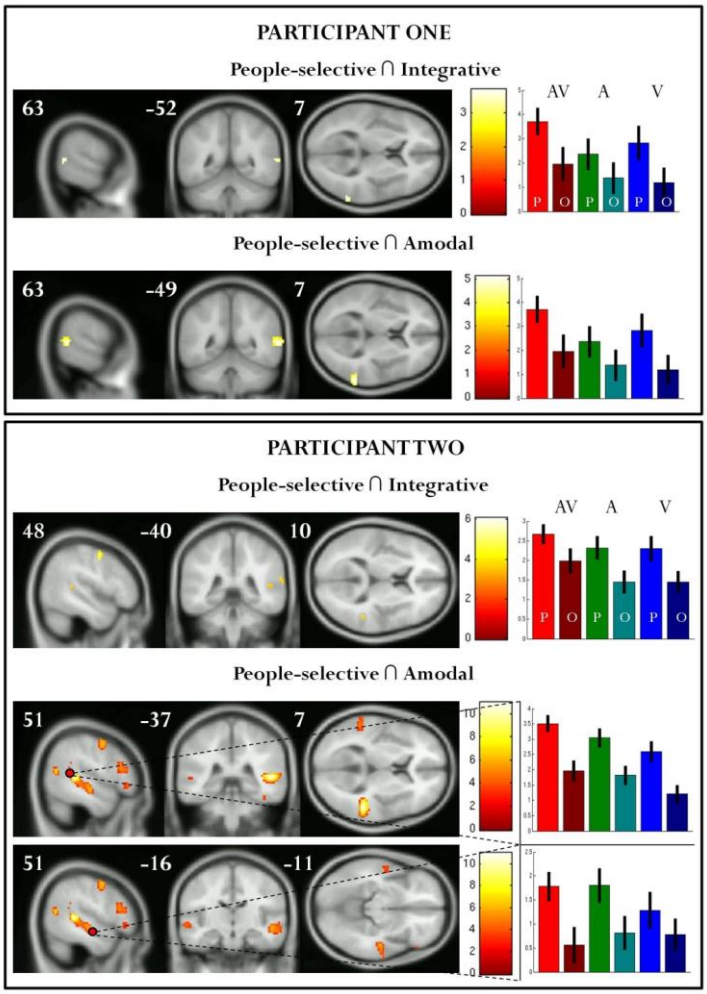


Figure 3.3. Results from individual participants: people-selective, integrative regions and people-selective, heteromodal regions.

For descriptive purposes, contrasts are height thresholded ($t = 3.12$) to display voxels reaching a significance level of $p < 0.001$ (uncorrected). MNI coordinates and *t*-scores are from global and local (Participant 2) maxima of STS cluster.

3.5 Discussion

The aim of this study was to examine the neural correlates of people-selectivity (i.e., regions that preferred face and voice information, regardless of condition), audiovisual integration (i.e. a significantly stronger response to audiovisual as compared to unimodal stimuli), and 'heteromodality' (i.e., a significant response to both vision and audition), specifically within the posterior superior temporal sulcus (STS). Participants were scanned during an 'audiovisual localiser' during which they passively viewed a series of audiovisual, visual and auditory stimuli of either people or objects; responses to each specific condition were compared and contrasted. Using a single dataset and ecological stimuli - dynamic movies of faces and voices - our results not only confirm the multisensory nature of the pSTS, but also that areas of this structure selectively processes person-related information irrespective of the sensory modality.

3.5.1 Face-selectivity, voice-selectivity and people-selectivity in the STS

We firstly examined voice- and face-selectivity in our participants by contrasting the response to voices as compared to non-vocal sounds, and faces as compared to visual representations of objects, respectively.

When we contrasted the response to auditory information against baseline, the broad auditory cortex was highlighted bilaterally. A voice-selective response was confined to the upper banks of the bilateral STS; regions that appear to correspond with the 'temporal voice areas' identified by Belin et al. (2000; 2004). Maximum voice-selectivity was found in the mid-STS, a result which has been found in a number of other studies (e.g., Belin et al., 2000; Belin et al., 2002; Kreifelts et al., 2009). The 'voice-selective' regions of the STS

tend to show a greater response to vocal sounds than to non-vocal sounds from natural sources, or acoustical controls such as scrambled voices or amplitude-modulated noise. This response is also observed for vocal sounds of non-linguistic content (Belin, et al., 2002; Belin et al., 2011), highlighting that these regions process more than just the speech content of voice. In a voice recognition study, Kriegstein and Giraud (2004) delineated three distinct areas along the right STS involved in different aspects of voice-processing: whereas the mid-anterior STS carries out a spectral analysis of voices, more posterior and anterior areas emphasise more paralinguistic voice processing – for example, identity. We also identified the right precuneus as a voice-selective region in this experiment. Although perhaps less commonly found than the TVA, activation of the precuneus has been apparent in a number of studies investigating voice perception (e.g. Von Kriegstein et al., 2003; Sokhi et al., 2005).

The visual vs. baseline contrast showed activation maps covering most of the visual ventral stream, including early visual cortex (V1:3), V4, V5/MT, the fusiform and parahipocampal gyri and an extensive part of the human inferior temporal (IT) gyrus. This is consistent with the vast majority of research studying visual responsiveness. Face-selectivity was found in a network of regions, including the extensive right STS, left pSTS to mid-STS, the middle frontal gyrus, precuneus and caudate – all regions which have been associated with either the core or extended face-processing system (e.g. Haxby et al., 2000; Rossion et al., 2003; Andrews et al., 2010). Notably, at the set-threshold for the group-level analysis, the commonly found fusiform face areas (FFA) did not emerge, although these regions – along with the bilateral occipital face areas (OFA) - did appear for a number of individual participants, as well as at the group level at an uncorrected cluster threshold. Instead, the strongest response appeared to be in the STG/STS – particularly, the right pSTS. In our experiment, we used only dynamic faces, in an attempt to maximise the ecological validity

of our stimuli. The pSTS is known to be involved in the representation of the dynamic properties of faces (Allison et al., 1999; Haxby et al., 2000; Haxby et al., 2002) such as mouth, eye and head movements (Lee et al., 2010) and facial expressions (Phillips et al., 1997): although it does respond to pictures of static faces (Kanwisher et al., 1997; Hoffman & Haxby, 2000), it shows a response of significantly greater magnitude (up to three times) to dynamic as compared to static faces (Pitcher et al., 2011). Thus, it could be that continuously presenting only moving faces heightened the response in the pSTS and attenuated the response in the FFA, as previously proposed by Hoffman and Haxby (2000).

We further generalized this approach to all conditions and identified 'people-selective' regions in our group of participants as those that responded selectively to social stimuli in all conditions, whether this was audiovisual, audio only or visual only. Such regions were found in bilaterally in the pSTS to mid-STS, in addition to the right aSTS, the inferior frontal gyrus, hippocampus and precuneus. In a pioneering study, Kreifelts et al. (2009) examined voice-selectivity, face-selectivity and integration of affective information within the STS. They found, using fMRI, that the neural representations of the audiovisual integration of non-verbal emotional signals, voice sensitivity and face sensitivity were located in different parts of the STS with maximum voice sensitivity in the trunk section and maximum face sensitivity in the posterior terminal ascending branch. These authors did not observe the large overlap as was seen in our study, and we can only speculate as to some of the possible reasons. We predict the large response of the STG was in part due to contrasting dynamic audiovisual presentations of people against audiovisual presentations of objects, plus unimodal face and voice information – thus, these would have activated the portions of the STG/STS responsive to audiovisual information, in addition to those responsive to dynamic face information and voice-selective regions. In the study by Kreifelts, face and voice selectivity were examined using separate localisers, which simply

contrasted the response to different sets of unimodal stimuli. What is more, in their face-localiser, the authors only used static faces. Although static faces can also activate the STS (Haxby et al., 2000; Kanwisher et al., 1997) dynamic faces are known to evoke a more pronounced response in this region.

In summary, we find that in this experiment, a large part of the STS - extending from pSTS to aSTS - was 'people selective' in all modalities: this is striking, considering that previous research has localised face-selectivity and voice-selectivity in different, mostly non overlapping portions of this regions, specifically the pSTS and mid-STS to aSTS, respectively.

3.5.2 Face-voice integration and the STS

We used a conjunction analysis and the classical 'max criterion' to define integrative, audiovisual regions in our study. This analysis highlighted the bilateral thalami and the bilateral pSTS as regions responding more to audiovisual information as compared to both visual information and audio information alone.

Both the thalamus and the pSTS are well described as playing a role in multimodal processing. There is now converging evidence that not only sensory non-specific, but also sensory specific, thalamic nuclei may integrate different sensory stimuli and further influence cortical multisensory processing by means of thalamo-cortical feed-forward connections. Some studies provide evidence of thalamic influence on multisensory information processes in rats (Komura et al., 2005) and humans (Baier et al., 2006) and others link modulations of neuronal activity in subcortical structures with behavioural consequences like audiovisual speech processing (Bushara et al., 2001) and multisensory attention tasks (Vohn et al., 2007). Kreifelts et al. (2007) also reported in humans an

enhanced classification accuracy of audiovisual emotional stimuli (relative to unimodal presentation) and linked this increase in perceptual performance to enhanced fMRI-signals in multisensory convergence zones, including the thalamus.

The upper bank of the superior temporal sulcus has also emerged as a crucial integrative area, particular the pSTS. This region is known to have bidirectional connections with unisensory auditory and visual cortices (Cusick, 1997; Padberg et al., 2003) and to contain around 23% of multisensory neurons (Barraclough et al., 2005). Ghazanfar et al. (2005) showed that the STS was involved in speech processing when monkeys observed dynamic faces and voices of other monkeys. Consistent with findings from animals, the human pSTS also becomes active when processing audiovisual speech information (Calvert et al., 2001), in addition to presentations of tools and their corresponding sounds (Beauchamp et al., 2004), letters and speech sounds (van Atteveldt et al., 2004), and faces and voices (Beauchamp et al., 2004; reviewed in Hein and Knight, 2008). Recently – and also using the max criterion – Szycik et al. (2008) found the bilateral STS to be involved in face–voice integration. Crucially, this was observed using markedly different stimuli to ours - firstly, they presented a static face in their unimodal condition and secondly, they added white noise to their auditory and audiovisual stimuli. The fact that the activation of this region is preserved across stimulus types and sets underlines its importance in the integration of faces and voices. Previously, the hippocampus has also been implicated as key region in the integration of face and voice information (Joassin et al., 2011b). At the set threshold, this region did not emerge: however, as in a recent study by Love et al. (2011), the left hippocampus did emerge at less conservative, uncorrected significance level. This lends further support to the importance of this region; albeit, in a more minor role within this context.

Our conjunction of people-selective and integrative responses highlighted a cluster in the right pSTS, which was more responsive to people-related information – whether this was faces and voices, faces only or voices only. In addition, this region showed a significant preference for audiovisual information, as compared to both audio only and visual only information. Interestingly, this analysis removed the activation previously seen in the thalamus and the left pSTS, suggesting that these regions may be either more 'general' – or even, 'object-selective' – integrative regions. The right pSTS has been found in previous studies examining audiovisual integration (e.g., Werner and Noppeny, 2011; Hagan et al., 2009; Ethofer et al., 2006; Kreifelts et al., 2010; also reviewed in Calvert, 2001) but crucially, these have generally compared audiovisual to unimodal responses within independent stimulus sets, without contrasting activation to different stimulus categories. To our knowledge, this is the first study that directly looks at person-selectivity of audiovisual integrative regions and we therefore propose that the right pSTS could have a crucial role in combining 'socially-relevant' information across modalities.

3.5.3 'Heteromodality' and the STS

Further, we examined responses across modalities: 'heteromodal' regions were defined as those that simply responded significantly to both audio and visual information as compared to baseline, irrespective of what their response to the AV condition was. Thus, along with potentially highlighting regions which integrated face and voice-information (i.e., showed a significantly stronger response to audiovisual information), this criteria was also able to identify regions which responded to both faces and voices, but did not necessarily integrate this information. This analysis isolated regions in the right pSTS to mid-STS, left pSTS, bilateral inferior frontal gyrus and putamen. The bilateral pSTS proved to be an audiovisual, integrative region, overlapping with the regions found in our previous analysis. However, activation continuing down the trunk region of the STS appeared to be

genuinely heteromodal: the response to audiovisual information that was not significantly more than either audio or visual presentation, but the auditory and visual responses to the unimodal stimuli were significantly greater than baseline.

When we looked specifically at people-selective portions of these regions, activation followed the line of the posterior to mid STS. The peak of activation, in the pSTS, again overlapped with people-selective integrative regions. Kreifelts et al. (2010) also observed a sensitivity to voices as well as faces in the right pSTS, which they suggest might be conceived as an essential characteristic of the neural structures subserving the audiovisual integration of human communicative signals. However, they also point out that in their study, given the differences in control stimuli for the separate voice and face-sensitivity experiments, they refrain from any direct comparisons between the two qualities.

In the mid-STS, there was a stronger response to faces and voices together, as opposed to faces and voices alone – however, this difference was not significant. Outwith the STS, in the middle frontal gyrus, there was an equal response to both face-voice combinations and voices alone, but a lesser response to faces alone. Interestingly, this 'heteromodal' analysis highlighted a multitude of regions that did not emerge using our integrative criterion. We propose that the 'heteromodality' criterion, which does not make any assumption on what the response to combined stimuli should be but simply requires a response in both modalities, could act as an interesting complement to the typical analyses used when defining audiovisual regions, especially as some of these defining statistical criteria are recognised as being particularly stringent (Beauchamp et al., 2005; Love et al., 2011).

3.5.4 People-selectivity and the right hemisphere

In our study we found a strong right-hemispheric response to people-selective information. Although we found an initial people-selective response in both right and left hemispheres, conjunction analyses show lateralised integrative and heteromodal effects in the right hemisphere, specifically the right pSTS to mid-STS, and not in the left hemisphere. Given previous findings on face- and voice-selectivity, this dominance is perhaps unsurprising.

Although studies on face perception have reported face-selective regions in the fusiform gyri of both the left and right cerebral hemispheres, fusiform activations for faces are often found to be greater in the right than in the left (De Renzi et al., 1994; Kanwisher et al., 1997; McGarthy et al., 1997; Le Grand et al., 2003), and previous psychophysical investigations with split brain patients also suggest lateral asymmetry in face processing and encoding (Gazzaniga and Smylie, 1983; Miller et al., 2002). In a recent study (Ming Meng et al., 2012), the authors found that face-selectivity persisted in the right hemisphere even after activity on the left had returned to baseline.

Similarly, studies which have examined voice-selectivity – although smaller in number – also suggest a preference of the right hemisphere. For example, in Belin et al. (2000), the authors observed that averaged in a group of subjects, voice-sensitive activity appeared stronger in the right hemisphere. It appears this asymmetry may be particularly specific to the non-linguistic aspects of voices. In one functional magnetic resonance imaging (fMRI) study (von Kriegstein et al., 2003), it was shown that a task targeting on the speaker's voice (in comparison to a task focussing on verbal content) lead to a response in the right anterior temporal sulcus of the listener. In further study by Belin et al. (2002), it was shown that temporal lobe areas in both hemispheres responded more strongly to human voices than to other sounds (e.g., bells, dog barks, machine sounds) but that, again, it was

the right aSTS that responded significantly stronger to non-speech vocalisations than to scrambled versions of the same stimuli. In our experiment, we found bilateral face and voice-selective responses – however, for both of these effects the strongest activation was in the right hemisphere. Given the fact that the linguistic content of our stimuli were kept to a minimum, and that participants passively viewed and heard the visual and auditory information, this right dominance could possibly be expected.

We further identified both integrative and heteromodal regions bilaterally, in the STS and the thalamus (for the former analysis only). However, it was only in the right hemispheres that these effects showed a heightened preference for face and voice information. This extends on the multitude of research that suggests that there is right hemispheric functional asymmetry in response to social information. Indeed, the right hemisphere shows a preference for not only faces and voices, both also other socially relevant information such as biological human motion and sex pheromones. For all of these functions, stronger involvement of the right hemisphere in coding some aspects of person perception seems to be the rule, whereas involvement of the left hemisphere appears to sometimes be a shared role, and only exceptionally a main role. The reason to why this 'social asymmetry' exists in the first place still remains an open question, although there have been a number of possibilities postulated (see Brancucci et al., 2008). Additionally, whether the right hemisphere also prefers to integrate these other types of 'people-selective' information will only be answered with further investigation.

3.6 Conclusion

Our results build on previous research suggesting that the STS is a 'social-information processing' region, by clearly delineating 'people-selective' regions that respond discerningly to both face and voice information, across modalities. Furthermore, this study also provides the first evidence of a *'people-selective' integrative* region in the right pSTS. Future directions could involve exploring selectivity for other types of socially-relevant information in the STS, inter-individual variability of STS functionality, and further investigating the nature of neuronal populations in 'people-selective' STS regions.

4. Audiovisual integration of gender from the face and voice: a behavioural investigation

4.1 Abstract

Both the face and the voice provide us with not only linguistic information, but also a wealth of paralinguistic information, including gender cues. However, the way in which we integrate these two sources in our perception of gender has remained largely unexplored. In the following study, we used a bimodal perception paradigm in which varying degrees of incongruence were created between facial and vocal information within audiovisual stimuli. We found that in general, participants were able to combine both sources of information, with the perception of gender reflecting a contribution of information from both modalities. However, this combination was not symmetrical: in this experiment voice appeared to exert a stronger influence on gender perception. This finding was supported by results from conditions that directed attention to either modality: participants were unable to ignore the gender of the voice, even when instructed to, but were able to ignore the face. The dominance of vocal information in this experiment is discussed with respect to task and stimulus selection.

4.2 Introduction

In addition to communicating linguistic information, both faces and voices provide a rich source of information regarding a person's biological characteristics, including gender and unique identity. The ability to not only recognise this information in each sensory

modality, but also integrate these into a unified percept is a crucial part of social interaction. However, despite our natural, bimodal perception of paralinguistic information such as this, the overwhelming amount of literature on identity and gender recognition has concentrated on unimodal face and voice cues.

Although integration of some information is reasonably well-researched (e.g. visual and audio temporal and spatial cues), less is known of how we combine more socially-relevant, bimodal signals (namely, faces and voices). Perhaps the most researched area in the field of audiovisual person perception has been face-voice speech perception, most famously demonstrated in the 'McGurk effect' (McGurk and MacDonald, 1976). Furthermore, evidence suggests that audiovisual integration in speech perception can occur early in time, (van Wassenhove et al., 2005), thus indicating that such integration could possibly be mandatory, free from any voluntary control (Green et al., 1991).

Only a handful of studies to date have focussed on how facial and vocal non-speech information is combined: and in particular, within this area, the research on audiovisual gender perception is scarce. However, evidence from the few studies that exist on this topic suggests that gender information in the face and voice interacts in a similar way to other paralinguistic information, such as identity (e.g. Schweinberger et al., 2007; Kamachi et al., 2003; Sheffert and Olsen, 2004) and emotion (e.g. Massaro & Egan, 1996; de Gelder & Vroomen, 2000; Ethofer, 2006).

In a pioneering study, Smith et al. (2007) showed that auditory and visual information interacts during face gender processing. In their experiment, participants were instructed to categorise androgynous faces according to their gender, while the faces were coupled with pure tones in the male or female fundamental-speaking-frequency range. They found that

faces were judged as more male when coupled with a pure tone in the male fundamental frequency range, and vice versa, showing that auditory information does indeed interact with facial cues in gender categorisation. Facilitation effects have also been observed in the case of face-voice gender associations, with congruent bimodal audiovisual stimuli resulting in faster classification of gender, as compared to presentations of face and voice alone (Joassin et al., 2011).

Latinus et al. (2010) expanded on this research by investigating crossmodal interactions in gender categorisation. Subjects performed three gender judgement tasks: in the first, they judged if the gender of a static face and voice were congruent or incongruent; and in the second and third, they categorised the bimodal stimuli by gender, in one case attending only to voices and in the other only to faces. The directed attention aspects of the task allowed the authors to determine the influence of top-down modulation on multimodal processing (i.e., effects due only to the task), whereas the use of congruent and incongruent stimuli provided information on bottom-up stimulus-dependent processing. They found that an incongruent face disrupted the processing of voice gender (indicated by significantly lower categorisation 'hits'), suggesting an automatic integration of the two inputs in this task. However, an incongruent voice had a lesser, non-significant effect on the perception of face gender – as compared to an incongruent face on voice gender - suggesting that in their experiment, vision dominated over audition in terms of overall gender categorisation.

These studies have established the foundation for further studies in this area. However, all the aforementioned studies used static portraits of faces coupled with recordings of voices. Although integration effects have undoubtedly been observed with such relatively crude stimuli, this is always going to provide a somewhat unrealistic experience for the

participant: in everyday life we almost constantly see a dynamic presentation of audio and visual information, synchronised in time. This is the case in nearly all our social interactions, where we perceive others as dynamic and multimodal stimuli, with only a handful of unimodal examples such as speaking over the telephone where only the information from the voice is available. Articulatory movements of the face are especially related to speech perception, due to physical changes in the face occurring during vocal production (Munhall et al., 2006).

With regards to person perception, studies by Schweinberger et al. (2007) and Kamachi et al. (2003) have both shown that articulating faces cause differential effects as compared to static. Kamachi et al. (2003) found that unfamiliar face-voice matching in their experiment only occurred when the faces were moving, highlighting the importance of articulatory movements for integration of identity cues. In Schweinberger et al. (2007), benefits from corresponding faces were significantly larger for dynamic faces—which were synchronized with a familiar voice—than for the static faces. Although corresponding static faces also caused a significant facilitation in reaction times relative to baseline, it was significantly smaller than the one caused by dynamic faces. Furthermore, non-corresponding static faces did not cause any significant costs in familiar-voice recognition performance. By contrast, non-corresponding dynamic faces did cause substantial costs in performance, indicating that participants were not able to ignore faces as soon as they were presented in time synchrony with the acoustic stimulus. These results suggest that use of dynamic, multimodal stimuli may allow us to observe audiovisual integration effects which otherwise might not be seen.

The aim of the following experiment was to provide a fuller account of the audiovisual interactions that occur during gender processing of faces and voices. We created a unique

set of stimuli by parametrically morphing both faces and voices between genders, allowing us to create more subtle variations of incongruence within audiovisual stimuli. We also examined the effect of attention on the perception of bimodal face-voice stimuli, using three gender judgement tasks. In the first task, participants were not directed to attend to any modality. In the second two tasks, attention was directed to either face or voice. The same stimuli were used in the three tasks, allowing us to determine effects due only to the task. This paradigm allowed us to further investigate sensory dominance and its influence on gender categorisation. Additionally, we utilised both dynamic and static face stimuli: by doing so we were able not only to provide a more ecological approach to the face-voice integration process, but also directly compare responses to articulating and static faces.

We hypothesised that gender categorisation would differ between incongruent and congruent face-voice pairings, and that this difference would parametrically increase in accordance with the degree of incongruence. In accordance with previous literature, we also assumed that fully incongruent audiovisual stimuli would take longer to categorise than congruent 'end point' audiovisual stimuli (i.e., female face-female voice; male face-male voice) and that reaction times would lengthen as facial and vocal information became increasingly incongruent. Finally, we hypothesised that if a modality dominance existed, gender information from the dominant modality would interfere with gender categorisation, even when participants were instructed to ignore this modality; and that additionally, any categorisation shifts or costs (i.e. heightened reaction times) caused by face-voice incongruence would be smaller for the less dominant modality. Finally, we hypothesised, in line with findings by Schweinberger et al. (2007), that observed costs for incongruent stimuli and facilitation for congruent stimuli might be larger for dynamic stimuli, as compared to static stimuli.

4.3 Materials and Methods

4.3.1 Participants

Twenty one English speaking participants (3 non-native speakers; 12 females; all right handed; mean age= 22 years) participated in the study. All had self-reported normal or corrected vision and hearing. The study was approved by the ethical committee from the University of Glasgow. All volunteers provided informed written consent and received payment at the rate of £6 per hour for participation.

4.3.2 Stimuli

Video recording and editing

10 actors (5 males and 5 females, selected to match in age) were video-recorded saying the word 'had' multiple times. Actors were shown a template video of someone uttering the word "had", and were instructed to match their vocalisation as much as possible to the duration of the template video (~1 second), in order to minimise variation in length for our sets of vocalisations. All participants were native speakers of the English language. The males were clean-shaven, and the females wore no make-up. None had any distinctive facial markings or piercings. This ensured that morphs of the faces would not contain any cues which related to the gender of either individual. Recordings took place in the television studio at the Learning and Teaching Centre of the University of Glasgow, and actors were paid at the rate of £6 per hour. The actors were filmed under standard studio lighting conditions (standard tungsten light) and against a black background, and sat 235cm away from the camera, directly facing it. Videos were recorded with 25 frames per second (40ms per frame) using a Panasonic DVC Pro AJD 610 camera, fitted with a Fujiform A17 x 7.8 BERM-M28 lens, and transferred and edited using Adobe Premiere

Elements. Sound recording was captured directly from the video camera's microphone, and was recorded with 16-bit resolution at a sampling frequency of 44100 Hz. Videos were edited so that every pronunciation of the words by all male and female formed a separate clip. One clip from each volunteer was selected for use. Each of the clips was then separated into their visual and audio components.

Face morphing

In all clips, seven important temporal landmarks that best characterised the facial movements related to the vocal production were determined, and the frames at which they occurred were identified. These landmarks were the first movement of the chin, first opening of lips, maximum opening of the mouth, first movement of the lips inwards, time point at which the teeth met, closing of the lips, and the last movement of the chin. The theoretical average frames for these landmarks were then calculated, and the videos edited so the occurrence of these landmarks matched in all clips. Editing consisted of inserting or deleting video frames during fairly motionless periods. Due to the speakers pronouncing the word with standardised timing, little editing was necessary. The editing produced ten adjusted clips, each 36 frames (1440 ms) long. These were then used to create 'composite' male and female face frames (i.e. an average of the 5 female faces and 5 male faces, respectively). We reasoned that averaging would allow us to create gender-specific faces closer to a 'prototype', or a central average. The morphing software 'Psychomorph' (Tiddeman and Perrett, 2001) was used to generate the average morphs in each of the 36 frames for both the average male and average female face. The morphing software 'Videomorph' (pilot software created by Bernard Tiddeman) was then used to create a morphed continuum for each frame, which extended from 90% (average) female to 90% (average) male, in 10% steps. This provided 9 different face gender stimuli for each frame. Each of these morphed faces was therefore a ten-face composite (5 male faces, 5 female)

with a different weighting of the average male and female faces. All images were all converted to greyscale, matched for luminance, and an oval mask fitted around each face so to conceal potential gender cues such as the hair. New videos were then created using these masked frames. Corresponding gender morphs for each frame were edited together in Adobe Premiere (e.g. 10% female for Frame 1, 2, 3 and so forth). In order to create the static, control videos, we first identified the frame in which the mouth had a maximal aperture. In each of the videos, this frame (18th) was then selected and lengthened to last 36 frames.

Auditory morphing

Auditory stimuli were edited using Adobe Audition 2.0 (Adobe Systems Incorporated, San Jose, California). Stimuli were firstly RMS normalised in Adobe Audition (REF). In order to generate the auditory components to the "morphed-videos" a similar procedure was used. The voices were initially edited according to the theoretical 'average' frames generated by the video morphing procedure. While video editing consisted of inserting or deleting video frames around the identified landmarks, audio editing consisted of inserting or deleting equivalent lengths of vocalisation at these time points. This ensured that all our audio samples were an identical duration to one another, and to our videos. The morphing procedure was performed using STRAIGHT (Kawahara & Matsui, 2003) in Matlab (The MathWorks, Inc., Natick, MA). STRAIGHT performs an instantaneous pitch-adaptive spectral smoothing in each stimulus to separate the contributions of the glottal source (including F0) versus supralaryngeal filtering (distribution of spectral peaks, including the first formant, F1) to the voice signal. Voice stimuli are decomposed by STRAIGHT into 5 parameters: fundamental frequency (F0), formant frequencies, duration, spectrotemporal density, and aperiodicity; each parameter can be independently manipulated. Anchor points, that is, time–frequency landmarks, were identified in each sound on the basis of

landmarks easily recognizable on each spectrogram. Temporal anchors were beginning of the production, beginning and end of the voicing of "HA(-d)", as well as the plosive "(ha-)D" and the end of the production. Frequency anchors were first, second, and third formants at onset of phonation, onset of formant shift, that is the points where each formant lowered in amplitude and at the end of phonation.

An average female voice was then generated by resynthesis based on a logarithmic interpolation of the female voices temporal and frequency anchor templates to the 50% average of the female voices. The same procedure was used to generate a male average voice. Then, a morph continuum between the two average voices was generated using a resynthesis based on a logarithmic interpolation of the average female and average male anchor templates in steps of 10%. This resulted in 9 different voice stimuli, extending from 90% female to 90% male. The different weighting of the average male and female voices was equivalent to the weightings used for morphing the faces.

Audiovisual video production

162 audiovisual videos were produced by pairing static and dynamic face videos with the morphed voices (every pairwise combination of the 18 faces videos (nine face morphs, static and dynamic) with each one of the nine voices) covering the whole space of audio-visual face-voice gender allowed by our independent visual and auditory morphing procedure. Altogether, this provided a variety of congruent and incongruent stimuli. The audiovisual videos were then cut from the 10th frame to the 30th frame. This was in order to largely remove periods of a static face at the beginning and end of the clip, where the lips were closed. The videos started at the frame before movement of the lips occurred. It should be noted that in our original videos, the onset of the faces preceded the onset of the audio speech. Indeed, the first facial movements typically precede vocalisation in natural

utterances. Therefore, the onset of visual articulation did not correspond with the first frame of speech production. Instead, the vocalisation (defined by the first burst of the 'a' of 'had') began approximately 120 ms after visual onset, and 80 ms after the first movement of the lips. This auditory delay was matched in the static videos. After this final editing stage, videos lasted 800 ms.

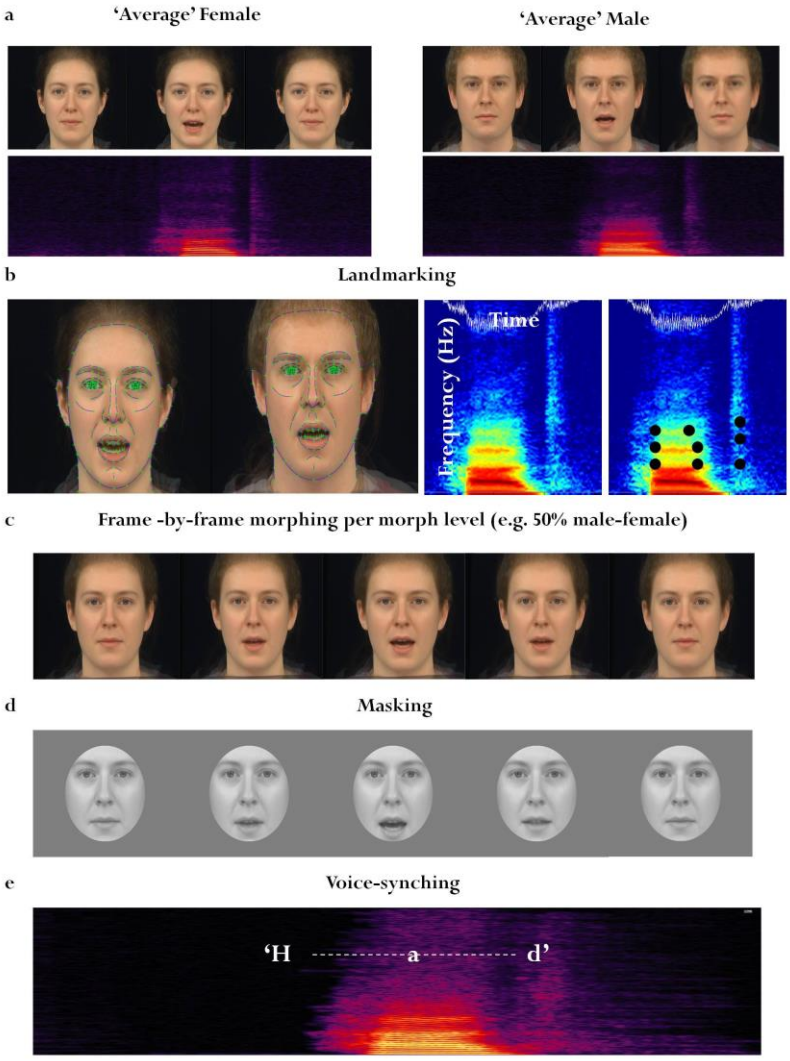


Figure 4.1. Making of audiovisual stimuli. a. 'Average' male and female videos (each composed of averaging videos from five individuals); b. Spatial landmarking of faces; and temporal/frequency landmarking of voices; c. A morphed video for a 50% male-female face; d. Face converted to greyscale and masked; e. Time-synchronised voice paired with masked video.

4.3.3 Design and Procedure

All videos were presented at 720 x 576 pixels, using Matlab 2007b and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions running on a PC. The auditory stimuli were presented in mono, via Beyerdynamic DT 770 headphones at approximately 70 dB as measured using a Lutron Sl-4010 sound level meter. Participants saw and heard all stimuli in a soundproof booth. Instructions were given to the participants before each condition. All participants undertook all three of the following tasks:

1. Audiovisual (uncontrolled attention)

Participants were instructed to watch the screen and listen to the presented voices, and asked to indicate their gender decision via a two choice button press. Participants were not instructed to attend to any modality in particular, but rather to simply pay attention to both the face and the voice. Before the experiment began a fixation cross appeared on the screen for 2 seconds. Each Audio-Visual (AV; 162 in total) stimulus was repeated 10 times during the course of the experiment. The experiment was divided in 10 blocks distributed over 2 sessions (5 blocks in each). In each block, the 162 AV stimuli were presented in a randomised trial order. Breaks were given between each block. Of these 162 AV stimuli, there were 18 completely congruent stimuli. If the participant indicated their response during the movie presentation, the next movie was presented one second after the end of that movie presentation. If the participant indicated their response after the movie presentation, the next movie was played one second after their response.

2. Audiovisual (attend to face)

In this condition the same stimuli were used as in condition 1, but this time, the participants were instructed to focus their attention on the faces to classify the gender of

the stimulus and ignore the gender of the voices. A randomised order was used again, and timings were the same as in condition one.

3. Audiovisual (attend to voice)

Again, the same stimuli were used as in condition 1 and 2; but the participants were instructed to focus their attention on the voices to classify the gender of the stimulus and to ignore the gender of the faces. They were also explicitly instructed not to close their eyes when presented with the stimulus.

The order of tasks 2 and 3 were counterbalanced between participants. This was in order to remove any possible effects of always directing to one modality first, and then the other (e.g. remaining attention bias of the previous task).

Average gender classification ratings and reaction times – for each participant and for each stimulus - were calculated and submitted to the following analyses. It should be noted that, in all analyses, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (epsilon < 0.75).

4.4 Results

We initially submitted categorisation and reaction time data to two separate ANOVAs, each with Movement (dynamic or static), Attention (Uncontrolled attention, Attention to Face, Attention to Voice), Voice (voice morph 1 (90% female) – 9 (90% male)) and Face (face morph 1 (90% female) – 9 (90% male)) as within-subjects factors (2x3x9x9 ANOVA). For gender categorisation, the effect of Attention, Voice and Face were all significant ($F(1.29, 23.1)=4.713$, p,0.04; $F(2.39, 43.0)=242$, $p<0.0001$; $F(1.36,$

24.49)=77.6, p<0.0001). There were also significant interactions between Attention and Face (F(2.23, 40.1)=66.6, p<0.0001), Attention and Voice (F(2.67, 48.1)=84.6, p<0.0001) and Voice and Face (12.03,216.59)=2.54, p<0.005). However the effect of Movement was not significant (F(1,18)=0.003, p=0.959). There were also no significant interactions between Movement and any of the other three factors. Regarding reaction time data, the effect of all factors were significant (Movement: F(1,20)=28.7, p<0.0001; Attention: F(1.34, 26.8)=4.60, p<0.04; Voice: F(2.55, 51.1)=29.1, p<0.0001; Face: F(2.32, 46.5)=4.14, p<0.02). There were also significant interactions between Attention and Movement (F(1.57, 31.4)=11.6, p<0.0001), Attention and Voice (F(4.91, 98.1)=13.1, p<0.0001, Attention and Face (F(3.46, 69.3)=5.80, p<0.002, and Voice and Face (F(11.6, 231)=3.35, p<0.0001).

We further examined categorisation and reaction time data for each condition. The dynamic and static categorisation data points were averaged, as there was no overall significant effect of Movement for these set of results.

4.4.1 Audiovisual condition – uncontrolled attention

Firstly we compared categorisation ratings obtained in the different AV conditions when subjects were not instructed to attend to a particular modality. Figure 4.2a shows a 3-D plot of the average ratings for the 9x9 morph steps in the audiovisual condition. Here it can be seen that although both face and voice morph caused shifts in categorisation ratings (indicated by change in colour) these changes were not symmetrical between the two modes – voice shows a stronger visible effect. Data was submitted to an ANOVA with Face (1-9) and Voice (1-9) as within-subject factors. The main effect of voice was significant (F(1.80,35.9) = 126, p<0.0001), as well as that of the Face (F(1.11,22.1) = 8.23, p=0.007), indicating that both face and voice gender affected overall gender ratings. The

Voice x Face interaction was also significant ($F(9.4,188) = 2.27$, $p=0.018$), demonstrating that the effect of one modality depended on values in the other modality, and that these effects were not purely additive across the modalities, but rather interacted. The effect of voice was larger overall – indicated by a greater main effect - highlighting that subjects were on average weighting the auditory modality more when making the gender judgment (Figure 4.2b). Figure 4.2c suggests, however, that not all subjects showed this effect; indeed two individuals weighted the face modality more than the voice, and three participants presented an entirely balanced strategy.

Additionally, in a series of planned comparisons, we examined at which points there were significant differences in categorisation ratings between stimuli. We earlier suggested that maximum incongruence between Face and Voice (i.e. 80% difference) would cause significant shifts in categorisation, as compared to 'end point' congruent stimuli. In order to test this we therefore compared categorisation values between these maximally incongruent and congruent stimuli in the following paired sample t-tests:

i) 90% female voice-90% female face vs. 90% female voice-10% female face

ii) 90% female voice-10% female face vs. 10% female voice-10% female face

iii) 10% female voice-10% female face vs. 10% female voice-90% female face

iv) 10% female voice-90% female face vs. 90% female voice-90% female face

After a Bonferroni correction for multiple comparisons, comparisons i), ii), and iv) remained significant ($t(20)=3.03$, $p=0.007$; $t(20)=9.57$, $p<0.0001$; $t(20)=-12.7$, $p<0.0001$ respectively). Comparison iii) was not significant at either the corrected p ($p<0.01$) nor at the standard $p<0.05$ level ($t(20)=-1.95$, $p=0.065$).

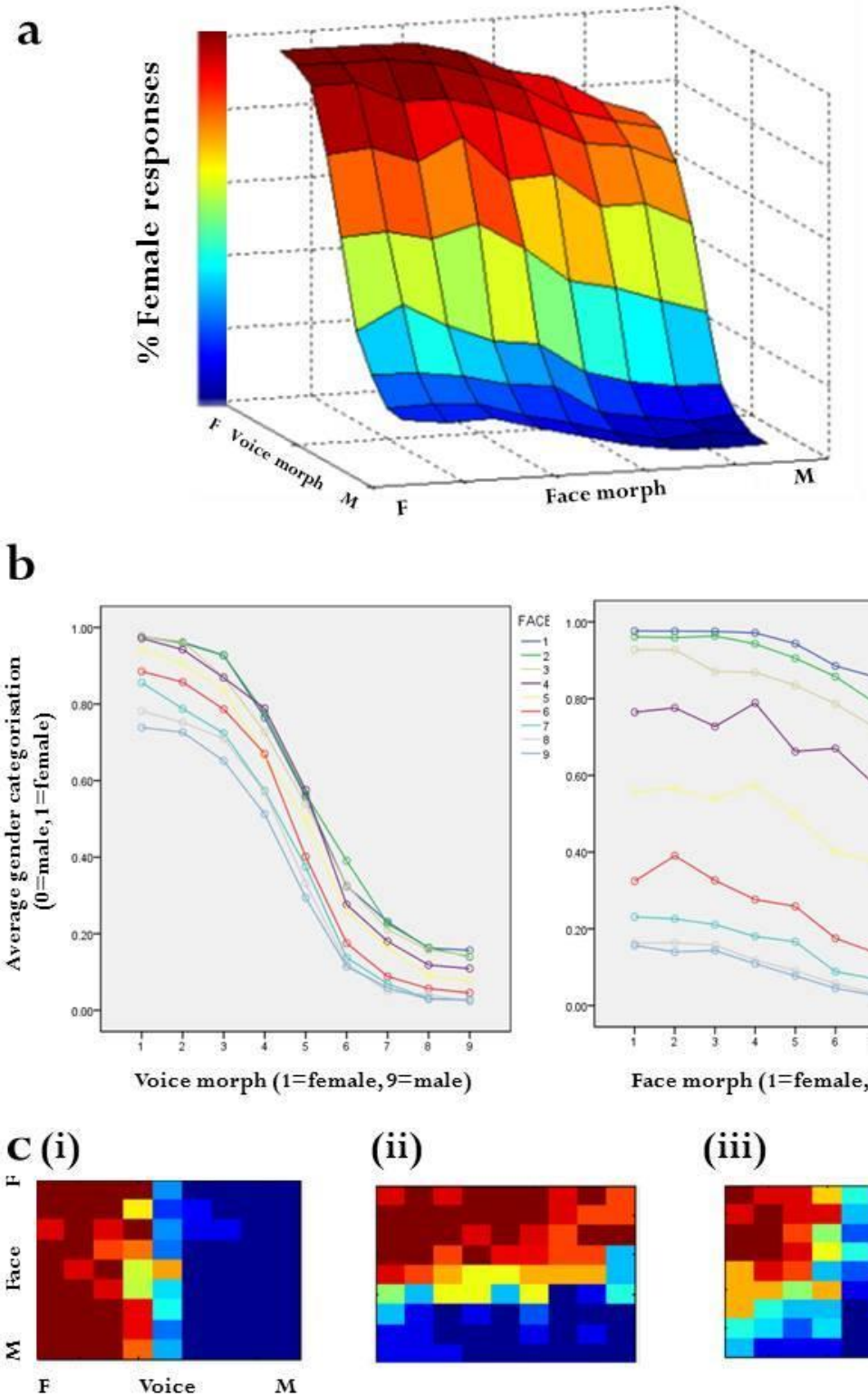

a
% Female responses
F
Voice morph
M
F
Face morph
M
b
Average gender categorisation
(0=male,1=female)
FACE
VOICE
Voice morph (1=female, 9=male)
Face morph (1=female, 9=male)
c (i)
(ii)
(iii)
F
Face
M
F
Voice
M

Figure 4.2 (previous page). Results from uncontrolled attention task (average categorisation ratings).
a. 3D plot of average categorisation responses across 21 participants. Average gender categorisation: Dark red = 100% female; Dark blue = 100% male;
b. 2D plots of average categorisation responses across the same set of participants. Face/Voice morph: 1 = 90% female information, 9 = 90% male information; Average gender categorisation: 0 = male, 1 = female; Colour scale: Navy blue = 90% female; Green = 80% female; Beige = 70% female; Purple = 60% female, Yellow = 50% female; Red = 40% female; Cyan = 30% female; Grey = 20% female; Blue = 10% female.
c. Individual participant categorisation strategies. i. Voice dominant (highlighted by little graduation in colour for Face stimuli, against majority of Voice stimuli); ii. Face dominant (highlighted by little graduation in colour for Voice stimuli, against majority of Face stimuli); iii. Balanced integration of both modalities (highlighted by graduation in colour for both Voice and Face stimuli).

Secondly, we compared reaction times obtained in the audiovisual condition where subjects were free to attend to any modality. Data was submitted to two ANOVAs (dynamic and static) with Face (1-9) and Voice (1-9) as within-subject factors. In the dynamic face ANOVA, the main effect of Voice was significant ($F(3.31,66.3) = 15.1$, $p<0.0001$), but not that of Face ($F(3.81,76.3) = 0.516$, $p=0.715$). However, the Voice x Face interaction was significant ($F(11.4,228) =1.89$, $p<0.0001$). Similarly, in the static face ANOVA, the main effect of Voice was significant ($F(2.80, 56.1)=17.1$, $p<10\text{-}4$), but not that of Face ($F(2.99, 59.8)$, $p=0.606$). The Voice x Face interaction was also significant ($F(12.8, 255)$, $p=0.04$). Plots of reaction time data for this condition can be seen in Figure 4.3. For both static and dynamic conditions, reaction time largely appeared to be a function of voice morph, with generally little change in reaction time as the face pairing moved from congruent to incongruent.

For the reaction times of both dynamic and static faces, we compared the 'end point' congruent and maximally incongruent stimuli in a number of planned comparisons to test

our hypothesis that incongruence between face and voice would result in categorisation costs (i.e. longer reaction times). Reaction times for male and female congruent stimuli were averaged for both dynamic and static face stimuli, as were reaction times for the maximally incongruent stimuli, as we observed no significant interaction with movement (Dynamic faces: 90% female face-90% female voice vs. 90% male face-90% male voice: $t(20)=-1.73$, $p=0.098$; 90% female voice-10% female face vs. 10% female voice-90% female face: $t(20)=-1.27$, $p=0.218$; Static faces: 90% female face-90% female voice vs. 90% male face-90% male voice: $t(20)=0.454$, $p=0.655$; 90% female voice-10% female face vs. 10% female voice-90% female face: $t(20)=-0.625$, $p=0.539$). We then performed a paired sample t-test of congruent vs. incongruent reaction times, for both dynamic and static faces. For dynamic face stimuli, this difference was significant ($t(20)=-2.65$, $p<0.02$) but for static face stimuli this difference was not ($t(20)=-1.95$, $p=0.066$).

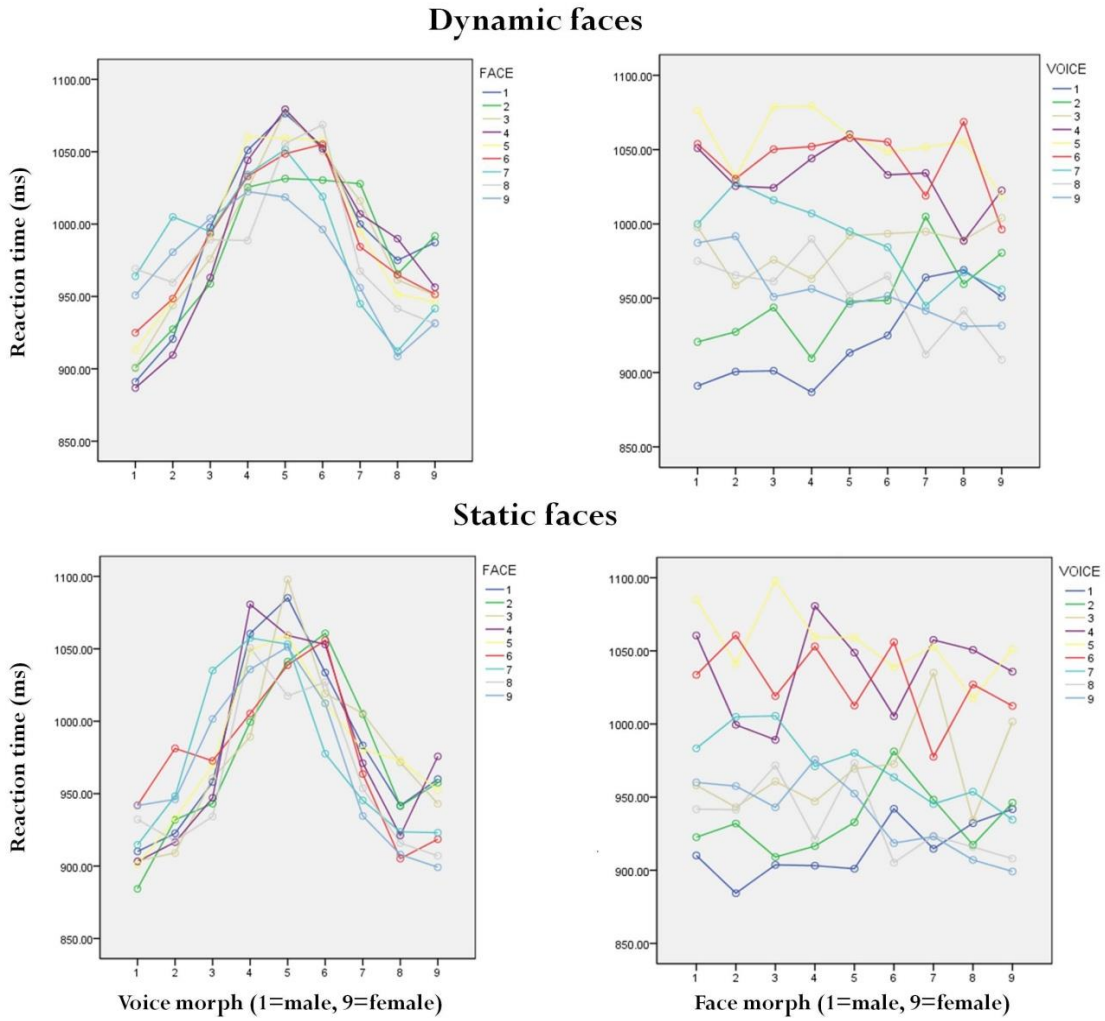


Figure 4.3. Results from uncontrolled attention task (average reaction times). Top left and right = Dynamic face information in audiovisual stimuli; Bottom left and right = Static face information in audiovisual stimuli. Face/Voice morph: 1 = 90% female information, 9 = 90% male information; Average gender categorisation: 0 = male, 1 = female. Colour scale: Navy blue = 90% female; Green = 80% female; Beige = 70% female; Purple = 60% female, Yellow = 50% female; Red = 40% female; Cyan = 30% female; Grey = 20% female; Blue = 10% female.

4.4.2 Audiovisual condition - attention to voice

Here participants were presented with a face-voice stimulus, but instructed to rate gender based only upon the voice. Categorisation data was submitted to an ANOVA with Face (1-9) and Voice (1-9) as within subject factors. The main effect of voice was significant

(F(1.86,35.4) = 295, p<0.0001) as expected, indicating adequate categorisation of the voice gender continuum. However, there was no significant effect of face gender, indicating a lack of influence of the visual modality on gender perception when attention was directed to the voice (F(2.10,39.9) = 2.81, p=0.07). This can be observed in Figure 4.4: the little visible difference between the curves as a function of Voice morph, and the lack of slope of the curves as a function of Face morph indicate the non-significant effect of face information. Additionally, there was no significant interaction between factors.

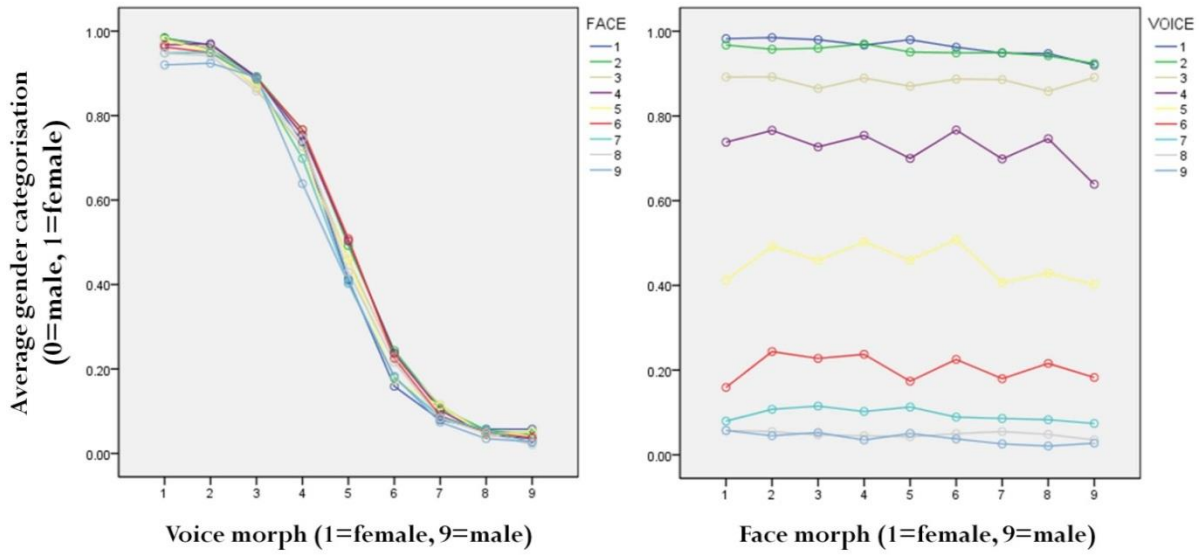


Figure 4.4. Results from attention to voice task (average categorisation ratings). Face/Voice morph: 1 = 90% female information, 9 = 90% male information; Average gender categorisation: 0 = male, 1 = female. Colour scale: Navy blue = 90% female; Green = 80% female; Beige = 70% female; Purple = 60% female, Yellow = 50% female; Red = 40% female; Cyan = 30% female; Grey = 20% female; Blue = 10% female.

Reaction time data was submitted to two ANOVAs (dynamic and static) with Face (1-9) and Voice (1-9) as within-subject factors. In the dynamic face ANOVA, the main effect of Voice was significant (F(3.28,65.6) = 24.3, p<0.0001), but not that of Face (F(5.15,103) = 0.996, p=0.425). The Voice x Face interaction was not significant (F(11.8,236) =1.05,

p=0.408). Again, in the static face ANOVA, the main effect of voice was significant (F(2.56,51.3) = 22.2, p<0.0001), but not that of Face (F(5.75,115) = 1.33, p=0.229). The Voice x Face interaction was not significant (F(11.8,236) =0.891, p=0.556). Results are shown in Figure 4.5.

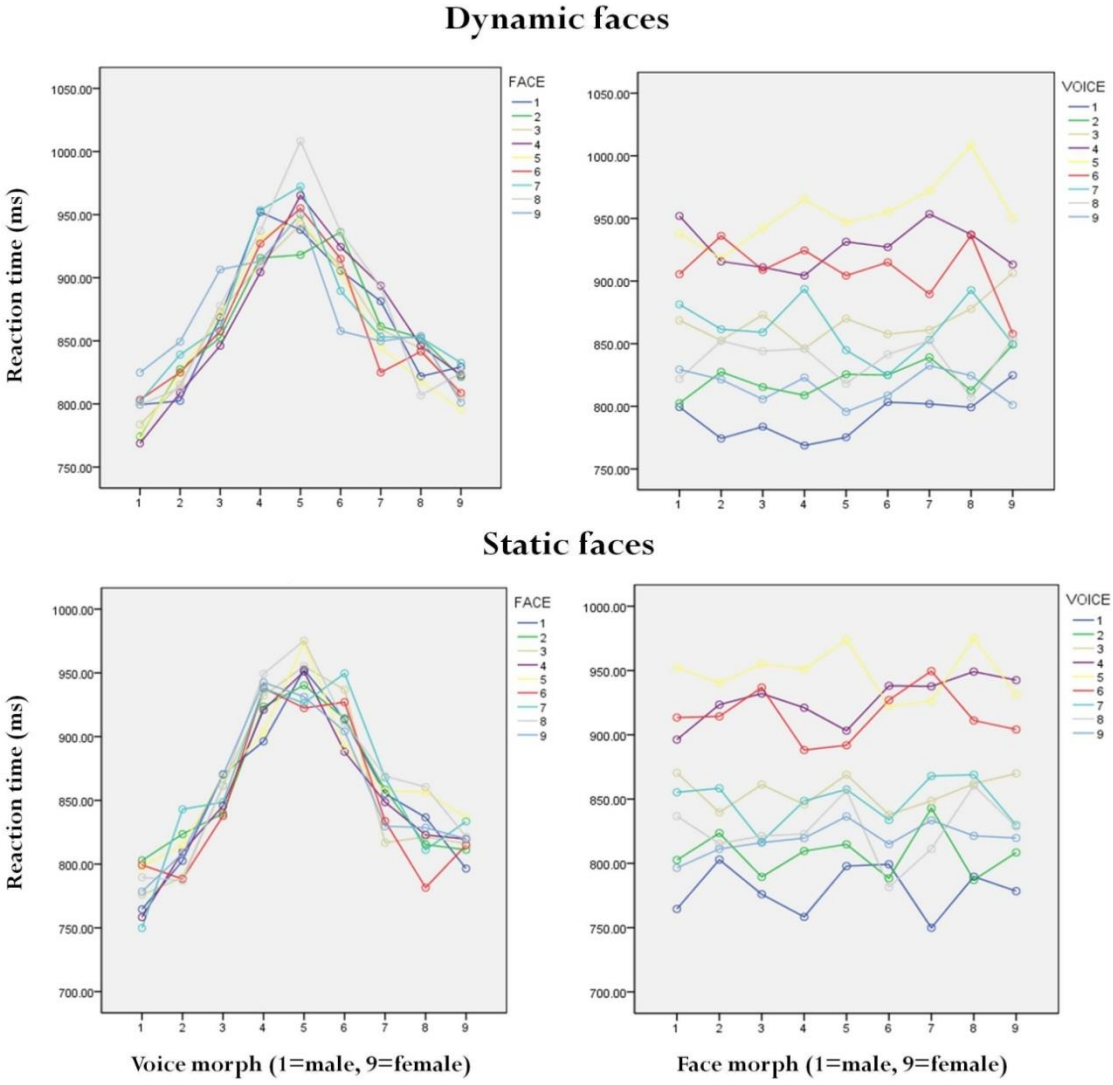


Figure 4.5. Results from attention to voice task (average reaction times): Top left and right = Dynamic face information; Bottom left and right = Static face information. Face/Voice morph: 1 = 90% female information, 9 = 90% male information; Average gender categorisation: 0 = male, 1 = female. Colour scale: Navy blue = 90% female; Green = 80% female; Beige = 70% female; Purple = 60% female, Yellow = 50% female; Red = 40% female; Cyan = 30% female; Grey = 20% female; Blue = 10% female.

4.4.3 Audiovisual condition - attention to face

As in the previous condition, participants' attention was directed to one modality; but in this case, they were instructed to focus on the face. Categorisation data was submitted to an ANOVA with Face (1-9) and Voice (1-9) as within subject factors. The effect of Face was, as expected, highly significant ($F(2.11,42.3) = 205$, $p<0.0001$). However, the effect of voice was also significant ($F(1.97,39.3) = 16.6$, $p<0.0001$), indicating a strong influence of the voice gender on face gender categorisation even under instructions to ignore the voice. The Voice x Face interaction was also significant ($F(12.6,252) =1.88$, $p=0.034$). This indicates that voice had a differential effect on categorisation ratings at various points along the face morph continuum. This influence of the voice can be seen in Figure 4.6: particularly, its notable effect on perceived face gender in the central, androgynous portion of the face continuum (red and yellow curves).

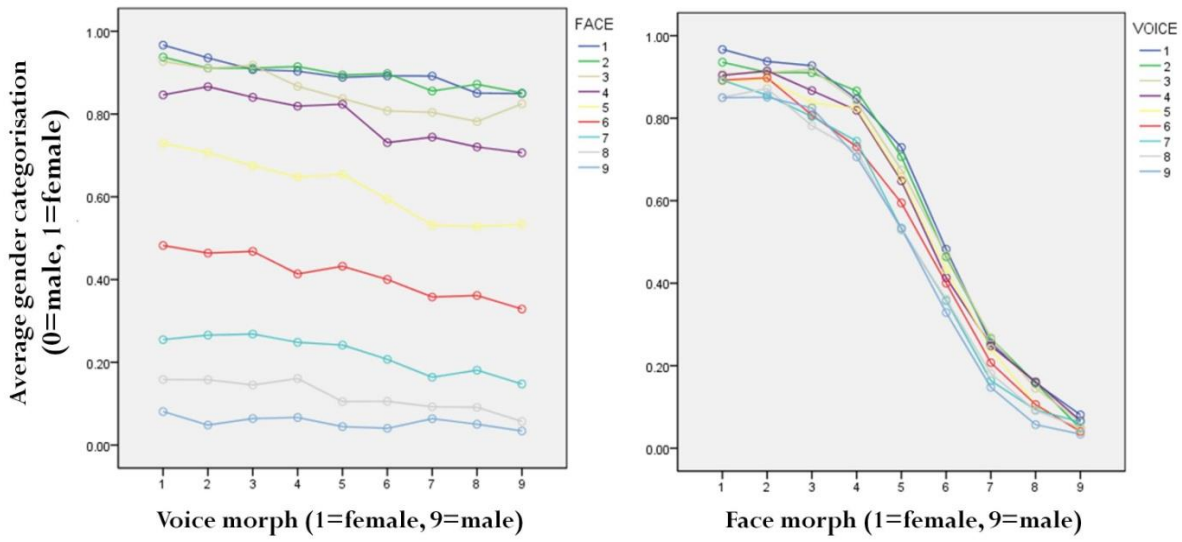


Figure 4.6. Results from attention to face task (average categorisation ratings). Face/Voice morph: 1 = 90% female information, 9 = 90% male information; Average gender categorisation: 0 = male, 1 = female. Colour scale: Navy blue = 90% female; Green = 80% female; Beige = 70% female; Purple = 60% female, Yellow = 50% female; Red = 40% female; Cyan = 30% female; Grey = 20% female; Blue = 10% female.

Reaction time data was submitted to two ANOVAs (dynamic and static) with Face (1-9) and Voice (1-9) as within-subject factors. As shown in Figure 4.7, in the dynamic face ANOVA, the main effect of Face was significant ($F(2.23,44.6) = 6.45$, $p=0.003$), but not that of Voice ($F(4.17,83.3) = 1.69$, $p=0.157$). However, the Voice x Face interaction was significant ($F(13.0,259) =2.28$, $p=0.007$), indicating that at some points in the continuum, an incongruent voice caused larger costs in reaction times than others. In the static face ANOVA, similar results were observed: the main effect of Face was significant ($F(2.46,41.8) = 4.95$, $p=0.008$), but not that of Voice ($F(5.16,87.7) = 2.00$, $p=0.08$); and the Voice x Face interaction was significant ($F(10.9,185) =1.89$, $p=0.044$.

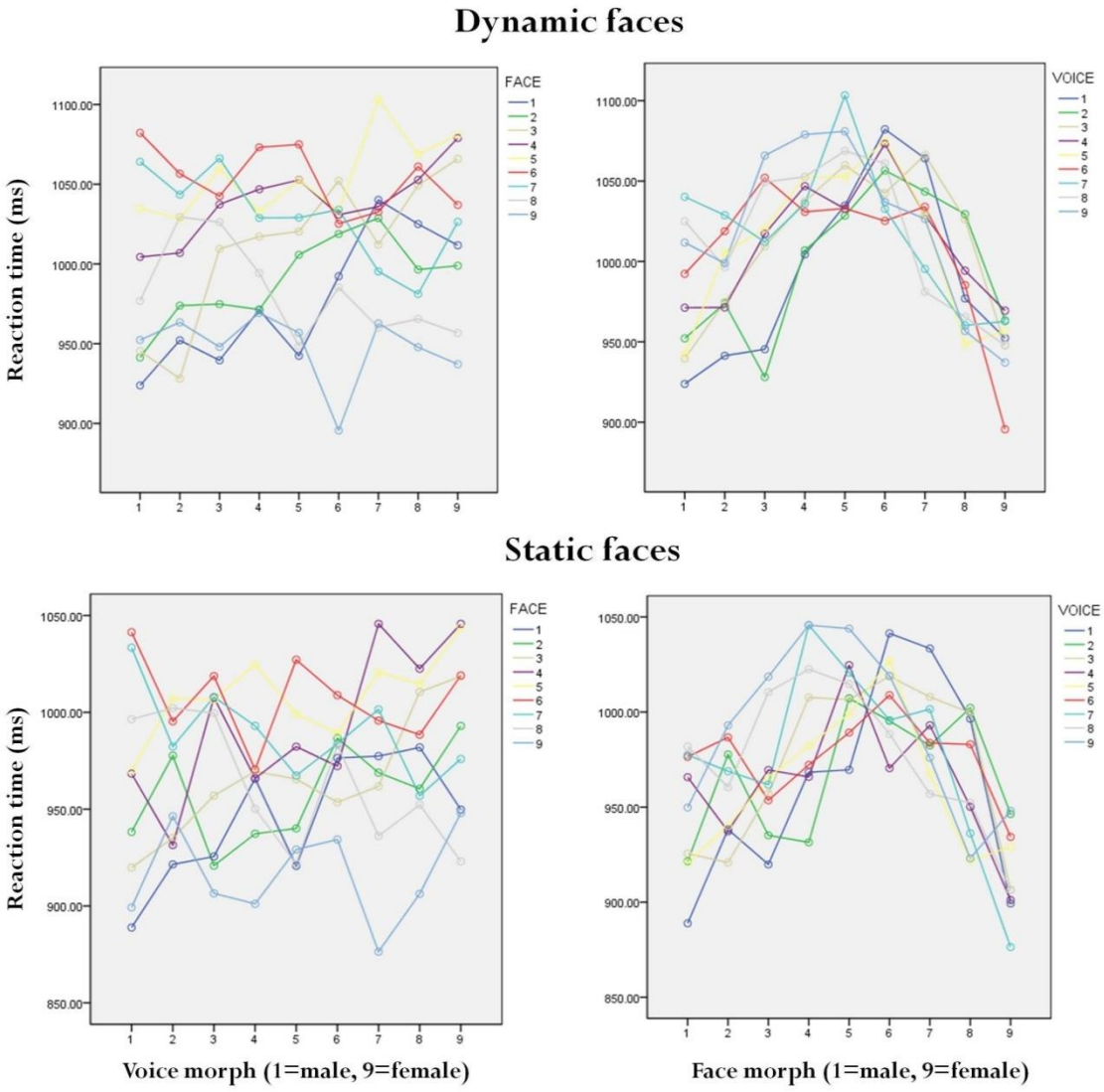


Figure 4.7. Results from attention to face task (average reaction times): Top left and right = Dynamic face information; Bottom left and right = Static face information. Face/Voice morph: 1 = 90% female information, 9 = 90% male information; Average gender categorisation: 0 = male, 1 = female. Colour scale: Navy blue = 90% female; Green = 80% female; Beige = 70% female; Purple = 60% female, Yellow = 50% female; Red = 40% female; Cyan = 30% female; Grey = 20% female; Blue = 10% female.

4.5 Discussion

The primary aim of the present study was to explore the combination of information from facial and vocal cues in the recognition of gender. Using state-of-the-art visual and auditory morphing technologies, we created a parametric space of gender stimuli consisting of dynamic and synchronous faces and voices. This was achieved via the independent parametric manipulation of gender. Overall, the experiment showed that both face and voice gender influenced overall gender ratings. However, the effect of voice gender in this experiment was stronger than that of face gender. This was confirmed by the results of the audiovisual conditions which controlled for attention: attending to voices resulted in the previous influence of face on gender categorisation disappearing, whereas attending to face still showed an influence of both modalities. Reaction time data also highlighted the strong effect of voice: in the audiovisual condition where attention was not directed to either modality in particular, there was a main effect of voice morph but not that of face. Reaction times appeared to follow the voice-morph level, with lower reaction times when the voice was unambiguous (90% male/female) and higher at the middle of the morph continuum.

4.5.1 Dynamic vs. static face information

In our experiment, we included both dynamic and static face stimuli. This was in order provide a more ecologically valid approach to the study of audiovisual gender integration, and also to directly investigate whether dynamic and static faces led to differential effects on participants' processing of face-voice gender information. Although arguably the dynamic quality of faces might be more important for a task such as processing of speech - an inherently dynamic process - results from behavioural studies investigating person

recognition (Kamachi et al., 2003; Schweinberger et al., 2007) suggested that articulating faces could elicit stronger audiovisual effects (e.g. more marked costs and gains of congruence and incongruence). Thus, we believed it was plausible that we might find differing effects due to articulation in the perception of gender.

Our results showed that participants' categorisation of gender was not dependent on whether the video contained an articulating or static face, in any of our conditions; however, movement information did significantly affect reaction times. The non-significant effect on categorisation could perhaps be expected, as both faces were offering exactly the same gender information (i.e., degree of gender morph). Further investigation of reaction times in our uncontrolled attention task showed that, although overall the main effect of face was not significant, between end-point congruent and maximally incongruent stimuli there was a significant difference in reaction times, but only for the dynamic stimuli. This is consistent with the aforementioned studies of Kamachi et al. (2003) and Schweinberger et al. (2007) who observed differential effects for dynamic and static faces. In this study we suggest there were two possible reasons for this effect: firstly, dynamic faces could simply offer *more* gender information; secondly, dynamic faces arguably attract more attention, and therefore could act as a stronger strategic gender 'cue'. If a dynamic face was unable to be ignored as easily, this could possibly account for the significant difference in reaction times between dynamic and static stimuli.

4.5.2 Role of attention: uncontrolled vs. directing to a modality

We then firstly compared ratings and reaction times within the audiovisual (uncontrolled attention) condition only. This analysis involved comparisons between all pairings of all face and voice morphs. We observed a main effect of both face and voice for gender categorisation, indicating that participants could combine data from the two sources to

arrive at a unique judgement on gender. However, the main effect of voice was greater, indicating that participants, on average, used vocal information more when categorising gender. The significant interaction between face and voice gender highlighted that our parametric shifts in gender, for both face and voice, exerted larger shifts in categorisation between certain points in the 3 dimensional audiovisual gender space, as compared to others. For example, incongruent facial information resulted in more pronounced effects at the 'female' and 'androgynous' regions of the voice continuum, as compared to when, for example, a female face was paired with a male voice. One reason for this could be that our cropped faces tended to look more male, perhaps in part because they were missing facial contour information that is a strong cue to gender.

With regards to reaction time data, for both dynamic and static stimuli the main effect of voice was significant but not that of face. Reaction times appeared to mainly be a function of voice morph, with androgynous morphs resulting in longer reaction times than morphs at either end point of the continuum, regardless of the face that was paired with the voice. However, as mentioned previously, maximal incongruence did cause significant costs for dynamic face stimuli.

We then investigated whether directing attention to a particular modality altered the previously observed integration patterns. Our reason for doing so was based on results of previous studies (de Gelder and Vroomen, 2000; Vroomen et al., 2001; Latinus et al., 2010), which found that bimodal information was processed regardless of whether it was required for the task performance – or indeed, explicitly instructed to ignore - suggesting an automaticity in face and voice processing, with a mandatory integration of inputs at some unconscious level. In our attention conditions, participants were still presented with audiovisual stimuli, but were instructed to ignore either the face or the voice, and make

their judgements purely on the basis of what they heard or saw in the other modality. We firstly examined gender categorisation, and found that participants were able to ignore the face when instructed to do so, indicated by no significant main effect of face, which had been observed in the audiovisual condition with uncontrolled attention. However, in contrast, participants were unable to ignore the vocal information. Although the effect of voice was notably smaller than in the uncontrolled attention condition, there were still significant shifts in categorisation depending on the degree of gender information contained within the voice of the audiovisual stimulus. Although overall, the main effect of voice which was seen within the ratings data did not manifest in the reaction time data, an incongruent voice pairing caused a visible cost in reaction times at certain points of the face morph continuum (noticeably, end points of the continuum). Generally, these results underline the strong effect of voice observed in the previous analyses, particularly in contrast to that of the face; and suggest that in this experiment, processing of voice gender was mandatory, and consequently it was automatically integrated with face gender information.

4.5.3 Auditory dominance in integration of face-voice gender?

Our results contrast somewhat with those from studies such as Joassin et al. (2011b) and Latinus et al. (2010), who reported a larger dominance of vision over audition. Specifically, these authors observed a significant faster classification of faces as compared to voices and additionally, Latinus et al. (2010) found that participants were not able to ignore face gender information, even when instructed to. However, it should be noted that both these studies are characterised by the use of un-manipulated face stimuli (e.g., their stimuli contained important additional gender cues such as hair) that could have introduced a larger amount of sexually-dimorphic physical differences between the conditions. Their results suggest that in everyday life situations the perception of gender from faces could

dominate over voices; however, it is unclear how the contribution of culture-specific variations such as facial hair, hair length, and make-up – factors which play a crucial part of gender discrimination – could have affected their results. We chose to investigate the perception of gender using more stimuli that were perhaps more gender constrained: these stimuli remained ecological, in that we used articulating faces with time-synchronised vocalisation, but a significant effort was made to remove potential "cultural" cues of gender from our face stimuli, as indeed suggested by Latinus et al. (2010) as a direction for future work.

Smith et al. (2007) previously demonstrated a strong effect of vocal information in the perception of gender, in that low-level auditory features strongly influenced the categorisation of face gender. In that study, the gender of the faces was ambiguous and thus, gender attribution was mostly based on auditory cues. This result can be accounted for by the information reliability hypothesis, which suggests that the dominant modality is whichever is more appropriate and the more efficient for the realisation of the task (Anderson, 2004). In Smith et al. (2007), the faces offered little – or confusing – gender information and thus the gender-specific pure tones were the most reliable source of gender cues. Although it might have been expected that vision would dominate over audition in the present experiment, it is perhaps not surprising that we found a stronger effect of voice. With regards to gender, voices arguably show greater dimorphisms than faces. For example, the fundamental frequency (f0), which determines the perceived pitch of a voice, is typically higher in females by one octave, as compared to male voices (Linke et al., 1973). It could be that when discrimination of faces becomes more difficult, sexual dimorphism in the voices provide a strong source of gender information which is used more by participants. However, due to limited scientific investigation into the area of integration of face and voice gender information, it is difficult at present to conclude

definitively whether one modality actually dominates over the other. This will only be clarified by using a range of controlled and uncontrolled stimuli, where the amount of information in one modality or the other is carefully modulated: for example, by using normalised faces and voices, employing masking techniques, or by controlling the timbre of individual voices. We suggest that this will have a crucial impact on modality dominance, and observed integration effects.

In this study we aimed to advance on the pioneering work already completed in this young field, in a number of ways. We have made an effort to improve the ecological validity of stimuli – specifically, by creating articulating faces with time-matched voices. Our inclusion of static portraits enabled us to directly compare, for the first time, whether there was a significant difference between processing of dynamic and still faces when integrating bimodal face-voice gender information. Our study also utilised morphing techniques in order to create parametric manipulations of both face and voice gender morph. Using these morphing techniques allowed us to also create ambiguous face–voice pairs, manipulate face-voice congruence in a more fine-grained manner, and to test, using controlled experimental manipulation, the respective influence of faces and voices in the multimodal processing of gender.

One limitation in our study is the fact that we used only bimodal stimuli, and thus were unable to compare responses between these and unimodal stimuli. However, it should be noted that the lack of unimodal conditions did not prevent us from drawing conclusions on effects of congruence, facial articulation effects on integration patterns and sensory dominance in the perception of gender; and secondly, that the large literature on both face and voice perception allows for at least an indirect comparison with existing studies.

Further related studies, however, could include unimodal conditions in order to directly quantify any gain of multimodal information in the perception of gender.

4.6 Conclusion

In conclusion, we found that overall, participants integrated gender information from the face and voice, with categorisation reflecting an input from both modalities. However, in conditions that directed attention to either modality, we observed that participants were unable to ignore the gender of the voice, even when instructed to. This strong effect of voice was also reflected in both categorisation and latency results from the condition which did not direct attention.

5. Audiovisual integration of face-voice emotion: an fMRI investigation

5.1 Abstract

In the everyday environment, non-verbal emotional communication is multimodal (e.g., affective tone, facial expression). Understanding these communicative signals and integrating them into a unified percept is paramount to successful social behaviour. While many previous studies have focused on the neurobiology of emotional communication in the voice or face alone, far less is known about integration of auditory and visual non-verbal emotional information. The present study investigated this process using event-related fMRI, in conjunction with novel morphed face-voice stimuli. Behavioural data revealed that participants took into account both face and voice when categorising emotion. Furthermore, we observed adaptation effects – both within and across modality – where preceding affective information presentation affected the speed of categorisation to a following stimulus. These adaptation effects were also mimicked at the neural level. In addition to modality-specific adaptation, we observed a crossmodal adaptation effect in the right pSTS, providing evidence that integration in this region might be subserved by multimodal neurons. Additionally, activity across the right STS was modulated by the level of congruency between the face and the voice. Overall, these results clearly support the role of the STS in multimodal emotion processing.

5.2 Introduction

Stimulation in natural settings usually recruits a number of different sensory channels simultaneously (Stein and Meredith, 1993). Particularly important with regards to social

interactions is the perception of emotional cues from the face and the voice. These auditory and visual cues are important for conveying the emotional state of an individual to other persons and the integration of such cues is an essential part of face-to-face social interactions (de Gelder and Vroomen, 2000). Indeed, emotional information in a social context is inherently multimodal. Congruency between facial expression and emotional prosody facilitates emotion recognition (de Gelder and Vroomen, 2000; Ethofer et al., 2006) and additionally, emotional prosody can alter facial emotion perception (Massaro and Egan, 1996), even with the explicit instruction to ignore this information (de Gelder and Vroomen, 2000; Collignon et al., 2008).

Over the past decade in particular, a number of studies have addressed the question of how emotional information from different modalities is processed, extending upon findings on integration of low-level audiovisual cues (Calvert and Thesen, 2004; Stein and Stanford, 2008). Earlier behavioral findings have more recently been paralleled by results from electrophysiological and functional brain imaging studies, providing new insights into the neural processes underlying multimodal emotion integration. Studies in nonhuman primates have revealed an ability to integrate socially relevant multimodal cues from conspecifics (Ghazanfar and Logothetis, 2003), which is characterised by responsiveness of the superior temporal sulcus (STS) (e.g. Ghazanfar et al., 2008), amygdala and auditory cortex (Ghazanfar et al., 2005; Remedios et al., 2009), and prefrontal cortex (Sugihara et al., 2006). In humans, a number of sites have been proposed to integrate affective information from the face and the voice, including the primary sensory cortices (de Gelder et al., 1999; Pourtois et al., 2000, 2002), temporal regions such as the superior temporal sulcus (STS) (Pourtois et al., 2000, 2005; Ethofer, 2006; Kreifelts et al., 2007, 2010; Robins et al., 2009), and affective structures such as the amygdala (Dolan et al., 2001; Ethofer et al., 2006a; Klasen et al., 2011).

Deciding whether a neuron is 'multisensory' on the basis of single cell recordings is relatively straightforward. Integration is thought to occur when the response to a combined stimulus (e.g., audiovisual) is different from the response predicted on basis of the separate responses (e.g., auditory and visual). The initially employed criterion was that a neuron's spike count during multisensory stimulation should exceed that to the most effective unisensory stimulus (Stein and Meredith, 1993). For 'true' multisensory integration to occur, information from the different sensory systems needs to converge on individual neurons (Meredith, 2002). When the inputs converge in the same area and also synapse on the same neurons, this is termed 'neuronal convergence'. This is in contrast to 'areal convergence', where different sensory inputs converge in the same region, but without synapsing on the same neurons. This intermingling of unimodal populations may occur in a number of regions, but it does not mean that these singular neurons actually integrate the sensory inputs.

When dealing with fMRI data, the decision of whether a voxel or region is multisensory becomes far more complicated. Firstly, one voxel contains hundreds of thousands of neurons, the activity of which is averaged out by the fMRI signal. This voxel may not contain a homogenous set of neurons (e.g., all multisensory): instead, the large sample of neurons can be made up of mixed unisensory and multisensory sub populations (Laurienti et al. 2005). For example, the heterogeneous nature of the multisensory STS was demonstrated in a high-resolution fMRI study which showed that it consisted of mixed visual, auditory and audiovisual sub-populations (Beauchamp et al. 2004). Furthermore, multisensory sub-populations can themselves consist of multisensory neurons with very diverse response properties (additive, super- or sub-additive; Perrault et al., 2005). Conventional fMRI approaches are unable to specifically 'tag' these different types of neurons. Secondly, the appropriateness of the criteria used within fMRI to infer regions as

integrative is still under debate (e.g. Ethofer, 2006; Beauchamp, 2005; Love et al., 2011; see also **Chapter 2** of this thesis), with certain criteria such as super-additivity proposed to be too stringent (e.g. Ethofer, 2006; Beauchamp, 2005) and others such as the 'mean criterion' too liberal (Beauchamp, 2005). For example, without an initial criterion requiring unisensory activation in each modality, the mean criterion can classify unisensory areas as multisensory (Beauchamp, 2005). Therefore, the regions highlighted in fMRI studies depends in part on the statistical analysis used to quantify integration, and which of these is chosen can result in markedly different patterns of activation (Love et al., 2011). Thus, it is questionable whether we have a completely clear idea of which regions integrate these two unimodal sources. Finally, to date, a plethora of terms have been used in the context of multimodal research (e.g. 'heteromodal', 'multimodal', 'crossmodal', 'multisensory', 'amodal', 'supramodal' etc.). In certain instances, some of these terms have been used interchangeably, despite the fact that in different contexts they can have quite distinctive meanings (Calvert et al., 2000). Consequently, a single term is sometimes applied to a number of regions that actually perform different functions, sometimes making interpretation of previous findings complex.

Here we used an efficiency-optimised functional magnetic resonance imaging (fMRI) adaptation (fMR-A; Grill-Spector et al., 1999) paradigm – the so-called 'continuous carry-over design' (Aguirre, 2007) - to explore face-voice adaptation to affective information. FMR-A allows the researcher to move beyond the limited spatial resolution imposed by fMRI, and to draw inferences on neuronal populations within voxels. We made use of the possibilities offered by the recent developments in both facial and auditory morphing techniques in order to create a range of novel audiovisual stimuli that were parametrically morphed in both modalities. Additionally, our face-voice stimuli were dynamic with time-synchronised vocalisations, so to provide an ecological experience that has rarely been

seen in previous experiments. Participants were scanned in a rapid event-related design while viewing the parametrically morphed audiovisual movies, and performing a 2-alternative forced choice (2AFC) emotion classification task. The continuous carry-over design allowed us to examine in an optimally efficient way the repetition–suppression effect - that is, the effect of one stimulus on the cerebral response of the one presented immediately after. Due to our use of bimodal stimuli, we were not only able to investigate 'unimodal' adaptation effects – albeit within a multisensory context – but also the effect of one modality upon another.

Our aim was to investigate whether fMRI adaptation would reveal not only voxels responding to within-modality adaptation, but also multisensory voxels showing crossmodal repetition suppression following the repetition of an emotion across modality. This builds on previous behavioural work which has suggested that information in the voice can prime face recognition (Ellis et al., 1997; Hills et al., 2010), and infant results from one other fMRI study which observed that visual information could adapt the neural response to objects experienced tactilely (Tal and Amedi, 2009). We predicted that if neurons truly integrate the information from the face and voice, presenting an emotion representation in one modality followed by the same representation of the emotion in the different modality would result in the suppression of the activation of these neurons, and hence lead to a reduced fMRI signal. This would be in contrast to a recovery of the signal, which would simply imply that this activity originates from a combination of neuronal populations, each tuned to the emotion exposure in the visual modality or the auditory modality. If this were the case, a presentation of information from a new modality would activate a new group of neurons, and the result would be a stronger, non-adapted fMRI signal. Furthermore, we proposed that crossmodal adaptation could potentially highlight

'supramodal' regions – that is, those containing neurons activated by a concept as compared to sensorial information.

Due to the unique nature of our experimental design, we were not only able to examine the effect of one stimulus upon another, but also the mean difference in response between different sets of stimuli (i.e., 'direct effects'). Therefore, we exploited our paradigm in order to investigate also the effects of stimulus congruence and ambiguity on brain activity. The study of congruence in particular is one which researchers have used to search for regions which play a role in audiovisual processing (e.g. Calvert et al., 2000), and thus in our experiment acted as a complement to our examination of crossmodal adaptation.

Studies of audiovisual emotion representation have typically compared congruent audiovisual stimuli with purely auditory or visual ones (Kreifelts et al., 2007, 2010; Robins et al., 2009). However, this introduces the confound of perceptual load, which is increased when there is bimodal stimulation. Therefore, it is important that studies employ experimental designs controlling for perceptual load to investigate audiovisual integration independently of this aspect of the integration process. One experimental approach would be to replace the unimodal conditions with bimodal conditions in which either the auditory or visual cues are 'scrambled' while conserving the perceptual complexity of the auditory/visual cue, thus resulting in experimental conditions with comparable perceptual load (i.e., AV, $A_{scrambled}V$, $AV_{scrambled}$). However, these artificial or 'scrambled' conditions may themselves trigger confounding crossmodal interaction processes. Thus, perhaps a better way to control perceptual load is to employ a congruence design where a condition with emotionally congruent bimodal stimulation is compared to emotionally incongruent bimodal cues. Because only congruent unimodal stimuli are thought to be able to 'bind'

together, a heightened response to congruent information is considered to be a reflection of audiovisual integration.

Dolan et al. (2001) compared congruent and incongruent audiovisual stimuli, and found that congruent information (specifically, fear) provoked an increased response in the left amygdale. However, the authors of this study combined emotional sentences with static faces. Because multisensory stimulus integration relies on spatial and temporal coincidence (Stein and Meredith, 1993), respective paradigms naturally require precisely matched dynamic stimuli; thus static faces paired with voices are sub-optimal stimuli. Additionally, correlates of successful emotion integration should be separated from those of audiovisual speech integration in general – thus, stimuli should be devoid of any linguistic content.

Klasen et al. (2011) conducted the first study investigating the multimodal representation of emotional information with dynamic stimuli expressing facial and vocal emotions congruently and incongruently. Using novel computer-generated stimuli, they combined emotional faces and voices in congruent and incongruent ways and measured brain responses using fMRI during an emotional classification task. Both congruent and incongruent audiovisual stimuli evoked larger responses in thalamus and superior temporal regions compared with unimodal conditions. Congruent emotions were characterized by activation in amygdala, insula, ventral posterior cingulate (PCC), temporo-occipital, and auditory cortices, and incongruent emotions activated a frontoparietal network and bilateral caudate nucleus. The PCC alone exhibited differential reactions to congruency and incongruency for all emotion categories (in addition to the amygdala for expressions of happiness), leading the authors to conclude that emotional information does not merge at

the perceptual audiovisual integration level in unimodal or multimodal areas, but in vPCC and amygdala.

In our study, by generating all possible pairings of face and voice morphs, we created a range of audiovisual stimuli that were parametrically varied in incongruence, allowing us to extend upon these previous results by examining perception of congruence in a more fine-grained manner. We also developed upon the study by Klasen et al. (2011) by attempting to disentangle the factors of task difficulty from those of emotional congruency by also taking into account stimulus ambiguity.

Finally, we were also able to investigate direct effects of face and voice emotion upon brain activity and compare these to any activity elicited as a result of emotional context (i.e., adaptation effects). What is more, our assessment of direct effects also enabled us to look for any interactions between the two modalities, which can highlight regions where activity is due to a unique combination of face and voice emotion. Overall, the approach of this study – investigating both direct and adaptation effects, in addition to emotional congruence – provided us with an opportunity to investigate audiovisual processing of emotion from a number of different angles within one experiment. Consequently, we were able to uncover not only potential networks for unimodal face and voice affective processing (albeit within a multisensory context), but also those responsible for combining information from the two modalities.

5.3 Materials and Methods

5.3.1 Participants

Ten English-speaking participants (4 males and 6 females; mean age 27 years (± 13 years)) took part in a pre-test of stimuli, in order to ensure there was appropriate categorisation of unimodal emotion, and a new group of eighteen participants (10 males, 8 females, mean age: 25 years (± 3.7 years)) were scanned in the main fMRI experiment. All had self-reported normal or corrected vision and hearing. The ethical committee from the University of Glasgow approved the study. All volunteers provided informed written consent before, and received payment at the rate of £6 p/hour for participation.

5.3.2 Stimuli

Video recording

Two actors (one male, one female) were recorded. Both had studied drama at University level. The actors were paid for their participation at the rate of £6 p/hour. Each actor sat in a recording booth, and was given instructions through an outside microphone connected to speakers within the booth. The actor wore a head cap, in order to hide the hair, and a marked head panel was fitted to the cap, which was used to determine head position. A Di3d capture system (see Winder et al., 2008) was used for the video recording. The actor sat between two camera pods, at a distance of 143 cm away from them both. Thus, each camera captured a slight side-view of the face, as opposed to a directly frontal view. Each pod consisted of a vertical arrangement of 3 different cameras. The top and bottom cameras were black and white, and were used to capture general shape information. The middle camera in each pod was a colour camera, used to capture texture and colour information. A lamp was placed behind each camera, and luminance kept constant at 21 amps. Video information was recorded by Di3D software on this PC as a series of jpegs at

high resolution (2 megapixels). Vocal sound-information was transmitted via a Microtech Geffell GMBH UMT 800 microphone – positioned above the actor - to a second PC outside the booth, and was recorded at 44100 Hz using Adobe Audition (Adobe Systems Incorporated, San Jose, CA).

The actors were instructed to express anger and happiness in both the face and the voice. The sound 'ah' was chosen as it contains no linguistic information. They were asked to sit as still as possible, in order to keep head movement to a minimum. Expressions were produced a number of times, with a pause of three seconds between each repetition. The actor clapped in front of their face before they produced each set of expressions, which provided markers when later matching the audio recording to the video.

Video processing

Video output was split into a number of different sequences, where each sequence was made up of a number of jpegs (frames) and each repetition of each emotional expression formed one sequence. Two final sequences were chosen for each actor. Using the Di3D software, 43 landmarks were placed around the face and facial features in the first and last frame of the sequence, forming a landmark-mesh. An existing generic mesh was applied to the beginning and the end of the sequence (i.e., first and last jpeg), which was then warped to fit the landmark-mesh. The first mesh was then used to estimate the mesh position in the second jpeg, which was then used to estimate the position in the third and so on. This forward tracking/mesh estimation was then carried out in the opposite direction (i.e. the last mesh was used to estimate the mesh position in the jpeg before it). The two side-views of the actor, one from each camera pod, were merged together, forming one directly frontal view of the face. We smoothed the converging line, which ran from the forehead to the chin down the middle of the face, using average facial texture information.

Any head movement was removed by tracking and aligning the eight marked points on the head panel, so that they were always in the same position throughout the sequence.

Audio processing

In addition to the original sound recording, a duplicate reduced-noise version was also produced. A recording made in the empty booth provided a 'noise-baseline', which was used to remove noise using a Fourier transform. The entire reduced-noise audio recording for each actor was then edited in Adobe Premiere (Adobe Systems Incorporated, San Jose, CA). Using the actor claps as markers for the start of each emotional expression, the audio sequences corresponding to the correct video sequence frames (at a frame rate of 25 frames per second) were identified and split into separate clips. The separate audio samples were then normalised for mean amplitude using Adobe Audition.

Video morphing

The video morphing was performed independently on the texture and shape components of the 4D models. The texture was warped onto a common template shape using the piecewise-affine warp and the morph was then performed as a weighted linear sum on the RGB pixel values; the shape was normalised for rigid head position (i.e. rotation, translation) using a combination of the ICP (Besl and McKay, 1992) and the RANSAC (Bolles and Fischler, 1981) methods and the morph was then performed as a weighted linear sum on the vertex coordinates. To account for timing differences between two expressions, pairs of matching anchor frames were selected in the two sequences corresponding to similar movement stages (for example, 'mouth first opens', 'maximum mouth opening', etc.). The sequence pairs were broken up into segment pairs between the anchor points and the lengths of the pixel and vertex timecourses for the segment pairs were rescaled to be equal for the pair using linear interpolation. The new length was

chosen as the average length of the segment over the pair. Finally, the segment pairs were reassembled into the full sequence pair and the morph was performed at each frame of the sequence. Five morph levels were chosen - ranging from 10% to 90% of one expression, in 20% steps - and the same morph level was used at each frame of the sequence, producing a total of five morph sequences which were rendered to video using 3DS Max.

Audio morphing

Auditory stimuli were edited using Adobe Audition 2.0. In order to generate the auditory components to the 'morph-videos' three temporal and three frequency points were identified and landmarks corresponding to these set in the MATLAB-based morphing algorithm STRAIGHT (Kawahara, 2003), which were then used to generate a morph continuum between the two affective vocalisations equivalent to the faces. Two continua of voices – one for each actor, and consisting of five different voices ranging from 90% angry to 90% happy in 20% steps - were then generated by resynthesis based on a logarithmic interpolation of the angry and happy voices temporal and frequency anchor templates to a 50% average.

Audiovisual movie production

The auditory and visual morphing procedures produced five dynamic face videos and five audio samples for each actor. Within actor, these stimuli were all equal length. In order to ensure all stimuli were of equal length, we edited video and audio clips between actors. In all video clips, seven important temporal landmarks that best characterised the facial movements related to the vocal production were determined, and the frames at which they occurred were identified. These landmarks were the first movement of the chin, first opening of lips, maximum opening of the mouth, first movement of the lips inwards, time point at which the teeth met, closing of the lips, and the last movement of the chin. The

theoretical average frames for these landmarks were then calculated, and the videos edited so the occurrence of these landmarks matched in all clips. Editing consisted of inserting or deleting video frames during fairly motionless periods. The editing produced ten adjusted video clips, each 18 frames (720 ms) long. The audio samples were then also adjusted in accordance with the temporal landmarks identified in the video clips, in order to create 10 vocalisations (5 for each actor) of equal length. Within actor, the five visual and five auditory clips were then paired together in all possible combinations. This resulted in a total of 25 audiovisual stimuli for each actor, parametrically varying in congruence between face and voice affective information.

5.3.3 Pre-test: stimulus validation

In a pre-test, using the separate group of ten participants, we investigated categorisation of our stimuli across the two actors, firstly in order to ensure that expressions were recognised as intended, and secondly to clarify that there were no significant differences in categorisation of expressions produced by different actors. Five participants were assigned to the expressions of the male actor, and another five were assigned to the expressions of the female actor. The stimuli were played to participants through a FLASH (www.adobe.com) object interface running on the Mozilla Firefox web browser. For each condition, stimuli were preloaded prior to running the experiment. The conditions were as follows:

1. Audio only

In this condition, participants heard a series of voices alone. They were instructed to listen to each voice, and make a decision on emotion based on the voice they had just heard. Again they indicated their decision via a button press. The five voice morphs were presented 10 times each, in a randomised order in one block consisting of 50 trials.

2. Video only

Participants saw all face videos, uncoupled with a voice. They were instructed to watch the screen and indicate their decision regarding emotion in the same way as before. The five faces were presented 10 times each, in randomised order in one block consisting of 50 trials.

Participants could respond either whilst the stimulus was playing, or after it ended. Regardless of when they responded, there was a 100ms wait until the next stimulus began playing. Conditions were counterbalanced between participants, with five participants for each of the two possible orders (collapsing across actor gender).

Categorisation data was submitted to two, two factor mixed ANOVAs. In the first, degree of face emotion morph was a within subject factor, whilst the actor (male or female) was a between subject factor. This analysis highlighted a significant effect of face emotion morph on categorisation ($F(1.35,10.8)=126$, $p<0.0001$). There was no effect of actor on categorisation ($F(1,8)=0.949$, $p=0.359$). Face categorisation results (averaged across actors) yielded the classic sigmoid-like psychometric function from the emotion classification task, with a steeper slope at central portions of the continua. The percentages of anger identification were 96% (± 2.23%) for the 90% angry face, and 2% (± 2.74%) for the 90% happy face. The 50% angry-happy face was identified as angry 53 times out of 100 (± 9.75%). In the second ANOVA, degree of voice emotion morph was a within subject factor, whilst the actor was the between subject factor. This analysis highlighted a significant effect of voice emotion morph on categorisation ($F(2.23,17.7)=127$, $p<0.0001$). There was no effect of actor on categorisation ($F(1,8)=0.949$, $p=0.575$). Again, voice categorisation results (averaged across actors) yielded the classis sigmoid-like psychometric function from the emotion classification task, with a steeper slope at central

portions of the continua. The percentages of anger identification were 96% (± 6.52%) for the 90% angry voice, and 0% (± 0.00%) for the 90% happy voice. The 50% angry-happy voice was identified as angry 32 times out of 100 (± 19.4%).

5.3.4 Design and Procedure

Continuous carry-over experiment

In the main experiment, stimuli were presented using the Psychtoolbox in Matlab, via electrostatic headphones (NordicNeuroLab, Norway) at a sound pressure level of 80 dB as measured using a Lutron Sl-4010 sound level meter. Before they were scanned, subjects were presented with sound samples to verify that the sound pressure level was comfortable and loud enough considering the scanner noise. Audiovisual movies were presented in two scanning runs (over two different days) while blood oxygenation-level dependent (BOLD) signal was measured in the fMRI scanner. We used a continuous carry-over experimental design (Aguirre, 2007). This designs allows for measurement of the direct effects (i.e., that of face and voice emotion morph) and the repetition suppression effect, which can be observed in pairs of voices or faces (like the typical fMRI adaptation experiments).

The stimulus order followed two interleaved N=25 Type1 Index 1 sequences (one for each of the speaker continua; ISI: 2s; Noyane and Theobald, 2007), which shuffles stimuli within the continuum so that each stimulus is preceded by itself and every other within-continuum in a balanced manner. The sequence was interrupted by seven 20s silent periods, which acted as a baseline, and at the end of a silent period the last 5 stimuli of the sequence preceding the silence were repeated before the sequence continued. These stimuli were removed in our later analysis. Participants were instructed to perform a 2AFC emotion classification task using 2 buttons of an MR compatible response pad (NordicNeuroLab, Norway). They were also instructed to pay attention to both the face

and voice, but could use the information presented in whatever way they wished to make their decision on emotion. Reaction times (relative to stimulus onset) were collected using Matlab with a response window limited to two seconds.

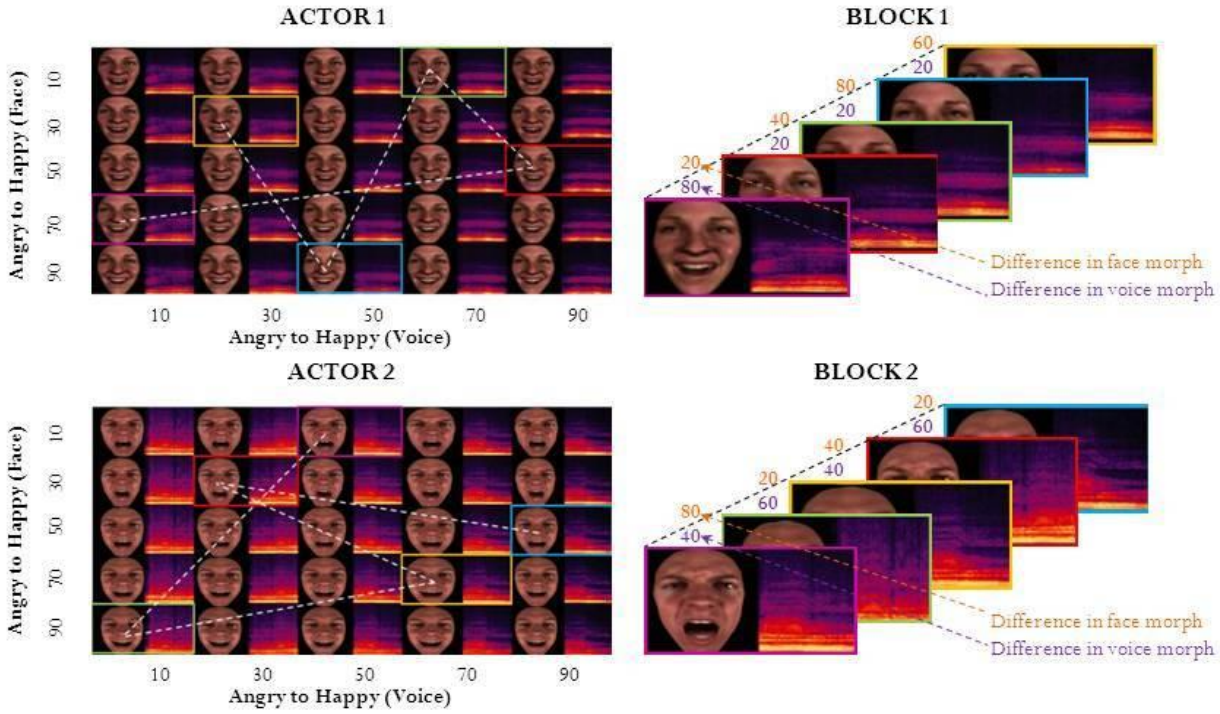


Figure 5.1. Stimuli and continuous carry-over design. Anger and happiness expressions produced by two actors were morphed between 10% and 90% anger, in 20% steps, creating 25 different audiovisual stimuli per actor. Expressions from each actor were presented in two interleaved Type1 Index1 (n=25) continuous carry-over sequences, over two experimental runs. Each block contained expressions from only one actor, and blocks were alternated between actor.

Localisation of the temporal voice areas (TVA; Functional Localiser Experiment)

A functional localiser of the temporal voice areas (TVA; Belin et al., 2004; Belin et al., 2011) was conducted for each subject. This consisted of a 10 minute fMRI scan measuring the activity in response to either vocal or non-vocal sounds (Belin et al. 2000; Pernet et al. 2007) using an efficiency-optimized design. Briefly, the voice localiser involved participants listening passively to 8-sec blocks from either vocal or non-vocal sound

categories presented with a 33% proportion of silent blocks in an efficiency-optimized pseudo-random order. Vocal sounds were either speech (for example, isolated words, connected speech (e.g. phrases) in several languages) or non-speech (such as laughs, sighs and coughs) produced by several speakers of different gender and age. Non-vocal sounds included natural sounds, animal cries, mechanical sounds, instrumental sounds. Stimuli are available at http://vnl.psy.gla.ac.uk. The response to vocal as compared to non-vocal sounds can be contrasted, in order to localise the TVA. The independent functional localiser was used in voxel selection/region of interest (ROI) definition. Generally, its aim was to identify whether voice-specific statistical maps from our audiovisual carry-over experiment overlapped with the TVA.

Localisation of the fusiform face area (FFA) and face-selective network (Functional Localiser Experiment)

As for voices, a functional localiser of the fusiform face area (FFA; Kanwisher et al., 1997) was conducted for each subject. Participants viewed alternating blocks of faces, houses and noise; noise patterns constructed from the two other conditions (Vizioli et al., 2010). All images were shown as uniform grey presented on a white background, and measured 11.25 degrees of visual angle: faces were cropped using an elliptical annulus to remove neck, ears and hairline from the images. Blocks of the three categories lasted for 18s and were made up 20 image presentations lasting 750ms, separated by 250ms of blank white screen. Five blocks of each category were shown. Each scan began with 12s of fixation cross on a uniform background at the start of the each run, and again between each condition block. Participants completed 2 runs of the FFA localiser, each lasting 456s, using a fixed order: 1) faces, noise then houses and 2) reverse order. No task was given other than to attend displays and maintain fixation. Parallel to the voice-localiser, the response to face as compared to non-face information was contrasted, in order to localise face-selective

regions – particularly, the FFA. The independent functional localiser was used in voxel selection/region of interest (ROI) definition. Similarly to the voice-localiser, its aim was to identify whether face-specific statistical maps from the audiovisual carry-over experiment overlapped with the FFA.

5.3.5 Imaging parameters

Functional images covering the whole brain (slices=32, field of view=210x210mm, voxel size=3x3x3mm) were acquired on a 3T Tim Trio Scanner (Siemens) with a 12 channel head coil, using an echoplanar imaging (EPI) sequence (interleaved, TR=2s, TE=30ms, Flip Angle=80 degrees) were acquired in both the carry-over and localiser experiments. In total, we acquired 1560 EPI image volumes for the carry-over experiment, split into two scanning sessions consisting of 780 EPI volumes; 336 EPI volumes for the voice-localiser; and 512 EPI volumes for the face-localiser, split into two experimental runs of 255 EPI volumes. For both the carry-over experiment and experimental localisers, the first 4s of the functional run consisted of 'dummy' gradient and radio frequency pulses to allow for steady state magnetisation during which no stimuli were presented and no fMRI data collected. MRI was performed at the Centre for Cognitive Neuroimaging (CCNi) in Glasgow, UK.

At the end of each fMRI session, high-resolution T1-weighted structural images were collected in 192 axial slices and isotropic voxels (1 mm3; field of view: 256x256 mm^2, TR=1900ms, TE = 2.92ms, time to inversion = 900ms, FA = 9 degrees).

5.3.6 Statistical Analysis

Behavioural Data Analysis

a) Categorical data

Each participant's mean categorisation values for each audiovisual emotion morph stimulus (collapsed across actor) was submitted to a two factor (face morph and voice morph), repeated measures ANOVA, with 5 levels per factor (percentage of 'anger' information in the morph). This was in order to assess the overall contributions of face and voice emotion morph on categorical response.

b) Reaction time data

Effect of degree of morph

Each participant's mean reaction time values for each stimulus (collapsed across actor) were firstly submitted to a two factor (face morph and voice morph), repeated measures ANOVA, with 5 levels per factor (percentage of 'anger' information in the morph). As with categorical data, this was in order to assess the overall contribution of face and voice emotion morph – or the 'direct effects' of face and voice morph - on reaction times.

Effect of ambiguity and congruence

Secondly, we computed a multiple regression analysis to investigate the relative contribution of ambiguity and congruence of our audiovisual stimulus on the reaction times in individual subjects. These values took into account the emotion morph contained in both the face and the voice. Congruence was defined as the absolute value of face morph level minus voice morph level. Therefore, the higher values indicated the highest degree of incongruence. However, we recognised that although completely congruent stimuli were all assigned the same value, some would presumably be easier to categorise than others (e.g., 90% angry face-90% angry voice as compared to 50% angry face-50% angry voice). Therefore, ambiguity took into account the clarity of the *combined* information of the face and the voice. To do this, we calculated the average percentage of 'anger' information

contained in the stimulus. For example, the 90% angry face-90% angry voice stimulus contained 90% anger informativeness, and the 10% angry face-90% angry voice contained 50% anger informativeness – as did the 50% angry face-50% angry voice stimulus. We then performed the following calculation:

$$\textbf{\textit{Ambiguity}} = (50\%-(\text{average \% anger information}))*2$$

This calculation measured the difference between the combined affective information in the stimulus and a completely ambiguous value of 50%. This resulted in ambiguity values which were a 90 degree rotation of our congruence values, where the values indicated the level of unambiguous affective information contained within the stimulus as a whole. The higher values indicated a clearer combined emotion representation (unambiguous) and lower values indicated an unclear combined emotion representation (ambiguous). For ambiguity and congruence values assigned to each stimulus, see Figure 5.2. It should be noted that there was a significant negative correlation between ambiguity and congruency values ($r=-0.556$, $p<0.0001$).

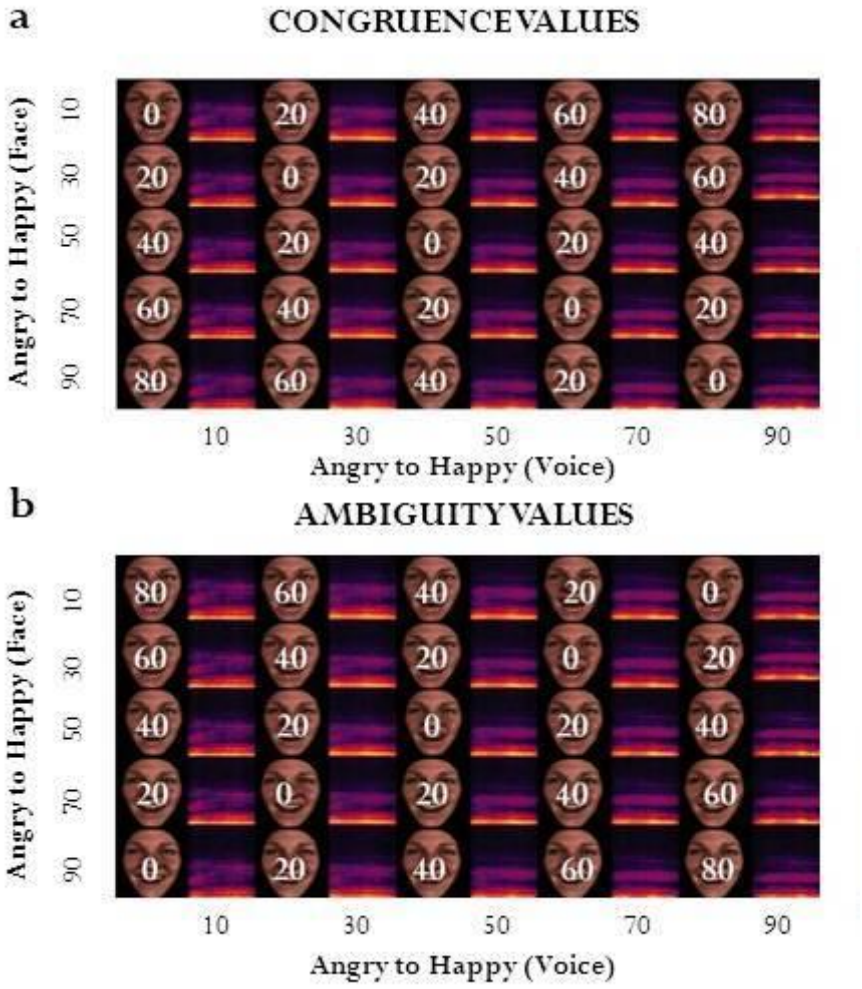


Figure 5.2. Congruence and ambiguity values assigned to stimuli. (a): Congruence values. Values represent the percentage of incongruence between the face and voice (highest values=highest level of incongruence). (b): Ambiguity values. Values represent the percentage of unambiguous information contained within the stimulus (highest values=least ambiguous) . Values are superimposed on Actor 1, but values were identical for the stimuli of both Actor 1 and 2.

Effect of physical distance

Finally, reaction time data was then submitted to two further ANOVAs, which evaluated contextual (or adaptation) effects on reaction times.

Unimodal

The first five stimuli in each block were removed. Each participant's mean reaction time values for each stimulus in every block (collapsed across actor) were then submitted to a

two factor (difference in morph between face of that stimulus and the preceding stimulus (i.e., face to face physical distance), difference in morph between voice of that stimulus and the preceding stimulus (i.e., voice to voice physical distance)) repeated measures ANOVA, with 5 levels per factor (percentage of morph difference between two stimuli). This allowed us to observe the bearing of unimodal face and voice 'carry-over' effects (the effect of one stimulus upon another; i.e., the physical distance between consecutive stimuli of the same modality) upon reaction times.

Crossmodal

The first five stimuli in each block were removed. Each participant's mean reaction time values for each stimulus in every block (collapsed across actor) were then submitted to a two factor (difference in morph between face of that stimulus and the voice of the preceding stimulus (i.e., voice-to-face physical distance), difference in morph between a voice of that stimulus and the face of the preceding stimulus (i.e. face-to-voice physical distance)) repeated measures ANOVA, with 5 levels per factor (percentage of morph difference between two stimuli). This allowed us to observe the bearing of face and voice crossmodal carry-over effects – or the physical distance between consecutive stimuli of different modalities - upon reaction times.

Imaging analysis

SPM8 software (Wellcome Department of Imaging Neuroscience, London, UK) was used to pre-process and analyse the imaging data. First the anatomical scan was AC-PC centred, and this correction applied to all EPI volumes.

Functional data were motion corrected using a spatial transformation which realigned all functional volumes to the first volume of the run and subsequently realigned the volumes to the mean volumes. The anatomical scan was co-registered to the mean volume and segmented. The anatomical and functional images were then normalised to the Montreal

Neurological Institute (MNI) template using the parameters issued from the segmentation keeping the voxel resolution of the original scans (1x1x1 and 3x3x3 respectively). Functional images were then smoothed with a Gaussian function (8mm FWHM).

EPI time series were analysed using the general linear model as implemented in SPM8. For each subject (first-level analysis), localiser and experimental data were modelled separately.

Localiser data

Within the voice localiser, voices and non-voices were modelled as events using the canonical haemodynamic response function (HRF), and one contrast per stimulus type was computed. A 'voice greater than non-voice' contrast was created for each subject, which was used at the group level (second-level analysis) in a one sample t-test to identify the voice selective regions (i.e., the TVA). Parallel to this, within the face localiser, faces, houses and noise were modelled as events and one contrast per stimulus type was computed. A 'face greater than non-face' contrast was created for each subject, which was used at the group level in a one sample t-test to identify face-selective regions (in particular the FFA and STS (whose activity is particularly relevant for processing of dynamic faces (e.g. Haxby et al., 2000)).

Localiser results were thresholded at $p<0.05$ (peak voxel FWE corrected). ROI analyses were conducted within MarsBar (Brett et al., 2002).

Direct effects of face morph and voice morph

In this first analysis, functional data were analysed in a two-level random-effects design. The first-level, fixed effects individual participant analysis involved a design matrix

containing brain activity - time-locked to stimulus onset and duration - modelled against 25 separate regressors, corresponding to each of our 25 stimuli. To account for residual motion artefacts the realignment parameters were also added as nuisance covariates to the design matrix. We then carried out a two factors, repeated measures ANOVA at the second level, within a RFX (random-effects) group analysis, which allowed us to observe regions in which brain activity was modulated by Face morph and Voice morph, in addition to any interactions between the two.

Effects of stimulus ambiguity and congruency

Functional data was analysed in two separate two-level random effects designs. Because ambiguity and congruence values were negatively correlated, we ensured that the variance explained only by either ambiguity or congruence was modelled by entering these values as the second parametric modulator in the respective design matrix, and the values we intended to covary out (either ambiguity or congruence) were entered as a first parametric modulator.

Ambiguity

Brain activity time-locked to stimulus onset and duration was modelled against the 1st (linear) expansion of two parametric modulators: congruence, then ambiguity. The linear expansion allowed us to investigate regions which responded more to ambiguous information as compared to unambiguous and vice versa, with an expected parametric linear modulation of signal by the degree of ambiguity in the stimuli. The contrast for the effect of the second parametric modulator - ambiguity - was entered into separate second-level, group RFX analysis. We then further looked at both positive and negative correlations of BOLD signal with ambiguity.

Congruence

Brain activity time-locked to stimulus onset was modelled against the 1st (linear) expansion of two parametric modulators: ambiguity, then congruence. The contrast for the effect of the second parametric modulator – congruence - was entered into separate second-level, group RFX analyses. We then further looked at both positive and negative effects of congruence.

Adaptation ('continuous carry over')

Functional data was analysed using four two-level random effects designs: two which examined unimodal carry-over effects, and two which examined crossmodal carry-over effects.

Unimodal

For both face and voice unimodal carry-over effects, brain activity time-locked to stimulus onset and duration was modelled in separate design matrices against one parametric modulator, which accounted for the absolute difference between the a) face or b) voice morph levels of consecutive bimodal stimuli.

Crossmodal

Brain activity was modelled against three parametric modulators: the first accounted for the absolute difference between the face morph levels of consecutive bimodal stimuli; the second accounted for the absolute difference between the voice morph levels of consecutive bimodal stimuli; and the third accounted for the crossmodal carry-over effect, which was either the absolute difference between the a) face morph of a stimulus and the voice morph of the preceding stimulus (i.e., voice-to-face physical distance), or b) the absolute difference between the voice morph of a stimulus and the face morph of the preceding stimulus (i.e., face-to-voice physical distance). Our design matrices ensured that

any crossmodal carry-over effects observed were not a result of unimodal carry-over effects, as the variance explained by these values was essentially regressed out.

In all four of our design matrices, a linear expansion allowed us to investigate regions where the signal varied in account with the physical difference between stimuli, with a hypothesised linear modulation of signal as the degree of morph level difference increased parametrically. Contrasts for the effects at the first level for each design matrix were entered into four separate second-level, group RFX analysis, in which we conducted a one-sample t-test.

Reported results from the experimental run are from whole-brain analyses, masked by an experimental audiovisual vs. baseline contrast thresholded at $p<0.001$ (peak voxel uncorrected), and are reported descriptively at a threshold of $p < 0.05$ (FWE cluster size corrected) in combination with $p< 0.001$ (peak voxel uncorrected), unless stated otherwise.

5.4 Results

5.4.1 Behavioural results

a) Categorical data

The percentages of anger identification were of 96.3% (± 4.7%) for the 90% angry face-90% angry voice stimulus and 2.78% (± 3.59%) for the 90% happy face-90% happy voice stimulus. The percentage of anger identification for the 50% ambiguous angry-happy stimulus was 49.4% (± 16.9%).

The repeated measures ANOVA highlighted a main effect of voice morph ($F(1.14, 19.4)=15.3$, $p<0.002$) and of face morph ($F(2.02, 34.3)=348$, $p<0.0001$), and also a significant voice x face interaction (($F(5.78, 98.1)=6.78$, $p<0.0001$)). The 3-D and 2-D

categorisation curves are shown in Figures 5.3a) and 5.4, respectively. These results highlighted that – at least in some points in the 3-D emotion space –the participants' decision on emotion was due to a combination of information from the two modalities. Figures 5.3a) and 5.4 illustrate that the voice had the greatest modulating effect at the ambiguous point in the face morph continuum, whereas face produced a strong modulating effect at all points of the voice morph continuum. Overall, it was clear that that information from the face exerted the strongest effect on categorical response.

b) Reaction time data

Effect of Degree of Morph

The mean reaction times for the two end point congruent stimuli were 813ms (± 67.4ms) and 779ms (± 64.4ms) (for 10% angry face-10% angry voice and 90% angry face-90% angry voice, respectively). For the 50% angry face-50% angry voice stimulus the mean reaction time was 895ms (± 100ms). Finally, for the two most incongruent stimuli (10% angry face-90% angry voice; 90% angry face-10% angry voice) these reaction times were 822ms (± 101ms) and 829ms (± 92.5ms).

The ANOVA of reaction time data highlighted a main effect of voice morph ($F(2.91,49.6)=11.8$, $p<0.0001$) and of face morph ($F(2.34,39.7)=70.6$, $p<0.0001$), and also a significant interaction between the two modalities ($F(2.90,39.4)=7.40$, $p<0.0001$). The 3-D and 2-D illustrations of reaction times are shown in Figures 5.3b) and 5.4, respectively. These results highlighted that the speed of the participants' decision on emotion was affected by the information contained in both modalities. For example, as can be seen in Figure 5.4, when 90% happy face was combined with a 90% happy voice (congruent), the reaction time was visibly quicker than when it was combined with a 90% angry voice (incongruent). Overall however, and as with categorical data, face morph had a stronger influence on reaction times. This can be seen in Figures 5.3b) and 5.4, where there was an inverted 'U' shape for reaction times along the face morph continuum.

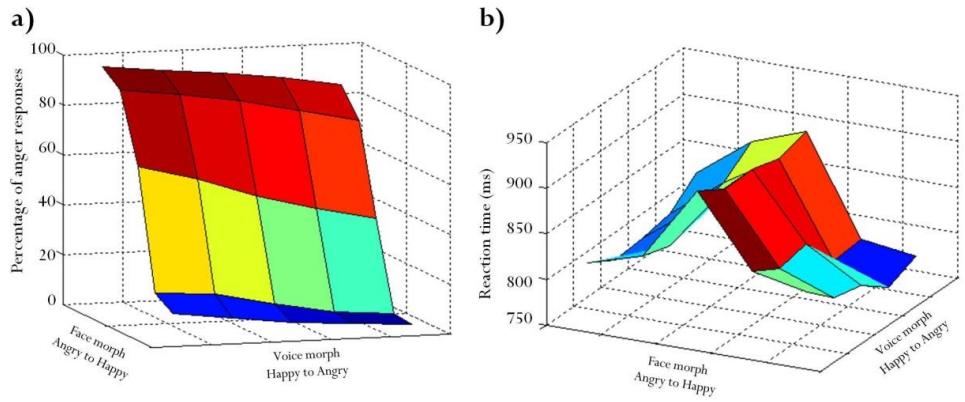


Figure 5.3. On-line behavioural results. a: 3D representation of categorisation results; b: 3D representation of reaction time results. Colour scale: red – blue=angry – happy responses

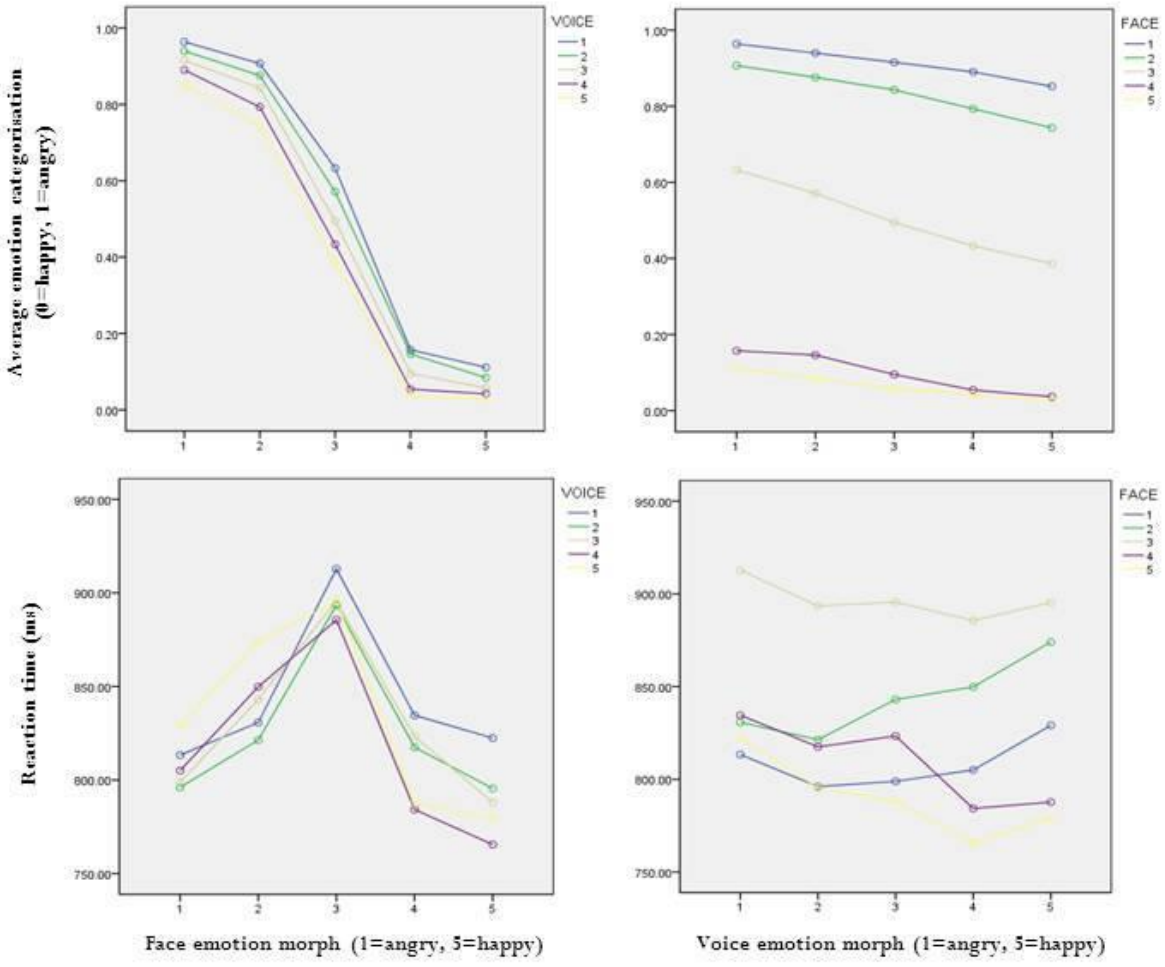


Figure 5.4. Direct effects of face and voice emotion morph (categorisation and reaction time results). Top left and right: categorisation results (0=0% angry, 1=100% angry); Bottom left and right: reaction time results. Emotion morph: 1=100% angry, 5=0% angry; Blue=100% angry, Green=70% angry, Beige=50% angry, Purple=30% angry, Yellow=10% angry.

Effect of ambiguity and congruence

The multiple regression analysis indicated that ambiguity was significantly related to reaction time (B=-18.7, t=-4.43, p<0.0001), with a higher level of ambiguity resulting in longer reaction times, but that congruence was not (B=-7.78, t=-1.83, p=0.067).

Effect of Physical Distance

In addition to examining 'direct effects' of Face and Voice morph, we also investigated the effect of an emotion presentation on the response to a secondary emotion presentation: adaptation effects.

a) Unimodal

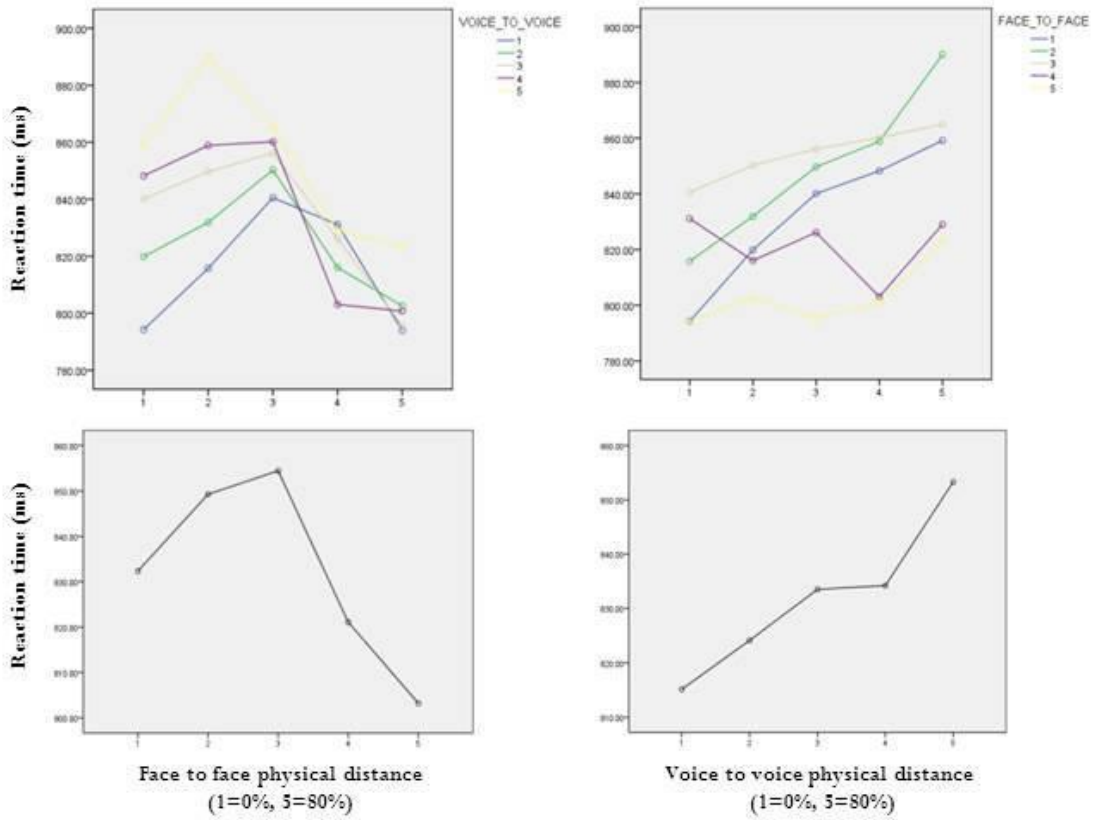
The two factor unimodal carry-over effects ANOVA (face-to-face physical distance, voice-to-voice physical distance) of reaction time data highlighted a main effect of both voice physical distance ($F(2.17,36.8)=16.2$, $p<0.0001$) and face physical distance ($F(2.84,48.2)=37.5$, $p<0.0001$), and also a significant interaction between both these factors ($F(4.76,90.0)=4.33$, $p<0.003$. The results from this ANOVA are shown in Figure 5.5a). These results are explained in more detail later in this chapter; however, overall they indicate an important influence of the previously heard voice on voice emotion identification, and similarly an influence of the previously heard face on face emotion identification. However, these contextual effects were different for the different modalities: generally, intermediate differences between consecutive face morphs resulted in the longest reaction times, whilst the largest differences between consecutive voice morphs resulted in the largest reaction times.

b) Crossmodal

The two factor crossmodal carry-over effects ANOVA (voice-to-face physical distance, face-to-voice physical distance) highlighted that there was a significant interaction between the two crossmodal effects ($F(6.54,111)=4.04$, $p<0.002$). There was a significant main

effect of voice-to-face physical distance ($F(2.62,44.5)=57.8$, $p<0.0001$) but no overall significant main effect of a face to voice physical distance ($F(2.94,50.1)=0.220$, $p=0.879$). These results are shown in Figure 5.5b). Again, these results are explained in more detail later in this chapter; briefly, they suggest that voice exerted a stronger adaptive effect on face than face did on voice. However, the significant interaction also indicates that at some point, the change in reaction times due to the voice-to-face physical difference depended on the face-to-voice physical difference occurring simultaneously, further showing that these two crossmodal effects were combined at some point to affect speed of emotion categorisation.

a) Unimodal adaptation



b) Crossmodal adaptation

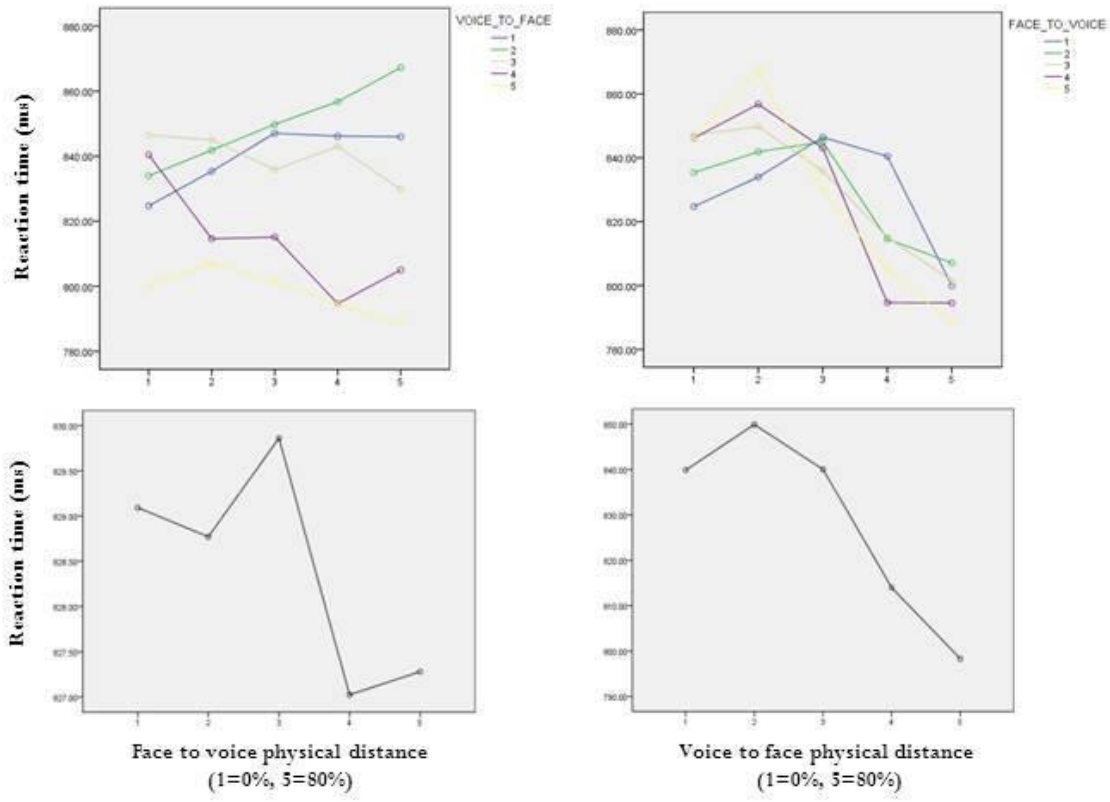


Figure 5.5 (previous page). Carry-over effects of face and voice emotion morph. a) Unimodal adaptation. Top and bottom left: unimodal face adaptation; Top and bottom right: Unimodal voice adaptation. b) Crossmodal adaptation. Top and bottom left: crossmodal effect of face upon voice; Top and bottom right: crossmodal effect of voice upon voice.
Physical distance between consecutive stimuli: 1=0%, 5=80% difference; Blue=0% difference, Green=20% difference, Beige=40% difference, Purple=60% difference, Yellow=80% difference.

5.4.2 fMRI results

Temporal voice areas

The TVA identified by the independent functional localiser were located as expected along the superior temporal gyrus (STG) and the STS. Two spherical regions of interest (10mm in radius) were created using MarsBar, centred around the peaks of right and left voice-selective activation, which were located in the bilateral superior temporal gyri.

Fusiform face area and face-selective regions

The face-selective regions identified by the independent functional localiser were located as expected in the bilateral fusiform gyrus (FG) and bilateral occipital gyrus (OG); again, two spherical regions of interest (10mm in radius) were generated, each centred around the peaks of right and left face selective activation, which were located in the bilateral FG (specifically, the FFA). At a slightly more liberal threshold than specified, the right STG/STS also emerged as a face selective region.

Brain regions	*Coordinates (mm)*			*k*	*t-statistic*
	x	*y*	*z*		
	a) Voice-localiser				
Superior temporal gyrus (STG)	63	-28	4	150	12.00
STG	-57	-25	4	133	11.80
	b) Face-localiser				
Fusiform gyrus (FG)*	42	-58	-17	97	11.57
Inferior occipital gyrus (IOG)	42	-82	-8	62	10.68
FG*	-36	-49	-17	7	7.43
IOG	-39	-85	-8	8	7.16
STS/STG**	45	-43	16	84	4.62

Table 5.1. Results from functional localiser experiments. a. Results of independently contrasting vocal sounds against non-vocal sounds. b. Results of independently contrasting faces against non-face visual stimuli.

Contrasts were thresholded to display voxels reaching a significance level of $p < 0.05$ with FWE correction and an additional minimum cluster size of greater than 5 contiguous voxels. MNI coordinates and t-scores are from the peak voxel of a cluster.

*** - The bilateral FFA were chosen as the peaks to base our ROIs around.**

**** - A cluster in the right STG/STS emerged at a threshold of $p<0.001$ (uncorrected). Due to our inclusion of dynamic faces and the STS's involvement in processing the changeable aspects of faces (e.g. Haxby, 2000) we based a third spherical ROI (10mm radius) around this region.**

Direct effects of face and voice emotion

We found a main effect of face emotion morph in the supplementary motor area, middle frontal gyrus (MFG), bilateral inferior frontal gyrus (IFG), and insula (Figure 5.6a, Table 5.2a)) . There was no main effect of face emotion morph in face-selective regions as identified by our functional localiser (i.e., no overlap with face-selective regions; no significant effect as indicated by an ROI analysis (left FFA: $F=1.71$, $p=0.272$; right FFA:

F=1.07, p=0.604; right STG/STS: F=0.43, p=0.788). The regions identified in our direct effects analysis appeared to show the greatest activation in response to ambiguous face information, as compared to angry and happy faces. Generally, there was also greater activation in response to faces on the 'happy' end of the continuum, as compared to those at the 'angry' end.

There was a direct effect of voice emotion morph in the bilateral STS/STG, the right temporal pole, inferior parietal lobule, cingulate gyrus and IFG (Figure 5.6b), Table 5.2c)). The regions in the STG/STS overlapped with those identified in the independent voice localiser (ROI analysis: left TVA: $p<0.001$, $t=4.15$; right TVA: $p<0.00005$, $t=5.48$ (Table 5.2d)). In contrast to the activation in response to faces, angry voices led to the largest increase in signal, with a step-wise decrease in response as the amount of 'happiness' information in the faces increased.

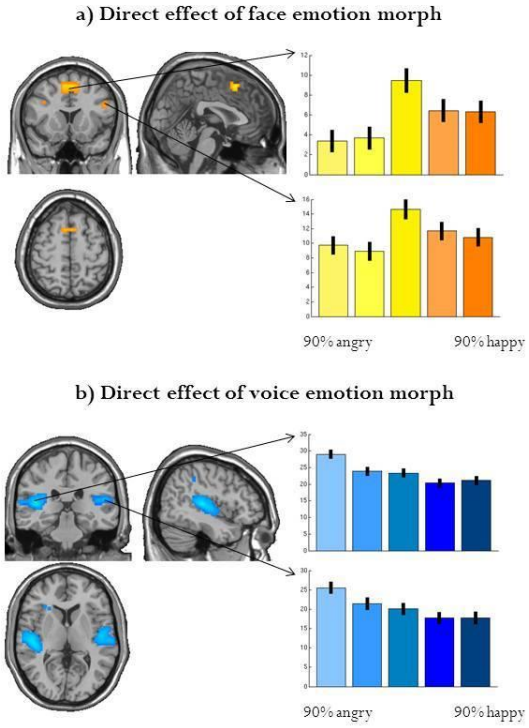


Figure 5.6. Direct effects of face and voice morph: a. Direct effect of face emotion morph. Effect of face morph plotted in the supplementary motor area and right middle frontal gyrus; b. Direct effect of voice emotion morph. Effect of voice morph plotted in the left (top) and right (bottom) STG./STS. Contrasts were thresholded to display voxels reaching a significance level of $p < 0.001$ (uncorrected) in conjunction with a cluster threshold of $p<0.05$ (FWE corrected), and an additional minimum cluster size of greater than 5 contiguous voxels. Contrasts were masked by an AV vs. baseline contrast ($p<0.001$ (uncorrected)).

Brain regions	*Coordinates (mm)*			*k*	*t-statistic*
	x	*y*	*z*		
a) Main effect of face emotion					
Supplementary motor area (SMA)	0	17	52	103	12.43
Middle frontal gyrus (MFG)	45	20	25	49	7.90
Inferior frontal gyrus (IFG)	-45	5	37	13	6.51
Insula	-30	23	4	24	6.10
IFG	45	8	40	6	6.00
b) Main effect of face morph (masked by face-selective regions)					
NO SIGNIFICANT CLUSTERS					
c) Main effect of voice emotion					
STG/Superior temporal sulcus (STS)	-45	-28	7	380	18.68
STG/STS	51	-13	1	308	11.79
STG/	51	5	-11	7	6.23
Temporal pole					
Inferior parietal lobule	-36	-37	40	9	6.04
Cingulate gyrus	-3	14	43	31	5.92
IFG	-33	26	7	9	5.35
d) Main effect of voice (masked by voice-selective regions)					
STG/STS	-45	-31	7	58	14.96
STG/STS	60	-22	4	48	9.92

Table 5.2. Direct effects of face and voice emotion morph: a,b. Main effect of face-emotion morph, masked by AV vs. baseline (a) and face-selective regions (b); c,d. Main effect of voice-emotion morph, masked by AV vs. baseline (c) and voice-selective regions (d).

Contrasts were thresholded to display voxels reaching a significance level of $p < 0.001$ (uncorrected) in conjunction with a cluster threshold of $p<0.05$ (FWE corrected), and an additional minimum cluster size of greater than 5 contiguous voxels. Contrasts were masked by an AV vs. baseline contrast ($p<0.001$ (uncorrected)). MNI coordinates and t-scores are from the peak voxel of a cluster.

Ambiguity and congruency

After removing the variance associated with congruency, a positive effect of ambiguity was found in the bilateral middle temporal gyrus (Figure 5.7, Table 5.3a)). These regions elicited more activation when the audiovisual stimulus was less ambiguous (as calculated by the specific combination of face and voice emotion morph). A negative effect was observed in the supplementary motor area, insula and precentral gyrus – here, there was greater activation for the more ambiguous types of stimuli (Figure 5.7, Table 5.3b)). After the variance associated with ambiguity values was regressed out, we found a positive effect of congruence across a wide region of the right STG/STS (Figure 5.7, Table 5.3c)). This region appeared to respond more to incongruent information, as compared to congruent.

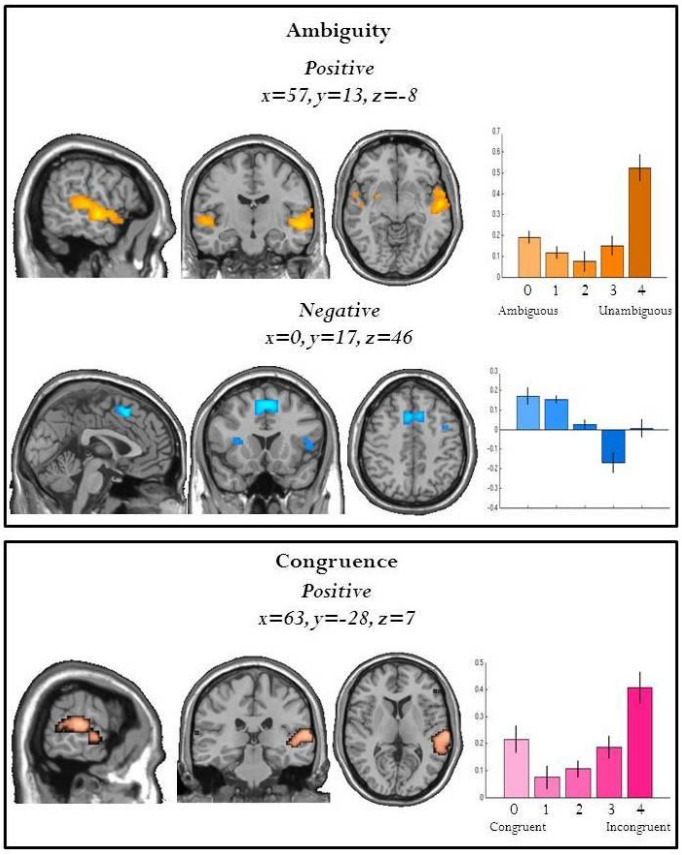


Figure 5.7. Effect of stimulus ambiguity and congruency. Contrasts were thresholded to display voxels reaching a significance level of p < 0.001 (uncorrected) in conjunction with a cluster threshold of p<0.05 (FWE corrected), and an additional minimum cluster size of greater than 5 contiguous voxels. Contrasts were masked by an AV vs. baseline contrast (p<0.001 (uncorrected)).

Brain regions	*Coordinates (mm)*			*k*	*t-statistic*
	x	*y*	*z*		
a) Ambiguity (positive effect)					
Middle temporal gyrus (MTG)/STS	57	-13	-8	313	9.14
MTG/STS	-51	-16	-5	59	6.31
b) Ambiguity (negative effect)					
Cingulate gyrus	0	17	46	136	8.92
Insula	-30	23	4	39	6.29
Precentral gyrus	-42	-1	28	36	4.58
c) Congruence (positive effect)					
STG/STS	63	-28	7	454	6.96
d) Congruence (negative effect)					
NO SIGNIFICANT CLUSTERS					

Table 5.3. Effects of stimulus ambiguity and congruence: a,b. Positive and negative effects of ambiguity value of stimulus; c,d. Positive and negative effects of congruence value of stimulus. Contrasts were thresholded to display voxels reaching a significance level of p < 0.001 (uncorrected) in conjunction with a cluster threshold of p<0.05 (FWE corrected), and an additional minimum cluster size of greater than 5 contiguous voxels. Contrasts were masked by an AV vs. baseline contrast thresholded at p<0.001 (uncorrected). MNI coordinates and t-scores are from the peak voxel of a cluster.

Adaptation ('continuous carry-over')

a) Unimodal adaptation

i) Face

We observed significant effect of face-to-face physical difference in a number of regions, including the bilateral FG, and the right cuneus, middle temporal gyrus (MTG) and inferior occipital gyrus (IOG) (Table 5.4a)). The observed effect overlapped with the independently localised face-selective regions: in these regions, the smaller the difference

in emotion of two consecutive faces, the lower/smaller was the BOLD signal and vice versa (ROI analysis: left FFA: $p<0.0.0001$, $t=5.08$; right FFA: $p<0.005$, $t=3.53$; right STG/STS: $p<0.05$, $t=2.35$) (Figure 5.8a), Table 5.4b)).

ii) Voice

We observed significant vocal repetition suppression effects in temporal and frontal regions, namely the bilateral STG/STS, and middle and inferior frontal gyri (Table 5.4c)). The observed effect in the STS overlapped with the independently localised voice-selective regions (ROI analysis: left TVA: $p<0.001$, $t=4.15$; right TVA: $p<0.00005$, $t=5.48$) (Figure 5.8a), Table 5.4d)).

Brain regions	*Coordinates (mm)*			*k*	*t-statistic*
	x	*y*	*z*		
a) Adaptation to face emotion					
Putamen	-21	8	10	70	7.46
FG	30	-52	-23	161	6.40
FG	-39	-52	-14	158	6.00
Cuneus	18	-97	16	114	5.52
Postcentral gyrus	-36	-28	43	74	5.43
Cingulate gyrus	-9	8	46	43	4.79
IOG	39	-76	-11	56	4.27
b) Adaptation to face emotion (masked by face-selective regions)					
FG	-36	-49	-17	7	5.90
FG	36	-40	23	41	5.95
IOG	39	-76	-11	20	4.27
c) Adaptation to voice emotion					
STG/STS	66	-28	1	404	7.98
STG/STS	-60	-37	7	303	7.27
Medial frontal gyrus	12	17	49	130	6.41
IFG	48	23	22	176	6.32
IFG	-39	5	37	177	6.12
d) Adaptation to voice emotion (masked by voice-selective regions)					
STG/STS	54	-22	1	137	7.98
STG/STS	-60	-37	7	92	7.37

Table 5.4. Unimodal adaptation results: a,b. Adaptation to face emotion, masked by AV vs. baseline (a) and face-selective regions (b); c,d. Adaptation to voice emotion, masked by AV vs. baseline (c) and voice-selective regions (d). Contrasts were thresholded to display voxels reaching a significance level of $p < 0.001$ (uncorrected) in conjunction with a cluster threshold of $p<0.05$ (FWE corrected), and an additional minimum cluster size of greater than 5 contiguous voxels. MNI coordinates and t-scores are from the peak voxel of a cluster.

b) Crossmodal adaptation

No crossmodal carry over effects were observed at the given threshold; however, at more liberal threshold of p<0.001 (uncorrected) with no cluster thresholding, one crossmodal carry over effect (voice-to-face physical distance) was observed in the posterior part of the STS (pSTS) (Figure 5.8b), Table 5.5b)). This effect did not overlap with either the face-selective or voice-selective regions obtained from our localisers.

Brain regions	***Coordinates (mm)***			***k***	***t-statistic***
	x	***y***	***z***		
a) Face-to-voice emotion adaptation*					
NO SIGNIFICANT CLUSTERS					
b) Voice-to-face emotion adaptation*					
STS	66	-43	7	10	4.15

Table 5.5. Crossmodal adaptation results: a. Face-to-voice adaptation; b. Voice-to-face adaptation. Contrasts were thresholded to display voxels reaching a significance level of p < 0.001 (uncorrected), and an additional minimum cluster size of greater than 5 contiguous voxels. Contrasts were masked by AV vs. baseline. MNI coordinates and t-scores are from the peak voxel of a cluster.

*** - This effect was not found at the set threshold of p<0.001 (uncorrected) in conjunction with a cluster threshold of p<0.005 (FWE corrected).**

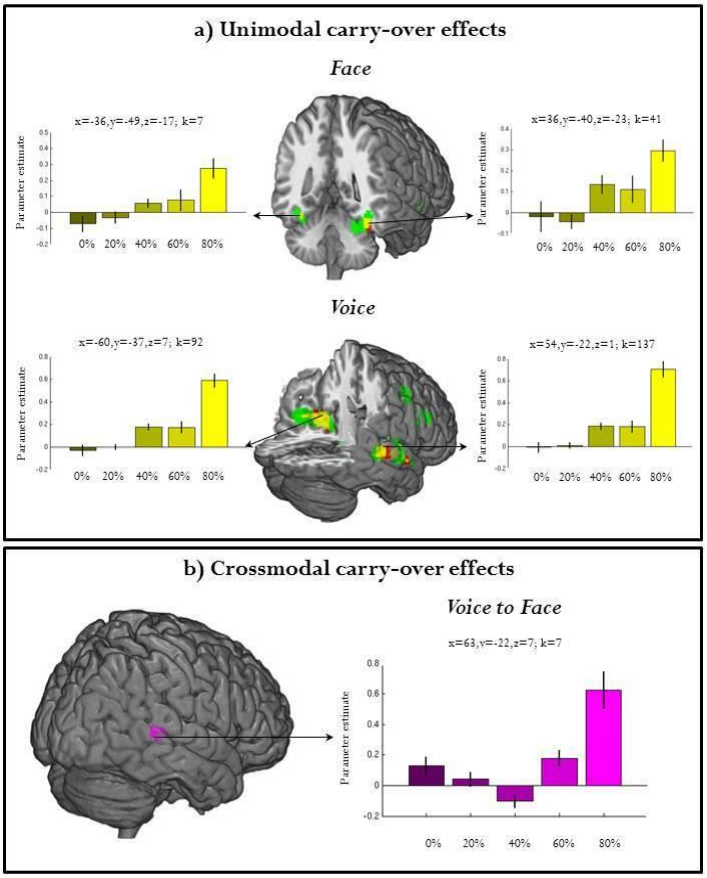


Figure 5.8. Unimodal and crossmodal carry-over effects: a. Unimodal carry-over effects. Green=Effect of face-to-face (top)/voice-to-voice (bottom) physical distance, Red=Regions localised by respective functional localiser, Yellow=overlap of experimental and localiser results.

Contrasts were thresholded to display voxels reaching a significance level of $p < 0.001$ (uncorrected) in conjunction with a cluster threshold of $p<0.05$ (FWE corrected), and an additional minimum cluster size of greater than 5 contiguous voxels. Contrasts were masked by an AV vs. baseline contrast ($p<0.001$ (uncorrected)).

b. Crossmodal carry-over effects (voice-to-face physical distance). Contrast was thresholded to display voxels reaching a significance level of $p < 0.001$ (uncorrected), and an additional minimum cluster size of greater than 5 contiguous voxels. Contrasts were masked by an AV vs. baseline contrast ($p<0.001$ (uncorrected)).

5.5. Discussion

We used visual and auditory morphing technologies to generate a range of face-voice stimuli parametrically varying in emotion, in conjunction with a continuous carry-over design so to examine the cerebral correlates of face-voice affect perception. Our main aim was to investigate the multimodal representation of emotion, in particular by searching for crossmodal adaptation effects occurring at the neural level. We supplemented this investigation by also examining the neural response to emotional congruency/incongruency in the face and voice, and exploring cerebral interactions between face and voice emotion.

5.5.1 Face-voice emotion behavioural effects

Categorical data

We observed a significant main effect of both face and voice emotion, indicating that – at least, at some points of the combined emotion space – participants took account of both sources of information to form a unique decision on emotion. However, the effect of face on categorisation was far larger, underlining that generally, participant used this modality more when categorising emotion. The significant interaction between these modalities highlighted that the observed effect of one modality depended on the morph step of another modality: for example, it was at the ambiguous points of the face and voice continua where information from the other mode had the largest effect on categorisation.

Previous behavioural work has shown that face and voice affective information can interact so to alter the perception of emotion. For example, in one of the first studies in this field, de Gelder and Vroomen (2000) found that affective ratings of facial stimuli in a morphed

continuum between two facial expressions were influenced by the concurrent presentation of a voice spoken in an affective tone. Furthermore, the authors found that participants were unable to ignore concurrently presented information in another modality, even when explicitly instructed to. Ethofer et al. (2006) also found that participants rated fearful and neutral facial expressions as being more fearful when presented in the presence of a fearfully spoken sentence, as compared to a condition where no voice was presented.

Due to the parametric morphing employed in *both* modalities we were able to develop on this work by investigating how the two modalities interacted within an extensive 3-D affective space. In their condition which did not direct to attention, de Gelder and Vroomen (2000) only used a morphed continuum of faces, and thus, it is unclear whether what the comparable effect of face on voice would be. It should be noted that in their condition which directed the attention to voice, the authors used a morphed continuum of voices. However, as the authors acknowledge, due to technical problems they could not develop a happy-sad continuum that would have been the natural counterpart of the face-continuum used in their other experiments, and had to use a more easily obtained continuum extending from happiness to fear. They found that face also had a significant effect on categorisation of the voice, which they took as indication that cross-modal biases between voice and face expressions were to a large extent bidirectional. However, their conclusions could not compare how the modalities interacted when attention was not manipulated. Our results show that although the modalities interacted, the interaction was not symmetrical.

Reaction time data

Effect of Degree of Morph

As with categorical data, we also observed main effects of both face and voice on reaction time, in addition to a significant interaction. The effect of face was far more pronounced, as can be seen by the general inverted 'U' shape of reaction time data along the face morph continuum, which remained relatively consistent regardless of the voice the face was paired with. However, at some points of the face continuum the voice did exert more of an effect, indicated by the significant interaction between the two modalities. Most noticeably, this was at the end points of the face morph continuum, where an incongruent voice led to an increase in reaction time. This was particularly apparent when an angry voice was paired with a happy face.

Researchers have generally compared bimodal to unimodal conditions when studying reaction time, mostly observing *faster* categorisation of emotion when expressed congruently in the face and voice, in comparison to one modality alone (e.g. Giard and Peronnet, 1999; de Gelder and Vroomen, 2000; Massaro and Egan, 1996; Ethofer et al., 2006). We were unable to compare conditions in this way, due to no inclusion of unimodal stimuli. However, we were still able to examine how different combinations of face and voice information (particularly, incongruence vs. congruence) affected reaction times. We observed that incongruence between the face and voice led to a general increase in reaction times (particularly at the end points of the face/voice continua), a result also observed in other studies (e.g. de Gelder and Vroomen, 2000; Collignon et al., 2008).

Effect of Physical Distance

Unimodal

Furthermore, we also observed a significant influence of context on the perception of face-voice emotion. This was indicated firstly by changes in reaction times according to the physical difference between consecutive stimuli of the same modality. We found that both face-to-face and voice-to-voice physical distance elicited a significant adaptation effect: however, notably, the physical distance between consecutive faces, and the distance between consecutive voices, had somewhat different influences on reaction times. For unimodal face adaptation, overall the smallest difference in physical difference resulted in a lengthening in reaction time in comparison to the largest differences (with the notable exception of when the voice physical difference was 0%). This is somewhat surprising, as we might have expected being previously exposed to a certain morph might have facilitated the consequent categorisation of that morph. Generally, repetition priming in face processing has been shown to facilitate the response to one stimulus following prior exposure to an identical (repetition) or related (associated) stimulus (e.g. Bruce and Valentine, 1985; Ellis et al., 1997; Ellis et al., 1990; Johnston et al., 1996)). However, in our experiment, reaction time generally peaked at the intermediate differences in morph (i.e., a 50% difference between consecutive face morphs).

Yet, it should be noted that there was a significant interaction between face and voice emotion morph. When the smallest physical difference in face morph (i.e., when the same face morph was presented consecutively), was paired with a large physical difference in *voice morph*, the reaction time increased (the larger the physical distance in voice morph, the longer the reaction time). For a face physical distance of 0% paired with a voice physical distance of 0%, reaction times showed an inverted 'U' shape, as opposed to when the same face physical distance paired with a voice physical distance of 80%, where they

presented a reverse sigmoid shape curve. This effect of voice physical distance was not so apparent at the other end of the face physical distance continuum: whether the voice physical difference was large or small, the mean reaction time generally remained constant and low.

With regards to voice adaptation, the general effect appeared to be almost the reverse to that of unimodal face adaptation: there was a steady increase in reaction times that paralleled an increase in voice physical difference. This indicates some difference in face and voice processing in this experiment, in that context affected the two modalities differentially. However, again it needs to be kept in mind that participants were presented with bimodal face-voice stimuli, and that there was a significant interaction between the two modalities. The significant interaction can also be seen when looking at the response caused by voice physical distance: for both large and small physical differences in consecutive voices, the closer the face physical difference was to that for voices, the shorter the reaction time was. Generally, this interaction provides evidence that effect of unimodal adaptation on reaction time is still in some way dependent on the physical distance between both consecutive face and voice values.

A recent study (Charest et al., 2012), investigating voice gender perception, observed a significant influence of context on the perception of voice gender indicated by changes in reaction time according to the physical difference between consecutive stimuli. Our results lend further support that the physical difference between different stimuli can either facilitate or hinder categorisation, whether this is gender or emotion perception. However, it should be noted that our results highlighted a somewhat different pattern of results from the aforementioned study. Generally, our finding was that the larger the difference between two consecutive voice morphs, the longer it took to categorise the stimulus. In contrast,

Charest et al. (2012) observed an inverted 'U' effect of physical distance, with the middle physical distances resulting in the largest increase in reaction time. The reasons for this may simply be both stimulus and task driven: we were investigating emotion, not gender perception; and our stimuli were not unimodal by nature, but rather face-voice stimuli. Thus, any unimodal effects we observed were always within a multimodal context – indeed, our two unimodal adaptation effects interacted. Further work should investigate the influence of context across a range of tasks and stimuli, in order to reach a more definite conclusion on the effect of context when looking at paralinguistic priming effects.

Crossmodal

Across modality, there was also a significant interaction between the two crossmodal effects. Generally, when face-to-voice and voice-to-face physical distance were congruent, reaction time was lower, and vice versa. However, at some points in the physical difference continuum there were exceptions to this rule. For example, when the distance between one voice morph and the next face morph was 80%, the other crossmodal effect (i.e., face-to-voice physical distance) appeared not to influence reaction time: reaction time consistently remained low. Overall, however, there was a main effect of voice-to-face physical distance but not of face-to-voice physical distance. This indicates some asymmetry in modality priming in this experiment, with voice apparently exerting a stronger priming or adaptive effect.

Very few studies have investigated cross-modal priming in the context of face-voice perception, and to our knowledge, none within the field of emotion perception. In one pioneering study, Ellis et al. (1997) observed that over short time-intervals, the presentation of a familiar voice-prime followed immediately by a face of corresponding identity resulted in a significant improvement in performance. Similar results were

demonstrated for face primes in relation to voice test stimuli. Furthermore, in a recent experiment, Hills et al. (2010) tested the magnitude of the face identity after-effect following adaptation to four modes of adaptors: faces, voices, names and occupations. The perceptual midpoint between two morphed famous faces was measured pre- and post-adaptation, and significant after-effects were observed for visual (faces) but also non-visual adaptors (voices and names). Our results build on this work by showing crossmodal adaptation effects (particularly the effect of voice on face) exist in the context of face-voice emotion perception, and furthermore that crossmodal priming effects can interact.

Effect of ambiguity and congruence

We also investigated the effect of both stimulus ambiguity and congruence on reaction time. Values for each of these dimensions were assigned based on where each stimulus lay in the 5x5 emotion space, where congruence value related to the degree of discordance between the emotion displayed in the face and voice, whereas ambiguity values referred to the clarity of the affective information in the *combined* stimulus. These values were also negatively correlated.

We observed a significant effect of stimulus ambiguity on reaction time, with the more ambiguous stimuli taking longer to categorise. However, there was no significant effect of stimulus congruence. In similar studies it has been observed that generally, the greater the incongruence between face and voice, the more time it takes to classify the emotion (e.g. de Gelder and Vroomen, 2001, Massaro and Egan, 1996). However, it should be noted that in our study, due to the parametric morphing and resultant combinations of face-voice emotion, that some stimuli with a small or no degree of incongruence would still have proved difficult for our participants to categorise – for example, those that had a pairing of

ambiguous information in both the face and the voice. Thus, in this study it is unsurprising that the level of stimulus ambiguity was more reflective of task difficulty.

5.5.2 Neural representation of facial emotion

We observed a significant main effect of face emotion morph in a number of regions, including the supplementary motor area (SMA), MFG, bilateral IFG and insula. In these regions, the strongest response appeared to be for ambiguous face information – that is, when there was a 50% mix of anger and happiness. This response profile is consistent with these regions' involvement in representing task difficulty. Heekeren et al. (2008) describes both the anterior insula and IFG as part of a system within a model of human perceptual decision making, responsible for detecting perceptual uncertainty or difficulty. They form this proposal on the basis of results from a number of studies (Binder et al., 2004; Thielscher and Pessoa, 2007; Grimband et al., 2006; Heekeren et al., 2006) which have found that in these regions the BOLD response was greater during difficult than easy trials (as indicated with a positive correlation with reaction time). The authors also suggest that these regions perform a further role in bringing to bear additional attentional resources in order to maintain accuracy in decision making when the task becomes more difficult. There is also evidence that the observed region of SMA activation is involved in performing complex tasks, and early learning stages, as opposed to the performance of simple or overlearned tasks (Picard and Strick, 1996; Fujii et al., 2002)

Additionally, there was generally a stronger response to happy information in these regions, as compared to angry (although perhaps not always statistically significant). While outside the basic model of face perception proposed by Haxby et al. (2002), several studies have shown that the IFG can be activated by expressive face processing (Carr et al., 2003; Montgomery and Haxby, 2008; Fusar-Poli et al., 2009). The MFG has also been shown to

be activated by facial expressions of emotion, and may play a role in emotion regulation (Eippert et al., 2007; Ochsner et al., 2002), via the orbitofrontal cortex (OFC), and its dense connections with the amygdala (McDonald, 1998). Fusar-Poli et al. (2009) suggest that is possible to speculate that while areas of the visual cortex are engaged in early perceptual processing of facial stimuli – potentially independent from emotional valence – the prefrontal cortex, on the other hand, participates in the conscious experience of emotion, inhibition of excessive emotion, or monitoring one's own emotional state to make relevant decisions.

Neurally, we found significant adaptation effects which mimicked those seen behaviourally. Unimodal face adaptation was observed in a network of regions, including the putamen, cuneus, cingulate gyrus, IOG and bilateral FG. All of these regions have previously been implicated in processing of facial affective information. For example, a recent meta-analysis (Fusar –Poli et al., 2009) of results from 105 fMRI studies linked processing of emotional faces to increased activation in the putamen; in particular, that of happy faces. The cuneus has been linked to tasks related to controlling attention, and the attentive processing of expressed emotions (Sreenivas et al., 2012) and the ventral anterior cingulate gyrus is important for autonomic function and emotional behaviour: indeed, the cingulate gyrus is an integral part of the limbic system, and has direct connections to the amygdala (Damasio, 1994).

An interesting result is that of the 'face-selective' regions – namely, the bilateral FG. There was a significant face adaptation effect within our face-selective regions identified using our separate functional localiser, including the bilateral FFA and right IOG, but not a direct effect.

In a multitude of studies, reliable BOLD signal in basic face processing areas has been evoked by emotional face perception. The FG, particularly the anterior region as it nears the parahippocampal gyrus, has been consistently associated with the perception of human faces (Haxby et al., 2000; Puce et al., 1995; Kanwisher et al., 1999), and has been shown to be more active during expressive (e.g., fearful) face processing than neutral faces (Morris et al., 1998, Vuilleumier et al, 2001, 2004). In a recent meta-analysis (Sabatinelli et al., 2011) clusters of activity identified by an ALE analysis of emotional face perception, and included regions implicated in the Haxby model of face processing (Gobbini and Haxby, 2007; Haxby et al., 2002), specifically the anterior FG, MTG, STG, and IOG.

The fact these regions did not emerge in our analysis of direct effects could potentially be a consequence of the approach of this analysis, which constrains the search to brain regions more sensitive to faces of one emotion over another. Typically in studies investigating facial emotion processing, emotional faces have been compared to neutral faces. The fact that we directly compared faces along an angry-to-happy morphed continuum could have potentially limited the activation we observed, as only regions which were significantly more activated to one of these emotion morphs over another would be highlighted. In contrast, comparing these emotions to neutral facial expressions may have activated a more extensive network of regions, including those face-selective regions. One proposition is that face emotion representation could potentially involve overlapping neuronal populations sensitive to angry or happy faces. Assuming equal proportions of angry- and happy-sensitive neurons in a given cortical area/voxel, the subtraction of angry- versus happy-related cerebral activity (or an analysis designed to find a main effect, such as an ANOVA) would fail to highlight them. Finally, it should be noted that any observed unimodal effects – whether these were direct or adaptive – were always within a

multimodal context. Thus, it is possible that audiovisual stimulation may have played a role in altering effects that would have been observed using only unimodal stimuli.

5.5.3 Neural representation of vocal emotion

In our analysis of direct effects of vocal emotion, we found a significant main effect of voice in the bilateral STS/STG, the right temporal pole, inferior parietal lobule, cingulate gyrus and IFG. The regions in the STG/STS overlapped with those identified in the independent voice localiser.

To date, only a small number of imaging studies of emotional prosody have been reported. Most of the early studies focused on whether a right hemisphere lateralization existed for the processing of emotional prosody (e.g. George et al., 1996, Pihan et al., 1997, Imaizumi et al., 1998). Very few studies have attempted to pinpoint more specific neural circuits underlying affective voice perception. Mitchell et al. (2003) found areas of the MTG and STS that activated more when attending to affective prosody as compared with semantic content of spoken words, and Grandjean et al. (2005*)* and Sander et al. (2005) reported fMRI data that revealed a region in STS that showed greater activation in response to angry speech as compared with neutral speech. In an fMRI study of five vocal emotions, Wildgruber et al. (2005) identified a right hemispheric network consisting of the pSTS, and dorsolateral and orbitobasal prefrontal cortex that showed selective activation during an emotion recognition task. However, differential activations for the five emotions were not observed. In a following fMRI study, Ethofer et al. (2006) identified regions in the right pMTG and STS and bilateral IFG and MFG that activated more when individuals identified affective prosody than when identifying the content of the spoken words. However, again no distinction was made between responses to the different expressed emotions studied. Similarly, electrophysiological findings (Paulman and Kotz, 2008)

demonstrated that early event-related potentials differ between emotional and neutral prosody but failed to identify differences between emotions. These findings seem to suggest that processing of emotional voices within the auditory cortex might primarily reflect a discrimination between emotional and neutral stimuli only, whereas categorisation of emotions might occur at later stages; e.g., within the frontal cortex.

To date, only one study has used conventional methods to test whether specific brain regions showed preferential engagement in the processing of one emotion over the other. Johnstone et al. (2006) conducted an fMRI study to examine the brain responses to vocal expressions of anger and happiness, and found that happy voices elicited significantly more activation than angry voices in the right anterior and posterior middle temporal gyrus, left posterior MTG and right IFG, suggesting a particularly salient role for vocal expressions of happiness.

However, as Ethofer et al. (2009) note, enhanced response often occurs irrespective of the specific emotion category, making it impossible to distinguish different vocal emotions with conventional analyses (e.g. Grandjean et al., 2005; Ethofer et al., 2006; Wiethoff et al., 2008). Therefore, conventional approaches can have important limitations for determining how information is represented within cortical areas. In their study, they presented pseudowords spoken in five prosodic categories (anger, sadness, neutral, relief, joy) during event-related fMRI, then employed multivariate pattern (MVPA) analysis to discriminate between these categories on the basis of the spatial response pattern within the auditory cortex. Their results demonstrated successful decoding of vocal emotions from fMRI responses in bilateral voice-sensitive areas, which could not be obtained by using averaged response amplitudes only. For all five categories, the most informative voxels were widely distributed. On average, these maps showed an overlap with each other for

approximately 50% of the voxels, and about 25% of these voxels were included in all five maps. These common voxels were mostly situated in the mid STG, confirming the key role of this region in processing emotion in voices. Categories that were either both high arousing (i.e., anger and joy) or both low arousing (i.e., sadness and relief) exhibited a stronger overlap than did emotional categories that differed in arousal or comparisons between individual emotional categories and neutral prosody. Likewise, pairwise comparisons between categories showed the greatest confusion between emotions with similar arousal (sad versus relief, joy versus anger) but good discrimination between emotions with a similar negative valence (anger versus sad) or a similar positive valence (joy versus relief). These findings concur with psychological and neural accounts postulating that arousal is a key dimension defining different emotion categories. However, the pairwise comparisons showed that each category could be classified against all other alternatives, indicating for each emotion a specific spatial signature that generalized across speakers. These results demonstrate for the first time that emotional information is represented by distinct spatial patterns that can be decoded from brain activity in modality-specific cortical areas.

Interestingly, in our study we did find a direct effect of vocal emotion, using a form of subtraction method, as did Johnstone et al. (2006). However, our effect was the reverse of that seen in this study: we observed a generally higher activation in response to angry, as compared to happy vocalisations. One reason for this could be stimulus differences, particularly acoustic parameters such as fundamental frequency (*F0*) and timbre, along with intensity and duration. *F0 is* an important parameter for expression of emotional arousal (Scherer, 2003), and it has been observed that there is better discrimination rates between emotions that strongly differ in *F0* converge. However, it should also be noted that in their study, Ethofer et al. (2009) could differentiate between emotions with similar

F0 (e.g., anger versus joy) indicating that decoding did not depend solely on *F0.* Additionally, in their study, Grandjean et al. (2005) found that the enhanced response to angry vocalisations was unrelated to acoustic amplitude or frequency of the prosody. Indeed, recent work (Hannerschmidt and Jurgens, 2007) demonstrates that *F0* is only one of the important parameters of denoting a specific type of emotion, with other features such as timbre playing an equal function. Moreover, previous fMRI results (Wiethoff et al., 2008) showed that the activation of STG was driven mainly by intensity and duration of stimuli, more than by their *F0*.

Although we might presume that acoustic parameters would be similar for the same vocalisations across our and Johnstone et al.'s (2006) study, our stimuli were briefer, and also contained no speech information. In vocalisations, where information is contained within dynamic cycles of information, factors such as length and speech content could have caused significant differences between the two experimental sets. Furthermore, the fact we were aware our emotional vocalisations were to be morphed placed some constraints on how the actors could produce the sounds (for example, we were not able to have the actors produce a laugh – perhaps one of the most distinctive auditory indicators of happiness – as this would have resulted in a broken vocalisation which would have been unable to be morphed effectively). Only future studies using systematically manipulated stimuli might help to address the question of which parameters (or combinations thereof) are most important for recognizing a particular emotion at both behavioural and neural levels, and whether voice-selective areas are performing this emotion categorisation independent of such acoustical factors.

We also observed unimodal voice adaptation, overlapping with the functionally localised TVA, lending further support to the role of these regions in some form of emotion prosody

processing. Additionally, this activation was notably larger and stronger in the right hemisphere. Research using fMRI on emotional prosody classification has shown that the right hemisphere is particularly involved (Buchanan et al., 2000; Morris, Scott, & Dolan, 1999; Rama et al., 2001) and recent studies confirm the right liberalised activity in MTG and STG (Ethofer et al., 2006; Grandjean et al., 2005; Mitchell et al., 2003). This activation appears to be relatively independent of attentional demands (Ethofer et al., 2006) and low-level acoustic features such as frequency and amplitude of the sounds (Grandjean et al., 2005). However, additional neuroimaging studies have painted a more complex picture in which a more distributed, bilateral neural network is engaged when processing emotional prosody. Although the activity elicited in response to emotional prosody is often stronger on the right, bilateral TVA are typically active during the processing of affective compared to neutral vocalizations. Our results support bilateral vocal emotion processing, by showing that there is also an adaptive response to emotion in the voice in both the right and left TVA. Interestingly, our strong adaptive effect in the right hemisphere contrasted with results from the *direct* effects analysis, where the main effect of vocal emotion was actually stronger in the left hemisphere. This indicates there might be potential (albeit slight) differences in neural representation of vocal emotion across the two hemispheres. Finally, as with unimodal face effects, it is important to note that the aforementioned effects of voice morph (both direct and adaptive) were observed as part of a bimodal stimulation of emotion and the same effects may not have been observed when removing the simultaneous face presentation.

5.5.4 Multimodal representation of face-voice emotion: congruence, face-voice interactions and crossmodal adaptation

The overall aim of this experiment was to examine bimodal processing of face-voice emotion, particularly integration effects represented at the neural level. Our unique design allowed us to explore this audiovisual processing from a number of perspectives, including stimulus congruence effects, interactions between face and voice emotion, and crossmodal adaptation. These are described further below.

5.5.4.1 Ambiguity and congruence effects on brain activity

Parallel to our analysis at the behavioural level, we investigated the effect of stimulus congruence and ambiguity on brain activity. With regards to face-voice integration, previously researchers have proposed that a stronger effect for congruent information as compared to incongruent would represent a 'binding' of information from the two modalities (e.g. Calvert et al., 2000; Dolan et al., 2001; Klasen et al., 2011). We were also able to include stimulus ambiguity in this experiment as a more thorough indicator of task difficulty.

We found that there was an effect of both stimulus ambiguity and congruence on brain activity. We observed a positive effect of both ambiguity and congruence across the STG/STS, in addition to a negative effect of ambiguity in the anterior cingulate, insula and precentral gyrus. In these latter regions, there was heightened activation in response to ambiguous information, as compared to unambiguous information.

The cingulate gyrus (particularly, the anterior cingulate cortex (ACC)) is amongst the brain regions most frequently reported in the functional neuroimaging literature as being

significantly activated when engaging in attentionally or behaviourally demanding cognitive tasks (Paus et al. 1998). For example, Paus (2001) suggests that the outputs of cognitive processing performed elsewhere in the prefrontal cortex are combined in ACC with representations of emotional state to enable appropriate behavioural responses to internal or environmental events. A number of studies have also implicated this region in the detection of conflict between different possible responses to a stimulus, event, or situation (e.g. Carter et al., 1999; Kerns et al., 2004; Wendelken et al., 2008). In our study, stimuli of an ambiguous nature would have been more demanding to categorise, requiring more energy for decision making, and thus it is unsurprising that we observed heightened activity in the cingulate in response to this information.

The anterior insular cortex (AIC) is among the non-sensory brain regions most commonly found activated in functional brain imaging studies on visual and auditory perception. As mentioned previously, a number of fMRI studies have suggested that the AIC plays an important role in the perceptual decision process – particularly, activity in this region can be reflective of task difficulty. For example, Thielscher and Pessoa (2007) used a graded series of morphed emotional faces and asked participants to indicate the faces' emotional expression. They observed an inverted U-shaped correlation between reaction times and BOLD responses in the AIC and the ACC, using reaction time as an index of decision processes. In other words, longer reaction times, which indicate a more difficult perceptual decision, were associated with greater AIC and dorsal ACC activations. Along similar lines, difficulty of perceptual decisions was modulated by varying noise levels in an auditory discrimination task (Binder et al. 2004). While accuracy correlated positively with activity in the auditory cortex, reaction time as a marker of task difficulty correlated positively with the BOLD signal in AIC. Finally, EEG components that are related to difficulty in perceptual decision making correlate with fMRI signals in the AIC, the ACC

and dorsolateral prefrontal cortex (Philiastides and Sajda 2007). AIC activity related to task difficulty could hence reflect the degree of cognitive effort that is required for a task. In our study, ambiguity values assigned to our stimuli were significantly correlated to reaction times, and therefore can be seen as an index of task difficulty. Thus, our results provide yet further evidence that activity in the insula is a reflection of task demand. Craig (2009) also notes that the great majority of studies reporting AIC activation also report activation of ACC. There is now a wealth of evidence that anterior insular and anterior cingulate cortices have a close functional relationship, such that they may be considered together as input and output regions of a functional system.

A positive effect of ambiguity and congruence was seen across the MTG/STG and STS, bilaterally, although this activation was far greater in the right hemisphere. This region has been implicated in auditory-visual multisensory integration for both speech and non-speech stimuli (Calvert et al., 2000; Sekiyama et al., 2003; Beauchamp, 2005; Miller & D'Esposito, 2005). In our study, within overlapping regions there was an increase in activation in response to stimuli that were by nature less ambiguous, and interestingly, also an increase in response to stimuli that were classified as more incongruent.

With regards to congruency, we might have expected that this pattern would be the reverse, for two reasons. Firstly, the congruence and ambiguity values assigned to our set of stimuli were negatively correlated and so it would be reasonable to presume that where there was an increase in activation to less ambiguous stimuli, there might be an increase in activation in response to stimuli that were congruent in nature. Secondly, the initial claims for the STS as an audiovisual binding site came from Calvert et al. (2000) who contrasted audiovisual speech to each modality in isolation (i.e., heard words or silent lip-reading). This revealed a super-additive response (i.e., a heightened response relative to the sum of

the responses of audio and visual speech information presented alone) in the left pSTS when the audiovisual input was congruent but a sub-additive response when the audiovisual input was incongruent (i.e., showing a reduced response relative to the sum of the responses of audio and visual speech information presented alone).

However, it should be noted that a number of studies have produced conflicting results. Indeed, Hocking and Price (2008) stated that at that time they were unable to find any studies that replicated the Calvert et al. (2000) study showing enhanced pSTS activation for congruent relative to incongruent bimodal stimuli. In an fMRI study of the McGurk effect conducted by Jones and Callan (2003) greater responses in the STS/STG for congruent audiovisual stimuli were not observed over incongruent audiovisual stimuli, as one would predict for a multisensory integration site. A recent imaging study investigating emotional incongruence (Muller et al., 2010) also did not observe a greater effect of congruent affective information over incongruent information in this region, although they found that incongruence of emotional valence in audiovisual integration activated a cingulate-fronto-parietal network, related to error detection and conflict resolution.

Hocking and Price (2008) suggest that potentially, one reason for the inconsistent congruency effects could be due to the fact that attention to one modality only during bimodal presentation elicits sub-additive effects (Talsma and Woldorff 2005; Talsma et al. 2007). They argue that to minimise interference during incongruent audiovisual speech streams, participants may automatically or attentionally reduce visual processing (Deneve and Pouget 2004; Ernst and Bulthoff 2004), particularly in the study of Calvert et al. (2000) where congruent and incongruent conditions were presented in separate experiments with no instructions to attend to the visual stimuli. This would explain the

absence of congruency effects in studies that presented brief stimuli or forced participants to attend to the visual input during incongruent audiovisual conditions.

Hocking and Price (2008) found that when task and stimulus presentation were controlled, a network of regions, including the pSTS, were activated more strongly for incongruent than congruent pairs of stimuli (stimuli were colour photographs of objects, their written names, their auditory names and their associated environmental sounds). They suggest that activation reflects processing demand which is greater when two simultaneously presented stimuli refer to different concepts (as in the incongruent condition) than when two stimuli refer to the same object (the congruent condition). They also hypothesise that if participants were able to attend to one input modality whilst suppressing the other, then pSTS activation would be less for incongruent bimodal trials. In contrast, if subjects were forced to attend to both modalities then the pSTS activation would be higher for incongruent bimodal trials that effectively carry twice the information content as congruent trials.

In our study, a key point should be noted: values assigned to stimuli (specifically, those indicating congruence) on the basis of the face and voice morph information were not necessarily reflective of the perceptual difficulty of classifying emotion. Specifically, although congruence and ambiguity values were significantly correlated, only ambiguity values were correlated with reaction times. Thus, although some congruent stimuli resulted in shorted reaction times (e.g., 10% angry face-10% angry voice), some did not (i.e., 50% angry face-50% angry voice). Therefore, we can suggest that the heightened response to incongruent information across the right STS was possibly not due to the perceptual difficulty of classifying the stimulus or processing demand. In our study, participants were instructed to attend both modalities. Although we cannot be sure that participants definitely

attended to both modalities in the incongruent trials, our behavioural data does suggest that although participants did place greater weighting on the visual modality, they did integrate the two modalities to some degree (indicated by a significant interaction between face and voice emotion morph for both categorical and reaction time data, in addition to a main effect of both modality). Therefore, in line with the proposal of Hocking and Price (2008), a tentative explanation is that participants were attending to both modalities and thus the STS activation was higher for incongruent bimodal trials. This does not necessarily imply greater perceptual difficulty, but rather just some recognition that the auditory and visual inputs were different.

Overall, the positive response to unambiguous information could be related to this information providing a clear emotional percept. This would fit in line with studies which have shown an increased response to congruent vs. incongruent information (although, as previously described, there is still conflict within this literature), where congruent information offers an unambiguous and clearer representation of emotion. In our study, congruent information did not always provide a clear indication of audiovisual emotion, even though the morph values in the face and the voice were identical. This suggests that unambiguity, rather than congruence per se, might be more related to activation in the STS.

Finally, Klasen et al. (2011) argue that incongruent emotional information cannot be successfully integrated into a bimodal emotional percept, and propose that regions responding more to congruent information than incongruent is reflective of some manner of integrative process. However, Belin and Campanella (2007) suggested that conversely, it may be possible for incompatible affective information in the face and voice to be combined in such a way as to create an entirely new emotional percept, one independent of information contained in either modality – an ‘emotional McGurk’ effect. This would

imply some form of audiovisual integration, although perhaps one with a nature and mechanisms entirely different from the integration of emotionally congruent information. We are far from being able to conclusively answer this question; nonetheless, our results point to a strong activation in the STG/STS region in response to incongruent information, that cannot be explained simply by task difficulty or processing demand.

5.5.4.2 Face-voice interactions

In our analysis of direct effects we also looked for any interactions between the main effect of face emotion morph and voice emotion morph expressed at the cerebral level. Such an effect would imply that the neural response to emotion expressed in one modality depended on the emotion expressed in the other modality; indeed, this result was observed in our analysis of behavioural data. In the previously described study by Johnstone et al. (2006), the authors observed a significant interaction between facial and vocal emotion in the left MTG: specifically, a happy–angry vocal emotion contrast was significantly greater when vocal expressions were accompanied by happy facial expressions than when accompanied by angry facial expressions. However, we did not observe such an interaction, even at a more liberal threshold. Perhaps future work would involve searching for such an interaction within anatomically defined and well documented multisensory convergence zones such as the pSTS, amygdala or thalamus; and using a technique such as MVPA to uncover an effect that might not be seen using the relatively conventional analysis in our study. Limiting the search to a restricted set of voxels might have uncovered an interaction that cannot be seen when comparing between *all* the voxels activated by audiovisual stimulation, as were included in our mask.

5.5.4.3 Crossmodal adaptation

Finally, we investigated whether neurally, one modality could adapt the response to information presented in the other modality: a crossmodal adaptation effect. Overall, this was the main focus of our experiment. As previously mentioned, a small number of behavioural experiments have observed that information in the face can prime response to that in the voice, and vice versa, but this has never been explored at the neural level – in any context, emotion processing or otherwise. So far, fMR-A has successfully been used to explore many important issues in unisensory processing in humans, but has only once been used to study multisensory integration (Tal and Amedi, 2009), and this study investigated visuo-haptic perception. Thus, we used this multi-sensory paradigm to investigate whether fMRI adaptation would indicate multisensory voxels showing multisensory adaptation and to further speculate upon neuronal populations past the voxel-level.

At a slightly more liberal threshold than initially set, we observed a crossmodal adaptation effect in the right pSTS – a region which has been well documented as a multimodal region, both in humans (e.g. Beauchamp et al., 2004; Beauchamp, 2005; Kreifelts et al., 2007, 2010; Ethofer et al., 2006) and non-human primates (Barraclough et al., 2005; Benevento et al., 2007; Bruce et al., 1981). Significantly, this effect was independent of any variance elicited by either of the unimodal carry-over effects: we regressed out both unimodal face and voice physical distance values, ensuring that only the variance associated with crossmodal adaptation was modelled.

This result raises interesting questions regarding the nature of neuronal populations in the pSTS. We hypothesised that in regions where there were ‘true’ multisensory neurons – in other words, where information from each modality synapsed on the same neurons - presenting a particular morph followed by the same morph in the other modality would

result in the suppression of the activation of these neurons, and hence lead to a reduced fMRI signal. Similarly, presenting a morph followed by a different morph in the other modality would result in a greater activation of these multisensory neurons, and result in a heightened signal. This would be in comparison to inter-digitised groups of neuronal populations, each tuned to the morph exposure in the visual modality or the auditory modality. If this were the case, we would expect that there would be no crossmodal effect – information from each modality would activate separate groups of unisensory neurons (albeit within the same region), and the result would be a non-adapted fMRI signal, one that did not differ as a result of physical differences between morph steps of different modalities. Our results seem to provide a strong indication of multisensory neurons in the right pSTS.

Interestingly, the observed crossmodal effect was asymmetrical – activity in this cluster was driven by the difference between a voice and the following face, and not the difference between a face and the following voice. Therefore, it appears that voice exerted a stronger adaptive effect on face, than face did on voice. This result is intriguing, especially as face-morph exerted most of the modulating effects on both emotion categorisation and reaction times. However, it should be noted that this result at the neural level paralleled the results seen behaviourally: here, only voice-to-face physical difference significantly modulated reaction time values (although there was a significant interaction between the two crossmodal effects).

Furthermore, one might presume that if a neuron was multisensory and therefore coding for both stimulus dimensions, both voice-to-face physical difference *and* face-to-voice physical difference would have exerted similar effects on its response. We can only speculate as to the reasons why this was not the case. One proposal is that visual and

auditory inputs could synapse on the same neuron, but have differential modulating effects or weighting on the neural response. Another could be that this multisensory region contains multisensory neurons *intermixed* with unisensory neurons. In this case, a cross-modal repetition (e.g., visual–auditory) would suppress activity of multisensory neurons, but at the same time activate new pools of unisensory neurons in the same voxel, which could counteract the cross-modal suppression. If this area contained an unequal mix of unisensory neurons (i.e., more auditory than visual) this could potentially explain this asymmetrical crossmodal effect.

This study highlights the power of using fMR-A (specifically, the continuous carry-over paradigm) to study multimodal integration in humans, and may lead us to a better understanding of neural organisation and functional properties of cortical neurons beyond what standard imaging techniques can achieve. However, given the complexity of the neural processes and the fact these measurements are still only indirect, results from this type of design should always be treated with caution. Researchers using fMR-A must take into account a number of important issues, especially in a multisensory context.

Typically, a main concern is that different populations in a multisensory region may be selective to 'features' in different conditions: visual repetitions may adapt visual and audiovisual neurons, auditory repetitions may adapt auditory and audiovisual neurons. This is in contrast to, for example, the visual system where different neuronal subpopulations are selective to one specific feature (e.g., maximum response to a 90% angry affective morph) or are not selective (e.g., emotion-invariant). In regular fMR-A, this can be problematic because neurons are shown to adapt despite intervening stimuli (Grill-Spector, 2006), so stimulus repetitions in alternating modalities will also adapt unisensory neurons (although probably to a weaker extent). Due to our unique design – where bimodal stimuli

were presented in an unbroken stream – we were able to counteract this effect somewhat: not only were we able to examine crossmodal effects but also unimodal adaptation, albeit within a multisensory context. Therefore, as previously mentioned, we were able to remove the variance associated with unimodal adaptation within this region and focus on the effects that were only crossmodal. At the same time this was a particularly stringent analysis: it is possible that removing any shared variance between unimodal and crossmodal effects could have reduced what might have been previously stronger between modality adaptation. Nonetheless, we felt it was appropriate to regress out unimodal adaptation in order to be sure any observed effect was crossmodal, and the fact that a significant response remained is striking.

Finally, care should be exercised in mapping between neural firing and hemodynamic response: interpreting the meaning of BOLD changes in adaptation paradigms such as this is far from simple. For example, Tal and Amedi (2009) note that local field potentials (LFPs) have been shown to correlate with the BOLD signal better than multi- or single-unit activity in the macaque monkey (in Logothetis et al., 2001). Furthermore, as mentioned in **Chapter 2**, adaptation may reflect a proportional reduction in firing rate to repetitions of a specific stimulus, a change in the tuning of neural responses to repeated stimuli, or shortening of the processing time for repeated stimuli (reviewed in Grill-Spector et al. 2006) and fMR-A cannot differentiate these three forms of face-voice integration. Altogether, it is clear that the link between fMR-A and neuronal tuning is far from straightforward. Only direct measurements such as single-cell/unit studies will allow for a more definitive interpretation as to the activity of multisensory neurons. Nonetheless, if treated with the appropriate care and consideration, fMR-A and the continuous carry-over design could represent a step forward in our understanding of not just representation of face-voice emotion, but multisensory integration in general.

5.6 Conclusion

Within the context of face-voice emotion perception, we observed a crossmodal adaptation effect in the right pSTS. This adaptation effect was also found at the behavioural level. Furthermore, the role of the STS in audiovisual emotion processing was also inferred by a strong response to parametric variation of face-voice congruency. This work extends upon previous behavioural research which has evidenced crossmodal response priming in face-voice identity perception, and other studies which have implicated the pSTS as playing an important role in integrating affective information from the face and the voice.

6. General Discussion

The main aim of this thesis was to explore interactions between face and voice paralinguistic processing pathways, and the subsequent integration of this information. Such interactions were proposed as part of an early model of voice processing (Belin et al., 2004; see also Belin et al., 2011) but have mostly been investigated within the context of speech perception: evidence of integrative effects during person perception has, until recently, been relatively sparse and preliminary (Campanella and Belin, 2007). Person perception research has generally concentrated on single modalities and, furthermore, has mainly used static face presentations in showing crossmodal effects. Focussing on this thesis as a whole, I suggest there are three main conclusions to be drawn:

1. Audiovisual integration plays a notable role in person perception: we integrate paralinguistic information from the face and voice at both a perceptual and neural level, thus supporting the described interactions in the Belin et al. (2004) model and other experimental findings in this young field;
2. This integration is not always an equal combination of two information sources: one modality can appear to dominate over the other, cause differential modulating effects, and this seems to depend in part on the face-voice processing pathways under question (e.g. gender, emotion) and probably stimulus selection;
3. The right posterior superior temporal sulcus (pSTS) appears to play a crucial role in integrating paralinguistic information from the face and the voice, and may selectively integrate information from these modalities over information which is not person-specific.

In this General Discussion each of the three experiments within this thesis will be reviewed in turn, allowing for expansion of the main conclusions detailed above, in addition to consideration of experiment-specific conclusions.

6.1 Conclusions from Chapter 3

In the first experiment, we conducted a broad investigation of neural face-voice integration: we searched for brain regions which combined these two sources of information, under passive conditions, with no specific emphasis on linguistic, or one type of non-linguistic processing. I believe this offered a natural starting point for the more specific work in **Chapter 5**, which examines neural processing of face-voice emotion.

Participants were scanned using functional magnetic resonance imaging (fMRI) while they were presented with auditory, visual, or audiovisual stimuli of people or objects, with the intention of localising areas that were 'people-selective', regardless of modality; audiovisual regions designed to specifically integrate person-related information; and also areas where information from the two modalities might converge, but not necessarily bind together (i.e., 'heteromodal' regions). Furthermore, we decided to place a special focus on the STS region, due to a multitude of evidence pointing to its role in both social perception and audiovisual integration. Previous studies have examined activity to similar stimuli within the STS (e.g. Beauchamp et al., 2004; Kreifelts et al., 2009), but not within one experiment. It was on this basis that we conducted a single study which directly compared social and non-social stimuli, in addition to bimodal and unimodal information. We hoped that our approach could perhaps afford a clearer clarification on the functional role of this region.

Firstly, we found that a large part of the bilateral STS - extending from the pSTS to anterior STS (aSTS) - was 'people selective' in all modalities: a striking result, as previous research has generally localised face-selectivity and voice-selectivity in different portions of this region (specifically the pSTS for face perception (e.g. Haxby et al., 2000) and mid-STS for voice perception (e.g. Belin et al., 2000)). We propose that our direct analysis of 'person-selectivity' across all modalities (as compared to, for example, using separate face- and voice- localisers) plus our use of ecological, dynamic stimuli could account for the large response across this area. This result advances current disparate streams of findings by conclusively showing that the right STS has a strong preference for ecological, socially-relevant stimuli.

Secondly, using a conjunction analysis which required the audiovisual response was stronger than that in *both* of the unimodal conditions, we found audiovisual integrative regions in the bilateral thalamus and bilateral pSTS – regions which have previously been implicated in the integration of auditory and visual information (e.g. Baier et al., 2006; Kreifelts et al., 2007; Beauchamp et al., 2004; Calvert, 2001). However, a conjunction analysis with the previously localised people-selective regions removed activation in the bilateral thalamus and left pSTS, meaning only a small cluster in the right pSTS remained. This not only supports previous evidence of neural face-voice integration (see **Conclusion 1**), but the result really does underline the significant role of the right pSTS in person perception (see **Conclusion 3**). As mentioned earlier, to date different studies have provided evidence of the pSTS's role in integrating faces and voices (e.g. Calvert et al., 2000; Calvert, 2001; Kreifelts et al., 2007, 2009; Ethofer et al., 2006; Beauchamp et al., 2004): however, because integrative responses in this region have also been found for non-face/voice stimuli such as tools and letter forms/sounds (e.g. Beauchamp et al., 2004; van Atteveldt et al., 2004) it has been unclear whether this integrative response was selective.

For the first time, we show that there might exist a dedicated module for multimodal face voice processing in the right pSTS.

Finally, audiovisual convergence regions emerged in the right STS and left pSTS; however, again a conjunction with person-selective regions localised activity to the right hemisphere. As one would expect, this overlapped with the people-selective, integrative region found in the previous analysis, but people-selective convergence also occurred along a large portion of the STS, extending from the pSTS to just anterior of the mid-STS. Therefore, although this region showed a significant response to auditory *and* visual information, it did not integrate these inputs (at least, did not provoke an audiovisual response strong enough to be recognised by our integrative criterion). Overall, these results appear to show that although the right STS is responsive to multiple forms of person-specific information, neurons reactive to this information may be arranged differently throughout this structure: it some regions, the neurons seem to be arranged/connected in order to facilitate integration, and in other regions not.

6.2 Conclusions from Chapter 4

In this chapter we investigated face-voice gender integration within a behavioural experiment. Face-voice gender integration has received practically no attention until very recently: at the time of writing, I was able to locate only three relevant studies using adult participants (two of which were published in the last two years). Thus, there is much call for further study not only to corroborate work already been completed, but also to provide new avenues for further research.

Participants performed a forced choice gender decision on a series of face-voice video stimuli, parametrically morphed between gender in both the face and the voice. The parallel morphing of information from both modalities allowed us to create a set of dynamic and time-synchronised audiovisual stimuli subtly varying in face-voice incongruence, and to investigate how incompatible information in the two modalities affected behavioural responses. We also included static faces so to compare whether articulating faces facilitated gender recognition. Finally, we investigated 'top-down' effects (i.e., effects of task) by directing attention to one modality or the other.

We found that overall, participants were able to combine both sources of information (see **Conclusion 1**), with the perception of gender reflecting a contribution of information from both modalities - a behavioural effect which has been observed in the field of speech recognition (e.g. McGurk and MacDonald, 1976); identity recognition (e.g. Schweinberger et al., 2007); and the perception of emotion (e.g. de Gelder and Vroomen, 2000). However, the weighting of the two modalities on the eventual response was not equal (see **Conclusion 2**): we observed that information contained within the voice appeared to exert a stronger influence on the participant's perception of gender. This was apparent in both the condition where participant's attention was unconstrained, and where it was diverted to the face. In this latter condition, participants were unable to ignore the vocal content of the stimuli even when instructed to, and consequently their perception of facial gender was altered in the direction of gender information contained within the voice. Furthermore, there was not a significant main effect of face morph on reaction times (although there was a significant interaction between the two modalities, indicating that at least some points along the voice continuum face morph significantly modulated reaction times).

This strong effect of vocal information was an interesting result for us. As mentioned previously, vision typically dominates over audition in a number of situations (Spence, 2001), and a recent experiment on face-voice gender integration (Latinus et al., 2010) found a dominance of facial over vocal information. I suggest that stimulus differences played an important part in the difference between our results and that of Latinus et al. (2010). Indeed, the general issue of stimulus control in audiovisual studies is one that should be addressed with thought. Latinus et al. (2010) used face stimuli with little alterations (e.g. some male faces had facial hair, faces were not cropped), as compared to our study where faces were fitted with a mask to remove the hair, hairline (around which morphing had created artefacts) and jawline, males were clean shaven, and texture smoothing was applied to further remove facial hair as much as possible. This was particularly important as we were morphing between the two genders. Both of the approaches have pros and cons, and so largely it falls to the experimenter to decide which of these routes to take.

Arguably, in some respects the facial stimuli of Latinus et al. (2010) were more ecological than ours: the faces they used are ones which we would encounter in everyday life, with gender-specific hairstyles, male faces with facial hair, and female faces with make-up applied. However, culture-specific variations such as facial hair, hair length, and make-up also play a crucial part of gender discrimination, although they are unrelated to facial structure (e.g. Sugimura et al., 2006). Therefore, it is unclear in the experiment by Latinus et al. (2010) how this information might have interacted with internal facial features/structure information in the processing of gender. Our stimuli allowed for more experimental control of these factors. Indeed, as Latinus et al. (2010) write:

"Further study should investigate the perception of gender on more controlled stimuli: for example by using normalised faces and voices, or by controlling the timbre of individual voices, in order to make the tasks equally difficult across sensory modalities. We believe that this could be assessed by using faces in which all "cultural" cues of gender have been removed and by using vowels instead of words."

Internal facial features such as the eyes, brows and nose, and the ratio between these landmarks, have been shown to be crucial in the perception of gender (e.g. Brown and Perrett, 1993; Burton et al., 1993) and these all remained intact in our stimuli. However, the hairline and jawline were removed in our faces and this may have made categorising the gender of the faces more difficult. Furthermore, whilst the morphing procedures allowed for a unique exploration of face-voice gender processing, we cannot be sure that they did not affect the 'naturalness' of our stimuli: indeed, morphing was extensively applied as faces/voices were firstly morphed within-sex to obtain male and female 'prototypical' faces and voices, and then between-sex so to create a morphed continua. Averaging across the 5 male faces/voices may have smoothed the image/audio, and this may have affected the informativeness of the two modalities differently. It may have been that the face and voice averages that served as the basis for morphing did not equate in "gender strength", e.g. the perceived femaleness in the female (male) voice average was not equivalent to perceived femaleness in the female (male) face average. If face averages were more ambiguous than were voice averages, this might offer an alternative explanation for this result. Specifically, voices may then have overridden faces as the latter did not contain sufficient gender cues. Despite this though, as mentioned in **Chapter 4**, voices do arguably show more pronounced sex dimorphisms than faces. For example, the fundamental frequency (f0), which determines the perceived pitch of a voice, is typically higher in females by one octave, as compared to male voices (Linke et al., 1973).

One limitation of our study was that we did not use unimodal stimuli, and therefore we were unable to examine the speed of categorisation (an indication of categorisation difficulty) of unimodal stimuli, and directly compare these. This might have provided more information as to informativeness of our faces and voices (i.e., it might have been that faces were categorised slower, thus exerted a less strong effect than voices when combined in an audiovisual stimulus). However, it should be noted that naturally, the informativeness of two different sources is unlikely to ever be completely equal. For example, person recognition (i.e. recognising a familiar person) is typically faster and more efficient for pictures of faces than for voices (Ellis et al., 1997; Schweinberger et al., 1997). Despite this, integration effects have still been observed in the context of identity perception under simultaneous presentation (Schweinberger et al., 2007). Our study highlighted that even though face categorisation may have been harder than voice categorisation, there was still a significant interaction between the two modalities.

As described in the experimental chapter, future work should involve manipulating face and voice informativeness. Each of these modalities contains a huge number of information sources that can be independently altered and examined (e.g. for the voice, fundamental frequency and timbre; for the face, specific facial features and facial structure). Such research could potentially help us understand what information sources are crucial for not only unimodal gender perception but also those that affect the interactions with another modality, and overall modality dominance.

The conditions which directed to attention also allowed us to speculate on the automaticity of face-voice gender perception. Such an approach was taken by de Gelder and Vroomen (2000), who found that participants were unable to ignore affective information presented in another modality, even when instructed to. This finding led the authors to suggest that

face-voice emotion integration may be a mandatory process. Interestingly, their findings were slightly different to ours, in that crossmodal biases between face and voice expressions were largely bidirectional. Our results were unidirectional, as in it was only the voice that couldn't be ignored. This seems to suggest that automaticity seems to depend on the strength of information provided by each of the modalities.

Finally, our results showed that participants' categorisation of gender was not dependent on whether the video contained an articulating or static face, in any of our conditions; however, in only the condition where there was a dynamic face were reaction times affected. This finding is understandable, as dynamic and static information contained the same gender information – thus we would not expect a difference in categorisation of gender – but differences in articulation could mean that one type of stimuli offers the information faster/slower than the other. Further investigation of reaction times in our uncontrolled attention task showed that between end-point congruent and maximally incongruent stimuli there was a significant difference in reaction times, but only for the dynamic stimuli. This is somewhat consistent with the studies of Kamachi et al. (2003) and Schweinberger et al. (2007) who observed differential effects for dynamic and static faces. In particular, the latter study showed that costs and gains caused by face-voice identity incongruence and congruence were more marked for voices which were paired with articulating faces. Our work provides this first evidence that this might be able to be extended to gender perception.

So why do dynamic and static faces elicit these different effects? I would suggest that in this context, dynamic faces attract more attention, are difficult to ignore, and consequently when they are incongruent with the voice this causes slower reaction times. I would also propose that as facial articulation is time-synchronised with vocalisation, that these faces

could also draw more attention to the *voice*: when you see the lips of a face move, you generally anticipate some manner of vocalisation. I would also suggest that you would expect this vocalisation to contain congruent information as to that you are seeing. In other words, with dynamic faces you may pay more attention to the two information sources than you would to a static face paired with a voice, and are perhaps more likely to expect a viable audiovisual stimulus: when this assumption is violated, it causes a significant disturbance to the speed of categorisation.

However, it is also possible that dynamic stimuli could simply provide *more* gender information than static faces. It is conceivable that the extra facial information conveyed by moving faces provides a stronger strategic cue than a static face, and it is this which causes the patterns of costs and benefits previously suggested as evidence for face-voice integration. One way to investigate this would be to use backwards videos. Here, the motion information contained in the videos would be the same as forward playing videos and if facial motion *per se* accounts for the differences between dynamic and static stimuli, there should be no differences between forwards and backwards conditions.

After observing behaviourally that information from the face and voice could interact in order to affect gender recognition, we then targeted the other main paralinguistic processing pathway in the Belin et al. (2004) model: emotion. We extended our work on gender perception by investigating the integration of face and voice emotion using neuroimaging techniques, specifically, fMRI.

6.3 Conclusions from Chapter 5

Here we used fMRI techniques (as in **Chapter 3**) to explore the bimodal processing of affective information, but approached this from a number of different angles. We used a continuous carry-over design which promotes neuronal adaptation in a well-controlled fashion, in conjunction with dynamic face-voice stimuli morphed parametrically in two modalities - similar stimuli as in our study of gender. Our main aim was to develop previous behavioural work which provided evidence of crossmodal face-voice priming (e.g. Ellis et al., 1997; Hills et al., 2010), by testing whether such an effect would exist at a cerebral level. The continuous carry-over design also enabled us to not only examine adaptation effects, but also 'direct effects' of different stimuli on brain activity. Thus, we complemented our investigation of crossmodal adaptation by analysing the neural effect of face-voice congruence vs. incongruence and vice versa, and also searching for cerebral interactions between the two modalities. Overall, the findings from this experiment generally well supported the results from both **Chapter 3** and **Chapter 4**, and provided further grounds for **Conclusion 1**: both on-line behavioural and neuroimaging results from this experiment indicated some integration of affective information from the face and the voice.

Our behavioural results highlighted that emotion recognition was dependent on information contained in both modalities. However, the contribution of the two modalities, as in **Chapter 4**, was not equal (see **Conclusion 2**): in this experiment *face morph* exerted a far larger effect than voice, producing most of the variation in categorical responses and reaction times. This, interestingly, is in contrast to the results in the gender experiment,

where voice information was more dominant. I suggest differences in stimuli and perhaps task could, at least in part, account for this (**Conclusion 2**).

Firstly, it may simply be that faces offer more in the way of affective information than voices, whereas voice could provide more information on sex. Regarding emotion, categorisation has consistently been found to be more accurate and quicker for faces (e.g. Hess et al., 1988; Bänziger et al., 2009; de Gelder and Vroomen, 2000; Collignon et al., 2008; Kreifelts et al., 2007). Thus, the task/processing pathway under question could play an important role in which modality dominates and consequential patterns of interaction and integration of the two sources.

Regarding stimuli, as described previously, manipulation in the gender experiment could have potentially removed some facial gender cues thus making categorisation more difficult. In the experiment on emotion, the only stimulus manipulation with regards to elimination of information was removal of the hair (a cue not relevant to emotion categorisation). Additionally, although our stimuli in principle were extremely similar – dynamic, time-synchronised, and parametrically morphed – different methods were used to capture the raw recordings. These differences were particularly notable for processing of face stimuli. Thus, it is unclear how these differences (even those that were relatively subtle) could have had a part to play in determining the dominance of particular modalities.

We observed a crossmodal adaptation effect at both the brain and behavioural level. Regarding neuroimaging data, we regressed out neural unimodal adaptation effects (which were observed mainly in face- and voice-selective regions) and found that a crossmodal effect remained in the right pSTS. This strengthens further the results of **Chapter 3**, and provides yet more evidence for **Conclusion 3**: that the right pSTS appears to be

particularly important as a face-voice integrative region. It is notable that this area – and this area alone - remained after a relatively stringent analysis, where all variation due to unimodal adaptation was removed.

Furthermore, using adaptation techniques allowed us to infer the properties of neurons within this region, reaching beyond the limited spatial resolution imposed by fMRI. We tentatively proposed that our results provide evidence that at least some of the neurons within this region are truly multisensory (i.e., these singular neurons receive and combine information from these two modalities). This would be in comparison to *only* inter-digitised groups of unimodal neurons, which are found in audiovisual convergence regions ('areal convergence') but which are not singularly multisensory (note that it would still be possible for unisensory neurons to exist alongside multisensory neurons, as suggested by Barraclough et al. (2005) and Beuachamp et al. (2004)). However, it is important to note that such conclusions deduced from such a design can remain *only speculative*: only a technique such a single unit recording can provide a direct measurement of the physical properties of neurons, a possible direction for future work (discussed further below).

Interestingly, the observed crossmodal adaptation was not bi-directional: we found that voice had a stronger adaptive effect on face than face did on voice. This was highlighted by the fact that activation in the right pSTS only resulted from the voice to face physical difference, and that behaviourally, only voice to face distance had a significant effect on reaction times (although it should be noted that there was a significant interaction between the two crossmodal effects). This result highlights yet again that the effects of two modalities can be unbalanced (see **Conclusion 2**). This asymmetry is intriguing as behaviourally, face emotion morph exerted a far larger direct influence on both categorical and reaction time data, and neurally, we might have expected that if inputs were synapsing

on individual neurons (as inferred by a crossmodal adaptation effect), that the effects of these inputs would be similar: in other words, face emotion would adapt the response to voice emotion as much as voice emotion adapted the response to face emotion. The reason for this asymmetry will require much further investigation; however, one reason proposed is that potentially, even although inputs may synapse on the same neurons, they may have different modulating effects. Furthermore, if this region was also composed of unisensory neurons, an unequal mix of these (i.e., more auditory than visual neurons) could potentially 'cancel out' the effect of face emotion on the observed neural response to voice emotion. Finally, in essence the huge amount of acoustic information contained within voices may just naturally have a stronger adaptive nature as compared to faces. Either way, it appears that modality dominance is also represented at a neural level.

We also found a large effect across the right STS/STG of stimulus incongruence: the more incongruent the affective information in the face and the voice, the higher the response was. Incongruence was not significantly correlated with task-difficulty (i.e., reaction times), as compared to the other modulator, stimulus ambiguity, which was. There was also a response to ambiguity in the STS, but here information which was *less* ambiguous caused an increased response.

Untangling these results is complex, due to the interactions with task difficulty, the fact that ambiguity and congruence values were also correlated, and the conflicting findings that already exist on congruence effects within this region. Earlier research on speech perception proposed that a super-additive effect of congruent audiovisual information (as compared to unimodal), and a sub-additive effect of incongruent information was evidence for audiovisual integration (Calvert et al., 2000), a similar argument which has been made in a recent study on face-voice emotion integration (Klasen et al., 2011). However, this

finding has rarely been replicated and many studies have found converse results (see Hocking and Price, 2008 for a review). Hocking and Price (2008) propose that activation in response to incongruence reflects processing demand which is greater when two simultaneously presented stimuli refer to different concepts (incongruent condition) than when two stimuli refer to the same object (congruent condition). However, in this experiment a number of forms of congruent information (which overall produced less activation) contained two sources of relatively ambiguous information, which presumably would also require a greater processing demand. This seems to suggest that the relative ambiguity of the stimulus can interact with congruence, perhaps subduing a potentially strong response. I would also raise again the point I made earlier: participants perhaps assume congruence between face and voice information, especially as the faces are dynamic. When this is not met, this may not only cause a behavioural disruption to the speed of categorisation, but also potentially a neural effect. This may not necessarily have to reflect a greater processing demand, but rather a realisation that the two inputs are not matching. Overall, although in this study congruent information did not exert a stronger response in the STS compared to incongruent information, as might be predicted within an integration framework, we still observed that the level of incongruence modulated activity across this region. Thus, we can suggest that the STS still plays an important role in the representation of different forms of audiovisual information.

Although we observed a significant interaction between face and voice emotion morph at the behavioural level, this was not observed at the neural level (although there were significant effects of both face and voice morph). As described in **Chapter 5**, an interaction between the two modalities might only be uncovered using more stringent analyses such as MVPA, a searchlight procedure, or simply by restricting the search to

previously defined regions of interest where one might expect such an interaction to occur (i.e., multisensory convergence regions).

One limitation of this experiment, as in **Chapter 4**, was inclusion of only audiovisual stimuli. Although we were able to conduct a number of different analyses using this particular set of stimuli, the addition of unimodal stimuli to the design would have allowed for an additional assessment of multimodality. Specifically, one would be able to compare the response to bimodal as compared to unimodal emotional stimuli – both at a behavioural and neural level. Furthermore, inclusion of unimodal stimuli enables the experimenter to assess the response to faces and voices alone (i.e., reaction times, categorisation accuracy). Furthermore, to distinguish specific neural processes underlying emotion-specific and general audiovisual integration processes, a neutral category (with neutral faces and neutral prosody) could also have been included (e.g. as in Robbins et al., 2009).

I would also suggest that a complement to our forced choice, pre-test validation in this experiment would have been some manner of stimulus rating (e.g., using a Likert scale). Although categorisation using the forced choice approach was high, assessment using a rating scale would have provided an extra certification that when participants viewed the stimuli, they were powerful indicators of 'happiness' or 'anger'. Within for example, face perception it has actually been shown that more intense emotional faces lead to greater activity in FG relative to weaker emotional faces and neutral faces (Glaescher et al., 2004). However, Schultz (2005) does note that this may be due to the modulatory effects of attention, not because of a direct role for the fusiform in computations about facial expressions. Ensuring an intense presentation of emotion could have been achieved by using only stimuli categorised above say, 4 on a Likert scale. This approach could also

have been used in Chapter 3: indeed, this would have been useful in assessing whether the morphing techniques affected the 'femaleness' and 'maleness' of our morphed stimuli.

In summary, all three experiments highlighted that paralinguistic information from one modality can integrate with information contained in another modality, both at a behavioural and neural level. However, these patterns of integration can be flexible and often unsymmetrical. Understanding why this is a difficult undertaking, as one has to contend with the inherently different natures of faces and voices, their interactions with task, and experimental specific effects (e.g., stimulus manipulations). Regardless of these issues, I hope that the work in this thesis highlights that the final integrative percept is not always a 50% combination of the information contained in the face and voice. Referring to the Belin et al. (2004) model, we could imagine the interactions existing, for example, as such:

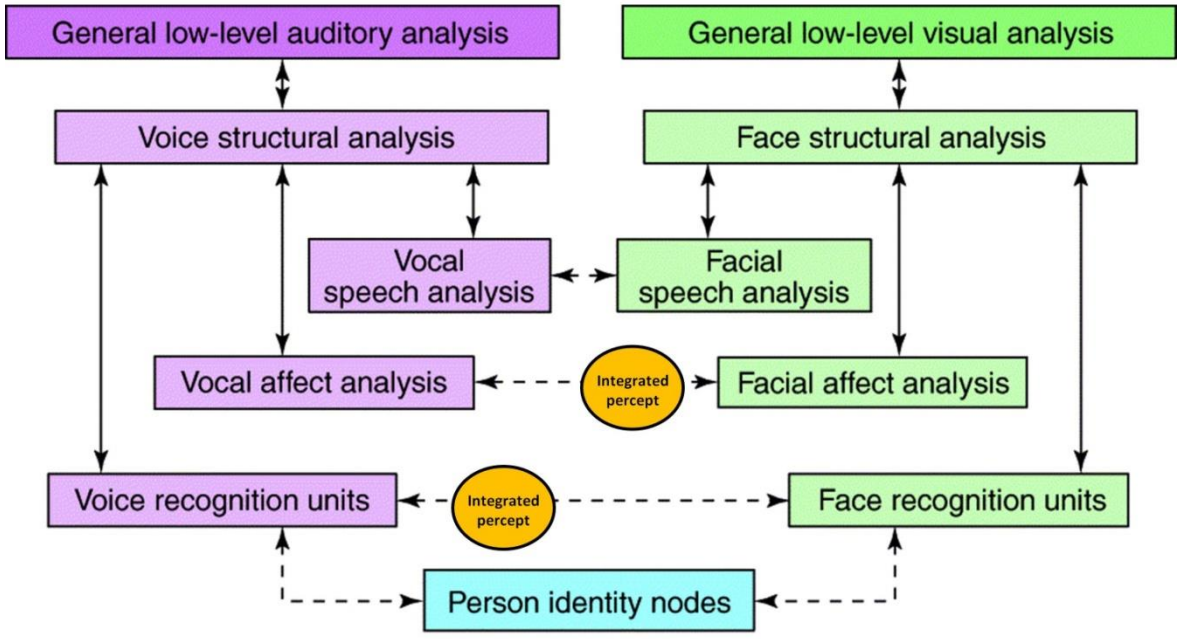


Figure 6.1. Potential modality dominance shown on Belin et al. (2004) model of voice perception.

Here, where the integrated percept lies would depend on task, experimental factors and so forth. Finally, **Chapter 3** and **5** highlighted the role of the pSTS in multisensory integration, and also specifically emotion perception. Both the experiments suggested that facial and vocal information in this region does not only converge, but is integrated or 'bound' together. The work in this thesis strongly suggests that this is achieved through singularly multisensory neurons, although of course this would require further corroboration using direct recordings of neuronal activity.

6.4 Future directions

I believe the work described throughout this thesis provides a springboard for further work. As mentioned previously, until recently little research had been conducted in the area of audiovisual person perception. There is much scope for investigation within this field, and below I provide descriptions of specific topics I consider are of importance.

6.4.1 Modality dominance and individual differences

As is apparent from the work in this thesis, one modality or sense can be weighted more than another, leading to observable differences in how we perceive not only other people around us, but the world in general. Modality dominance differs depending on task and stimulus: for example, temporal judgments made on audiovisual stimuli are based more on the auditory information, whilst visual information is used more in spatial judgments. Relevant to this thesis, we observed that generally, responses were based more on the face than on the voice when participants were asked to categorise emotion (**Chapter 5**); conversely, in our gender perception task we observed that responses were based more on the voice (**Chapter 4**), the reasons for which have been discussed above.

However, within each of these perception tasks we observed a range of strategies: some participants would use the face/voice only, integrate fully, or only integrate when information in their dominant mode was ambiguous. Examples of individual integration patterns are shown in **Chapter 4**. These results suggest that some individuals rely more on one sense to make certain social judgments, but that this dominant modality is also variable between individuals. An important question to ask is how a participant's preferred modality could affect the way in which they integrate audio and visual information. For example, behavioural responses to unimodal stimuli (e.g. speed of response, categorisation accuracy) could perhaps predict patterns of integration. This would be particularly interesting when using morphed stimuli. For example, categorisation of morphed unimodal stimuli usually produces a classic 'sigmoid curve' (see de Gelder and Vroomen, 2000; Charest et al., 2012; Bestelmeyer et al., 2010 for examples); however, there is presumably individual variation in these categorisation curves (e.g., the slope of the categorical curve; the point of subjective equality (PSE), the morph point which divides categorical perception). Could such unimodal patterns of categorisation have an influence when this modality is then combined with another? Similarly, the speed at which an individual is able to categorise two sets of unimodal stimuli is likely to affect the response to some combination of the two – one would presume that the quicker (and therefore easier) modality to categorise would exert a stronger influence when combined with one that was categorised slower.

Inter-individual differences could also be applied at the cerebral level, in attempt to establish links between neural and perceptual responses to audiovisual stimuli. Giard and Peronnet (1999) conducted one of the first studies which examined individual differences in multisensory integration. They ran an ERP study in order to investigate the processing stages and neural structures involved in multisensory object recognition. Auditory-visual

interaction components before 200 ms post-stimulus were observed in the visual areas, auditory cortex, and the right fronto-temporal area. Importantly however, when the subjects were separated into two groups according to their dominant modality to perform the task in unimodal conditions (based on a 'shortest reaction time' criteria), the integration effects were found to be similar for the two groups over the nonspecific fronto-temporal areas, but they clearly differed in the sensory-specific cortices, affecting predominantly the sensory areas of the non-dominant modality.

Nath et al. (2011) also noted that there can be inter-individual differences in audiovisual integration. They noted that the McGurk effect can be perceived by some children, but not by others. They observed that the STS of McGurk perceivers responded significantly more than that of non-perceivers to McGurk syllables, but not to other stimuli, and perceivers' hemodynamic responses in the STS were significantly prolonged. In addition to the STS, weaker differences between perceivers and non-perceivers were observed in the fusiform face area and extrastriate visual cortex. In a following study using adult subjects, Nath et al. (2012) again showed that the amplitude of the response in the left STS was significantly correlated with the likelihood of perceiving the McGurk effect: a weak STS response meant that a subject was less likely to perceive the McGurk effect, while a strong response meant that a subject was more likely to perceive it. All in all, these results suggest that the STS is an important source of inter-individual variability in audiovisual speech perception.

I would suggest functional connectivity, or psychophysical interactions analysis (PPI), could be an important tool when examining inter-individual variability. Given that the STS is a critical brain area for multisensory integration, and that it has been shown to have direct connections with the auditory and visual cortex, I think it would be of interest to examine whether the strength and number of these connections was related to an

individual's modality preference when integrating audio and visual information. For example, would someone that was visually dominant have greater connection between the visual areas and the STS? Nath and Beauchamp (2011) used a similar approach in an investigation of noisy audiovisual speech perception. They found increased functional connectivity between the STS and auditory cortex when the auditory modality was more reliable (less noisy) and increased functional connectivity between the STS and visual cortex when the visual modality was more reliable, even when the reliability changed rapidly during presentation of successive words. They therefore suggest that changes in STS functional connectivity may be an important neural mechanism underlying the perception of noisy speech. Future work could determine whether this approach could also be extended to examine modality dominance.

In order to examine modality preference at the neural level, I think it is crucial to first objectively assess participants' sensory dominance, possibly across tasks and stimulus types. This is important as a participant's sensory dominance is likely to interact with stimuli-driven sensory dominance, and therefore could bias the results of studies investigating multisensory integration. Although significant results have been observed using indicators such as reaction time (Giard and Peronnet, 1999), I propose that a more robust, validated assessment of modality preference could be developed. Such an assessment could also be related to and corroborated with auditory visual learners' questionnaires (e.g., see 'Types of Learners', *http://lyceumbooks.com/iHowToTeachEffectively.htm*). A further question regards whether this dominant sense could explain a participant's performance on different tasks and using different stimuli: for instance, whether someone who shows a visual dominance in, for example, identity recognition would be worse in temporal judgment tasks than someone

who relies more on their auditory sense? A validated measure of sensory dominance might be the first step in addressing this matter further.

6.4.2 Functionally localising multisensory regions in individual participants: a multimodal localiser?

Although inspection of individual activation effects – both behavioural and neural – is extremely important, some effects may only reach significance when performing group analyses. Moreover, random-effects group analysis is essential in order to reveal differences between groups. However, one problem has been that there is poor spatial correspondence of relevant areas using standard volumetric Talairach or MNI template brain matching techniques, further leading to suboptimal group results (Van Essen and Dierker, 2007).

Goebel and van Atteveldt (2009) suggest that surface-based techniques aligning gyri and sulci across subjects may substantially improve spatial correspondence between homolog multisensory cortical areas such as the STS/STG across subjects. They directly compared surface-based and volume-based (Talairach) registration of group data for an earlier investigation on multisensory investigation of letter-sound integration (van Atteveldt et al. 2004), and found that the analysis with cortex- based aligned data improved the statistics and provided more accurate localisation of multisensory effects in the auditory cortex and STS. Individual 'regions of interest' (ROIs) were selected based on individual anatomy, i.e., they were all located on the STS, and significantly, the authors found that comparison of the Talairach and cortex-based group statistical maps indicated that the averaged Talairach coordinates did not correspond to the location of the individual ROIs on the STS. Similarly, Kreifelts et al. (2009) used an individual mapping approach, in which an

anatomist set 27 consecutive measuring points (MP) along this structure based on a set-coordinate system by conducting individual mapping of this structure.

However, it is also important that functional-anatomical correspondence is taken into account. A complementary approach to account for individual variability is the use of functional localisers, which allows the researcher to 'functionally align' brains (Saxe et al., 2006) – take for example, the Temporal Voice Areas (TVA) localiser used in this thesis (Belin et al., 2000). Goebel and van Atteveldt (2009) suggest that future multisensory fMRI studies could use this approach to functionally localise integration areas, e.g., by using the max criterion. Additionally, Campanella and Belin (2007) propose that face-voice neuroimaging studies should always perform functional localisers of both face-selective and voice-selective areas. All-in-all, this would suggest that any audiovisual study should ideally complement the experiment with a face-localiser, voice-localiser, *and* multimodal-localiser. Although this approach is possible, it seems rather cumbersome. Additionally, given the differences in control stimuli for the separate voice and face-sensitivity experiments, one must refrain from any direct comparisons between the two qualities.

I would suggest that the work conducted in **Chapter 3** highlights that researchers in this field might be able to use a design that localised all three of these effects simultaneously: a multimodal *and* unimodal functional localiser. As well as presumably being more time-efficient and simpler to run, this would have the advantage of enabling direct comparisons between the different conditions. Such a localiser should be designed with care: for example, I would propose that as wide a set of stimuli as possible were used, perhaps even more varied than those used in **Chapter 3**. Additionally, in **Chapter 3** we noted that in our experiment, a face-selective contrast did not localise the fusiform face area (FFA). We

suggested that continuously presenting only moving faces heightened the response in the pSTS and attenuated the response in the FFA, as previously proposed by Hoffman and Haxby (2000). This would prompt the question as to the nature of the faces used within such a localiser – static, dynamic, or a mix of both. Of course, the ideal stimulus to use depends in part upon the researcher's question – however, as I believe that future research should always try to use ecological stimuli, I would still suggest that dynamic faces are used, or at least a mixture of both.

6.4.3 Multisensory and unisensory neurons: the nature of neuronal populations in integrative regions

Currently, it is still unclear how face-voice integration is represented at the neural level. I believe that the work presented in both **Chapter 3** and **Chapter 5** offers a foundation for future research using both single-cell recordings, and perhaps fMRI at a higher spatial resolution that is – at this moment – widely unavailable.

The increased response to audiovisual (or more widely, multisensory) stimulation observed at the voxel-level could have a number of different origins at the neural level. On the one hand, the audiovisual integration of emotional signals from the voice and face and voice (along with face sensitivity and voice sensitivity) could be embodied within a single population of multimodal neurons which integrate auditory and visual signals – in other words, 'true' multisensory neurons integrating stimulation from two or more sensory modalities. Indeed, at the single-neuron level, for a neuron to show an integrative response inputs from different sensory modalities need to directly synapse upon that one neuron (Meredith, 2002). Presumably, this combined sensitivity to human voices and faces would be a functional prerequisite for the audiovisual integration of human emotional signals. On the other hand, it is also conceivable that this effect might be explained by driving two

unisensory sub-populations instead of one – ones which may or may not interact during the perception of audiovisual emotion (Goebel and van Atteveldt, 2009; Tal and Amedi, 2010). If the latter scenario would be true, one might wrongly infer multisensory integration at the single neuronal level.

This begs the question: if a region contained *only* interdigitised unisensory populations – an overlap of face-sensitive and voice-sensitive neurons – what would be the mechanisms by which these neurons integrated the face-voice information and produced a significant audiovisual response, as measured by say, the max rule? It is important that this is resolved, as it in turn affects how results such as those from **Chapter 5** are interpreted. At this point, I would say that observation of crossmodal adaptation is strong evidence for multisensory neurons – neurons which have a conjoint representation of the emotion presented in both the face *and* the voice. The general consensus is that multisensory neurons should adapt to cross-modal repetitions (alternating modalities, e.g., A-V), while unisensory neurons should not or at least less (Tal and Amedi, 2010; Goebel and van Atteveldt, 2009). However, if interdigitised unisensory neurons can produce an enhanced BOLD signal in response to audiovisual stimulation, is it not also possible that the same populations could interact in order to adapt to information presented across modalities? At present, the mechanisms that would allow this are unclear: however, for example, one option might be some manner of direct /indirect connections between separate groups of unisensory neurons.

Furthermore, another question that requires resolution is this: if one observed two regions, both composed of only intermingled unisensory populations, what would be the reason for one of these regions producing significant activation as measured by an fMRI integrative criterion (super-additivity, ‘max rule’, ‘mean criterion’), and one not? In other words, what

would make one of the regions show a stronger response to audiovisual stimulation than the other? The results of **Chapter 3** also give rise to a similar question. A large portion of the right STS responded to a conjunction of A > baseline ∩ V > baseline ∩ 'people-selective' (i.e., was 'heteromodal' and people-preferring); however, only a restricted region in the right pSTS (which overlapped with the previously observed heteromodal activation) responded to a conjunction of AV > A ∩ AV > V ∩ 'people-selective'. What gives this region its integrative properties?

Evidence suggests that the human 'multisensory' cortex is most likely composed of a mixture of unisensory and multisensory subpopulations. For example, in a high-resolution fMRI study, Beauchamp et al. (2004) demonstrated that the human multisensory STS consists of visual, auditory and audiovisual 'patches'. They propose that the separate auditory and visual patches they observed were likely to represent concentrations of individual neurons that are receiving primarily auditory or visual inputs; and that the intervening multisensory patches that showed an enhanced (if not super-additive) response to audiovisual stimuli were likely to reflect concentrations of multisensory auditory-visual neurons. Thus, they suggest an integration model whereby auditory and visual inputs would arrive in the multisensory STS in separate patches, followed by integration in the intervening cortex. Further to this, Kreifelts et al. (2009) also noted that audiovisual integration of affective signals peaked in the anterior pSTS, but at an *overlap* of face- and voice-sensitive regions. We also proposed in **Chapter 5** that the observed asymmetrical crossmodal adaptation effect could be due to an unequal mix of unisensory neurons existing alongside multimodal neurons, thus cancelling out any face-to-voice crossmodal adaptation.

Increasing spatial resolution in fMRI studies may help to shed some light on the fine-grained functional organisation of small areas in the human brain (Logothetis, 2008). For example, the study of Beauchamp and colleagues (2004) used parallel imaging to achieve a spatial resolution of $1.6 \times 1.6 \times 1.6$ mm^3, providing a unique insight into the more detailed organisation of uni- and multisensory clusters in the pSTS. However, Goebel and van Atteveldt (2009) note that despite progress in high-resolution functional imaging, it is unclear what level of effective spatial resolution can be achieved with fMRI since the ultimate spatial (and temporal) resolution of fMRI is not primarily limited by technical constraints but by the spatial resolution of the vascular system, which is in the order of 1 millimetre (Duvernoy et al. 1981). Only future developments in this area will allow us to elucidate the extent to which the spatial resolution of fMRI can be pushed. Ideally, fMR-A with a powerful spatial resolution might be able to be utilised parallel to single-cell/unit recordings. Indeed, single-cell recordings still represent the only truly direct way to measure the activity of single neurons. The work of Stein and Meredith, and a number of other researchers have used this technique extremely successfully in order to heighten our understanding of multisensory processing. Hopefully this technique can be applied in the future to resolve the unanswered issues speculated on above.

6.5 General conclusion

Overall, this thesis adds to the rapidly growing body of knowledge of multisensory processes. Results highlighted integration of paralinguistic information from the face and the voice at both a perceptual and neural level, showing that combining information from two sources can significantly alter different aspects of person perception. I believe the results from the described experiments are a not only a valuable complement to work

already accomplished in this emerging field, but offer a starting point for a number of future studies. Certainly, if I am granted the opportunity to design and conduct multisensory integration experiments in the future, I will take the conclusions of this thesis into consideration.

References

Adolphs, R., Tranel, D., Damasio, A.R. (2003) Dissociable neural systems for recognizing emotions. *Brain and Cognition 52*: 61-9.

Aguirre, G.K. (2007) Continuous carry-over designs for fMRI. *Neuroimage 35*: 1480-494.

Alais, D., Burr, D. (2004) The ventriloquist effect results from nearoptimal bimodal integration. *Current Biology 14*: 257–62

Allison, T., Puce, A., McCarthy, G. (2000) Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences 4*: 267-78.

Allman, B.L, Meredith, M.A. (2007) Multisensory processing in "unimodal" neurons: Cross-modal subthreshold auditory effects in cat extrastriate visual cortex. *Journal of Neurophysiology 98*: 545–49.

Allman, B.L, Keniston, L.P, Meredith, M.A. (2008) Subthreshold auditory inputs to extrastriate visual neurons are responsive to parametric changes in stimulus quality: Sensory-specific versus non-specific coding. *Brain Research 1242*: 95–101.

Amedi, A., Malach, R., Hendler, T., Peled, S., Zohary, E. (2001) Visuo-haptic object-related activation in the ventral visual pathway. *Nature Neuroscience 4*: 324-30.

Andersen, T.S., Tiippana, K., Sams, M. (2004) Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research 21*: 301–08

Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z. (2010) Neural mechanisms for voice recognition. *Neuroimage 52*: 1528-540.

Andrews, T.J., Davies-Thompson, J., Kingstone, A., Young, A.W. (2010) Internal and external features of the face are represented holistically in face-selective regions of visual cortex. *Journal of Neuroscience 30*: 3544-552.

Arnal, L.H., Morillon, B., Kell, C.A., Giraud, A.L. (2009) Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience 29*: 13445–3453.

Attwell D, Iadecola C. (2002) The neural basis of functional brain imaging signals. *Trends in Neurosciences 25*: 621–25.

Baart, M., Vroomen, J. (2010) Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters 471*: 100–03.

Baier, B., Kleinschmidt, A., Müller, N.G. (2006) Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information. *Journal of Neuroscience 26*: 12260-2265.

Barraclough, N.E., Xiao, D., Baker, C.I., Oram, M.W., Perrett, D.I. (2005) Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience 17*: 377–91.

Bänziger, T., Grandjean, D., Scherer, K. R. (2009) Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion 9*: 691-704.

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., Martin, A. (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience 7*: 1190-192.

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.

Beauchamp, M.S. (2005) Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics 3*: 93-113.

Beer, A.L., Plank, T., Greenlee, M.W. (2011) Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex. *Experimental Brain Research 213*: 299-308

Behrmann, M., Moscovitch, M. (2001) Face recognition: evidence from intact and impaired performance. In F. Boller and J. Grafman (Eds.) Handbook of Neuropsychology (Vol. 4; pp. 181–206). North Holland, Amsterdam: Elsevier

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B. (2000) Voice-selective areas in human auditory cortex. *Nature 403*: 309-12

Belin, P., Bestelmeyer, P.E.G., Latinus, M., Watson, R. (2011) Understanding voice perception. *British Journal of Psychology 102*: 711-25

Belin, P., Zatorre, R.J., Ahad, P. (2002) Human temporal-lobe response to vocal sounds. Brain Research. *Cognitive Brain Research 13*: 17-26.

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*: 531-39

Belin, P., Zatorre, R.J. (2003) Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport 14*: 2105-109.

Benevento, L.A., Fallon, J., Davis, B.J., Rezak, M. (1977) Auditory–visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Experimental Neurology 57*: 849–72.

Bentin, S., Allison, T., Puce, A., Perez, E., McCarthy, G. (1996) Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience 8*: 551-65.

Bentin, S., Taylor, M.J., Rousselet, G.A., Itier, R.J., Caldara, R., Schyns, P.G., Jacques, C., Rossion, B. (2007) Controlling interstimulus perceptual variance does not abolish n170 face sensitivity. *Nature Neuroscience 10*: 801-02

Bertelson, P., de Gelder, B. (2004) The psychology of multimodal perception. In C. Spence and J. Driver (Eds.) *Crossmodal space and crossmodal attention* (pp. 141–177). Oxford: Oxford University Press.

Besle, J., Fort, A., Delpuech, C., Giard, M.H. (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience 20*: 2225-234.

Bestelmeyer, P.E.G., Rouger, J., DeBruine, L.M., Belin P. (2010) Auditory adaptation in vocal affect perception. *Cognition 117*: 217-23

Besl, P.J., McKay, N.D. (1992) A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14*: 239-55.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T. (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex 10*: 512-28.

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., Ward, B.D. (2004) Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience 7*: 295–301

Blank, H., Anwander, A., von Kriegstein, K. (2011) Direct structural connections between voice- and face-recognition areas. *Journal of Neuroscience 31*: 12906-2915.

Blum, P.S., Abraham, L.D., Gilman, S. (1979) Vestibular, auditory, and somatic input to the posterior thalamus of the cat. *Experimental Brain Research 34*: 1-9.

Bolles, R.C., Fischler, M.A. (1981) A RANSAC-Based Approach to Model Fitting and Its Application to Finding Cylinders in Range Data. *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, Vancouver, BC, Canada: 637-43

Bötzel K., Schulze S., Stodieck S. R. G. (1995) Scalp topography and analysis of intracranial sources of face-evoked potentials. *Experimental Brain Research 104*: 135–43.

Brancucci, A., Lucci, G., Mazzatenta, A., Tommasi, L. (2009) Asymmetries of the human social brain in the visual, auditory and chemical modalities. *Philosophical Transactions of the Royal Society of London*. *Series B, Biological Sciences 364*: 895-914.

Bredart, S., Barsics, C., Hanley, J.R. (2009) Recalling semantic information about personally known faces and voices. *European Journal of Cognitive Psychology 7*: 1013-021.

Bresciani, J.P., Dammeier, F., Ernst, M.O. (2008) Tri-modal integration of visual, tactile and auditory signals for the perception of sequences of events. *Brain Research Bulletin 75*: 753–60

Brett, M., Anton, J.-L., Valabregue, R., Poline, J.-B. Region of interest analysis using an SPM toolbox [abstract] Presented at the *8th International Conference on Functional Mapping of the Human Brain*, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage 16.

Brown, E., Perrett, D.I. (1993) What gives a face its gender? *Perception 22*: 829-40.

Bruce, C., Desimone, R., Gross, C.G. (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology 46*: 369–84.

Bruce, V., Young, A. (1986) Understanding face recognition. *British Journal of Psychology 77*: 305-27.

Bruce, V., Burton, A.M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., Linney, A. (1993) Sex discrimination: how do we tell the difference between male and female faces? *Perception 22*: 131-52.

Bruce, V., Valentine, T. (1985) Identity priming in the recognition of familiar faces. *British Journal of Psychology 76*: 363-83.

Buchanan, T. W., Lutz, K., Mirzazade, S., Specht, K., Shah, N. J., Zilles, K., J¨ancke, L. (2000) Recognition of emotional prosody and verbal components of spoken language: An fMRI study. *Cognitive Brain Research 9*: 227–38.

Burton, A.M., Bruce, V., Johnston, R.A. (1990) Understanding face recognition with an interactive activation model. *British Journal of Psychology 81*: 361–80

Burton, A.M., Bonner, L. (2004) Familiarity influences judgments of sex: The case of voice recognition. *Perception 33*: 747-52

Burton, A.M., Bruce, V., Dench, N. (1993) What's the difference between men and women? Evidence from facial measurement. *Perception 22*: 153-76

Bushara, K.O., Grafman, J., Hallett, M. (2001) Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience 21*: 300-04.

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., Woldorff, M. G. (2005) The spread of attention across modalities and space in a multisensory object. *Proceedings of the National Academy of Sciences of the United States of America 102*: 18751−8756.

Calder, A. J., Rhodes, G., Johnson, M. H., and Haxby, J. V. (Eds.) (2011) *The Oxford Handbook of Face Perception*. Oxford: University Press.

Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., Vatikiotis-Bateson, E. (2004) Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience 16*: 805-16.

Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C., Vatikiotis-Bateson, E. (2003) Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport 14*: 2213-217.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S. (1997) Activation of auditory cortex during silent lipreading. *Science 276*: 593-96.

Calvert, G.A., Campbell, R., Brammer, M.J. (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology 10*: 649–57.

Calvert, G., Campbell, R. (2003) Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience 15*: 57-70.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., David, A.S. (1999) Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport 10*: 2619-623.

Calvert, G.A., Hansen, P.C., Iversen, S.D., Brammer, M.J. (2001) Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage 14*: 427-38.

Calvert, G.A. (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex 11*: 1110–123.

Calvert, G.A., Thesen, T. (2004) Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology, Paris 98*: 191–205

Campanella, S., Belin, P. (2007) Integrating face and voice in person perception. *Trends in Cognitive Sciences 11*: 535-43

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J, Brammer, M.J., David, A.S. (2001) Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Brain Research*. *Cognitive Brain Research 12*: 233-43.

Cappe, C., Thut, G., Romei, V., Murray, M.M. (2010) Auditory-visual multisensory interactions in humans: timing, topography, directionality, and sources. *Journal of Neuroscience 30*: 12572–2580.

Carey, S. (1992) Becoming a face expert. *Philosophical Transcripts of the Royal Society London: B 335*: 95-102

Carr, L., Iacoboni, M., Dubeau, M.C., Mazziotta, J.C., Lenzi, G.L. (2003) Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceeding of the National Academy of Sciences of the United States of America 100*: 5497–502.

Carter, C.S., Botvinick, M.M., Cohen, J.D. (1999) The contribution of anterior cingulate cortex to executive processes in cognition. *Reviews in the Neurosciences 10*: 49–57

Charest, I., Pernet, C.R., Rousselet, G.A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.P., Belin, P. (2009) Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience 10*: 127.

Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P. (2012) Cerebral processing of voice gender studied using a continuous carryover fMRI design. *Cerebral Cortex*. (Epub ahead of print; published online 5th April 2012; retrieved 1st May 2012)

Chen, Y.H., Edgar, J.C., Holroyd, T., Dammers, J., Thönnessen, H., Roberts, T.P., Mathiak, K. (2010) Neuromagnetic oscillations to emotional faces and prosody. European *Journal of Neuroscience 31*: 1818–827.

Chudler, E.H., Sugiyama, K., Dong, W.K. (1995) Multisensory convergence and integration in the neostriatum and globus pallidus of the rat. *Brain Research 674*: 33-45.

Colavita, F.B. (1974) Human sensory dominance. *Perception & Psychophysics 16*: 409–12

Colavita, F.B., Tomko, R., Weisberg, D. (1976) Visual prepotency and eye orientation. *Bulletin of the Psychonomic Society 8*: 25–6

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., Lepore, F. (2008) Audio-visual integration of emotion expression. *Brain Research 25*: 126-35.

Craig, A.D. (2009) How do you feel now? The anterior insula and human awareness. *Nature Reviews Neuroscience 10*: 59–70.

Cusick, C.G. (1997) The superior temporal polysensory region in monkeys. In K. Rockland, J.H. Kaas, and A. Peters (Eds.) *Cerebral cortex* (Vol. 12; pp. 435–468). New York: Plenum Press.

Damasio, A.R. (1989) Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition 33*: 25-62.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A.M., Damasio, A.R. (1994) The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science 264*: 1102-105.

de Gelder, B., Vroomen, J. (2000) The perception of emotions by ear and by eye. *Cognition and 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*: 133–36

de Haan, M., Humphreys, K., Johnson, M. H. (2002) Developing a brain specialized for face perception: A converging methods approach. *Developmental Psychobiology 40*: 200-12.

Demb, J.B., Desmond, J.E., Wagner, A.D., Vaidya, C.J., Glover, G.H., Gabrieli, J.D. (1995) Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. *Journal of Neuroscience 15*: 5870–878

Deneve, S., Pouget, A. (2004) Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology (Paris) 98*: 249–58.

De Renzi, E., Perani, D., Carlesimo, G.A., Silveri, M.C., Fazio, F. (1994) Prosopagnosia can be associated with damage confined to the right hemisphere--an MRI and PET study and a review of the literature. *Neuropsychologia 32*: 893-902.

Desimone, R. (1996) Neural mechanisms for visual memory and their role in attention. *Proceeding of the National Academy of Sciences of the United States of America 93*: 13494-3499

Diamond, R., Carey, S. (1986) Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General 115*: 107-17.

Dolan, R.J., Morris, J.S., de Gelder, B. (2001) Crossmodal binding of fear in voice and face. *Proceeding of the National Academy of Sciences of the United States of America 98*: 10006–0010

Dolan, R.J., Vuilleumier, P. (2003) Amygdala automaticity in emotional processing. *Annals of the New York Academy of Sciences 985*: 348 – 55.

Driver, J., Spence, C. (2000) Multisensory perception: Beyond modularity and convergence. *Current Biology 10*: R731−R735.

Driver, J., Spence, C. (1998) Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society B-Biological Sciences 353*: 1319−331.

Driver J. (2001) A selective review of selective attention research from the past century. *British Journal of Psychology 92*: 53-78.

Driver, J., Noesselt, T. (2008) Multisensory interplay reveals crossmodal influences on 'sensory specific' brain regions, neural responses, and judgments. *Neuron 57*: 11–23.

Drucker, D.M., Kerr, W.T., Aguirre, G.K. (2009) Distinguishing conjoint and independent neural tuning for stimulus features with fMRI adaptation. *Journal of Neurophysiology 101*: 3310-324.

Duvernoy, H.M., Delon, S., Vannson, J.L. (1981) Cortical blood vessels of the human brain. *Brain Research Bulletin 7*: 519–79

Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., Gur, R.C., Mathiak, K. (2008) Recognition profile of emotions in natural and virtual faces. *PLoS One 3*: e3628.

Eifuku, S., De Souza, W. C., Tamura, R., Nishijo, H., Ono, T. (2004) Neuronal correlates of face identification in the monkey anterior temporal cortical areas. *Journal of Neurophysiology 91*: 358–71.

Eimer, M. (2000) Effects of face inversion on the structural encoding and recognition of faces. Evidence from event-related brain potentials. *Cognitive Brain Research 10*: 145–58

Eimer, M., Holmes, A. (2007) Event-related brain potential correlates of emotional face processing. *Neuropsychologia 45*: 15–31

Eippert, F., Veit, R., Weiskopf, N., Erb, M., Birbaumer, N., Anders, S. (2007) Regulation of emotional responses elicited by threat-related stimuli. *Human Brain Mapping 28*: 409–23.

Ellis, H.D., Jones, D.M., Mosdell, N. (1997) Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology 88*: 143-56

Ellis, A.W., Burton, A.M., Young, A.W., Flude, B.M. (1997) Repetition priming between parts and wholes: Tests of a computational model of familiar face recognition. *British Journal of Psychology 88*: 579-608.

Ellis, A.W., Young, A.W., & Flude, B.M. (1990). Repetition priming and face processing: Priming occurs within the system that responds to the identity of a face. *Quarterly Journal of Experimental Psychology 42A*: 495-512.

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

Ernst, M.O., Bulthoff, H.H. (2004) Merging the senses into a robust percept. *Trends in Cognitive Sciences 8*: 162–69

Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P. (2009) Decoding of emotional information in voice-sensitive cortices. *Current Biology 19*: 1028-033

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

Ethofer, T. Anders, S., Erb, M., Herbert, C., Wiethoff, S., Kissler, J., Grodd,W., Wildgruber, D. (2006) Cerebral pathways in processing of affective prosody: a dynamic causal modeling study. *Neuroimage 30*: 580-87.

Ethofer, T., Pourtois, G., Wildgruber, D. (2006) Investigating audiovisual integration of emotional signals in the human brain. *Progress in Brain Research 156*: 345–61.

Fairhall, S. L., Macaluso, E. (2009) Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *European Journal of Neuroscience 29*: 1247−257.

Falchier, A., Clavagnier, S., barone, P., Kennedy, H. (2002) Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience 22*: 5749–759.

Föcker, J., Hölig, C., Best, A., Röder, B. (2011) Crossmodal interaction of facial and vocal person identity information: an event-related potential study. *Brain Research 1385*: 229-45.

Fort, A., Delpuech, C., Pernier, J., Giard, M.-H. (2002) Early auditory-visual interactions in human cortex during nonredundant target identification. *Cognitive Brain Research 14*: 20–30.

Fu, K. G., Johnston, T. A., Shah, A. S., Arnold, L., Smiley, J., Hackett, T. A., Garraghty, P. E., Schroeder, C. E. (2003) Auditory cortical neurons respond to somatosensory input. *Journal of Neuroscience 23*: 7510–515.

Fuhrmann Alpert, G., Hein, G., Tsai, N., Naumer, M.J., Knight, R.T. (2008) Temporal characteristics of audiovisual information processing. *Journal of Neuroscience 28*: 5344–349.

Fujii, N., Mushiake, H., Tanji, J., 2002. Distribution of eye- and armmovement-related neuronal activity in the SEF and in the SMA and pre-SMA of monkeys. *Journal of Neurophysiology 87*: 2158– 166.

Fusar-Poli, P., Placentino, A., Carletti, F., Allen, P., Landi, P., Abbamonte, M., Barale, F., Perez, J., McGuire, P., Politi, P.L. (2009) Laterality effect on emotional faces processing: ALE meta-analysis of evidence. *Neuroscience Letters 452*: 262–67.

Gauthier, I., Anderson, A.W., Tarr, M.J., Skudlarski, P., Gore, J.C. (1997) Levels of categorization in visual recognition studied with functional MRI. *Current Biology 7*: 645–51

Gauthier, I., Behrmann, M., Tarr, M.J. (1999) Can face recognition really be dissociated from recognition? *Journal of Cognitive Neuroscience 11*: 349–70.

Gauthier, I., Skudlarski, P., Gore, J.C., Anderson, A.W. (2000) Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience 3*: 191–97.

Gazzaniga, M.S., Smylie, C.S. (1983) Facial recognition and brain asymmetries: clues to underlying mechanisms. *Annals of Neurology 13*: 536-40.

George, M.S., Parekh, P.I., Rosinsky, N., Ketter, T.A., Kimbrell, T.A., Heilman, K.M., Herscovitch, P., Post, R.M. (1996) Understanding emotional prosody activates right hemisphere regions. *Archives of Neurology 53*: 665-70.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthélémy, C., Brunelle, F., Samson, Y., Zilbovicius, M. (2004) Abnormal cortical voice processing in autism. *Nature Neuroscience 7*: 801-02.

Ghazanfar, A.A., Schroeder, C.E. (2006) Is neocortex essentially multisensory? *Trends in Cognitive Sciences 10*: 278–85

Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., Logothetis, N.K. (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience 25*: 5004–012.

Ghazanfar, A.A., Chandrasekaran, C., Logothetis, N.K. (2008) Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *Journal of Neuroscience 28*: 4457–469

Ghazanfar, A.A., Logothetis, N.K. (2003) Facial expressions linked to monkey calls. *Nature 423*: 937–38.

Giard, M.-H., Peronnet, F. (1999) Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Mental Processes and Brain Activation 11*: 473–90.

Glaescher, J., Tuescher, O., Weiller, C., Buechel, C. (2004) Elevated responses to constant facial emotions in different faces in the human amygdala: an fMRI study of facial identity and expression. *BMC Neuroscience 17*: 5

Gobbini, M.I., Haxby, J.V. (2007) Neural systems for recognition of familiar faces. *Neuropsychologia 45*: 32–41.

Goebel, R., van Atteveldt, N. (2009) Multisensory functional magnetic resonance imaging: a future perspective. *Experimental Brain Research 198*: 153-64.

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., Vuilleumier, P. (2005) The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nature Neuroscience 8*: 145–46

Grant, K. W., Walden, B. E., Seitz, P. F. (1998) Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America 103*: 2677–690.

Green, K.P., Kuhl, P.K., Meltzoff, A.N., Stevens, E.B. (1991) Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics 50*: 524-26

Grill-Spector, K., Knouf, N., Kanwisher, N. (2004) The fusiform face area subserves face perception, not generic within category identification. *Nature Neuroscience 7*: 555–62.

Grill-Spector, K., Henson, R., Martin, A. (2006) Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences 10*: 14-23

Grill-Spector, K. Malach, R. (2001) fmr-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychologica 107*: 293-321

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.

Grinband, J., Hirsch, J., Ferrera, V. P. (2006) A neural representation of categorization uncertainty in the human brain. *Neuron 49*: 757–63

Hagan, C.C., Woods, W., Johnson, S., Calder, A.J., Green, G.G., Young, A.W. (2009) MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America 106*: 20010-0015.

Halgren, E., Raij, T., Marinkovic, K., Jousmäki, V., Hari, R. (2000) Cognitive response profile of the human fusiform face area as determined by MEG. *Cerebral Cortex 10*: 69 – 81.

Hammerschmidt, K., Jürgens, U. (2007) Acoustic correlates of affective prosody. *Journal of Voice 21*: 531–40

Harris, A., Nakayama, K. (2008) Rapid adaptation of the M170 response: Importance of face parts. *Cerebral Cortex 18*: 467–76.

Hasselmo, M.E., Rolls, E.T., Baylis, G.C. (1989) The role of expression and identity in the face selective response of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research 32*: 203–18.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I. (2000) The distributed human neural system for face perception. *Trends in Cognitive Sciences 4*: 223–32.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I. (2002) Human neural systems for face recognition and social communication. *Biological Psychiatry 51*: 59–67.

Heekeren, H.R., Marrett, S., Ungerleider, L.G. (2008) The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience 9*: 467-79

Hein, G., Knight, R.T. (2008) Superior temporal sulcus--It's my area: or is it? *Journal of Cognitive Neuroscience 20*: 2125-136.

Heekeren, H. R., Marrett, S., Ruff, D. A., Bandettini, P. A., Ungerleider, L. G. (2006) Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. *Proceeding of the National Academy of Sciences of the United States of America 103*: 10023-0028

Henson, R.N.A., Rugg, M.D. (2003) Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia 41*: 263-70

Hess, U., Kappas, A., Scherer, K. (1988) Multichannel communication of emotion: Synthetic signal production. In K. Scherer (Ed.) *Facets of Emotion: Recent Research* (pp 161-182). Hillsdale, NJ: Erlbaum

Hills, P.J., Elward, R.L., Lewis, M.B. (2010) Cross-modal identity aftereffects and their relation to priming. *Journal of Experimental Psychology: Human Perception and Performance 36*: 876-91

Hocking, J., Price, C.J. (2008) The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex 18*: 2439–449.

Hoffman, E.A., Haxby, J.V. (2000) Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience 3*: 80-4.

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

James, T.W., Stevenson, R.A. (2012) The Use of fMRI to Assess Multisensory Integration. In M.M. Murray and M.T. Wallace (Eds.) *The Neural Bases of Multisensory Processes* (Chapter 8). Boca Raton (FL): CRC Press.

Jessen, S., Kotz, S.A. (2011) The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *Neuroimage 58*: 665 – 74.

Jiang, Y., Haxby, J.V., Martin, A., Ungerleider, L.G., Parasuraman, R. (2000) Complementary neural mechanisms for tracking items in human working memory. *Science 287*: 643–46

Joassin, F., Campanella, S., Debatisse, D., Guerit, J.M., Bruyer, R., Crommelinck, M. (2004) The electrophysiological correlates sustaining the retrieval of face-name associations: an ERP study. *Psychophysiology 41*: 625-35.

Joassin, F., Pesenti, M., Maurage, P., Verreckt, E., Bruyer, R., Campanella, S. (2011) Cross-modal interactions between human faces and voices involved in person recognition. *Cortex 47*: 367-76.

Joassin, F., Maurage, P., Campanella, S. (2011) The neural network sustaining the crossmodal processing of human gender from faces and voices: an fMRI study. *Neuroimage 54*: 1654-661

Johnston, R.A., Barry, C., Williams, C. (1996) Incomplete faces don't show the whole picture: Repetition priming from jumbled faces. *Quarterly Journal of Experimental Psychology 49A*: 596-615.

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*: 242–49.

Jones, E.G., Powell, T.P. (1970) An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain 93*: 793-820.

Jones, J.A., Callan, D.E. (2003) Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport 14*: 1129 –133.

Jonides, J. (1981) Voluntary vs. automatic control over the mind's eye's movements. In J.B. Long and A.D. Baddeley (Eds.) *Attention and performance IX* (pp. 187−203). Hillsdale, NJ: Erlbaum.

Kamachi, M., Hill, H., Lander, K., Vatikiotis-Bateson, E. (2003) "Putting the face to the voice": matching identity across modality. *Current Biology 13*: 1709-714.

Kanwisher, N. (2000) Domain specificity in face perception. *Nature Neuroscience 3*: 759-763.

Kanwisher, N., McDermott, J., Chun, M.M. (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience 17*: 4302-311.

Kanwisher, N., Stanley, D., Harris, A. (1999) The fusiform face area is selective for faces not animals. *NeuroReport 10*: 183–87.

Kawahara H. In: VoQual 03: *Voice Quality: Functions, Analysis and Synthesis*, 2003 August 27--29; Geneva (Switzerland): ISCA Tutorial and Research Workshop. 2003. Exemplar-based voice quality analysis and control using a high quality auditory morphing procedure based on straight.

Kawahara, H. (2006) Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology 27*: 349-353

Kayser, C., Logothetis, N.K. (2007) Do early sensory cortices integrate cross-modal information? *Brain Structure and Function 212*: 121-32.

Kayser, C., Logothetis, N.K., Panzeri, S. (2010) Visual enhancement of the information representation in auditory cortex. *Current Biology 20*: 19–24

Kerns, J.G., Cohen, J.D., MacDonald, A.W., Cho, R.Y., Stenger, V.A., Carter, C.S. (2004) Anterior cingulate conflict monitoring and adjustments in control. *Science 303*: 1023-026.

Kilts, C.D., Egan, G., Gideon, D.A., Ely, T.D., Hoffman, J.M. (2003) Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions. *Neuroimage 18*: 156–68.

King, A.J., Palmer, A.R. (1985) Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research 60*: 492–500

Klinge, C., Eippert, F., Röder, B., Büchel, C. (2010) Corticocortical connections mediate primary visual cortex responses to auditory stimulation in the blind. *Journal of Neuroscience 30*: 12798–2805.

Klasen, M., Kenworthy, C.A., Mathiak, K.A., Kircher, T.T., Mathiak, K. (2011) Supramodal representation of emotions. *Journal of Neuroscience 31*: 13635-3643.

Koelewijn, T., Bronkhorst, A., Theeuwes, J. (2010) Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica 134*: 372–384

Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V., Rizzolatti, G. (2002) Hearing sounds, understanding actions: action representation in mirror neurons. *Science 297*: 846-48.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Ono, T. (2005) Auditory thalamus integrates visual inputs into behavioral gains. *Nature Neuroscience 8*: 1203-209.

Koppen. C., Spence, C. (2007) Seeing the light: exploring the Colavita visual dominance effect. *Experimental Brain Research 180*: 737-54.

Koppen, C., Alsius, A., Spence, C. (2008) Semantic congruency and the Colavita visual dominance effect. *Experimental Brain Research 184*: 533–46

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*: 1445–456.

Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., Wildgruber, D. (2009) Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus. *Neuropsychologia 47*: 3059-066.

Kreifelts, B., Ethofer, T., Huberle, E., Grodd, W., Wildgruber, D. (2010) Association of trait emotional intelligence and individual fMRI activation patterns during the perception of social signals from voice and face. *Human Brain Mapping 31*: 979–91.

Lachs, L., Pisoni, D. B. (2004) Crossmodal source identification in speech perception. *Ecological Psychology 16*: 159-87.

Lachs, L., Pisoni, D. B. (2004) Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America 116*: 507-18.

Latinus, M., VanRullen, R., Taylor, M. (2010) Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neuroscience 11*: 36

Latinus, M., Crabbe, F., Belin, P. (2011) Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex 21*: 2820-828.

Laurienti, P.J, Perrault, T.J, Stanford, T.R, Wallace, M.T, Stein, B.E. (2005) On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research 166*: 289–97.

Lee, L.C., Andrews, T.J., Johnson, S.J., Woods, W., Gouws, A., Green, G.G., Young, A.W. (2010) Neural responses to rigidly moving faces displaying shifts in social attention investigated with fMRI and MEG. *Neuropsychologia 48*: 477-90.

Le Grand, R., Mondloch, C. J., Maurer, D., Brent, H. P. (2001) Early visual experience and face processing. *Nature 410*: 890.

Le Grand, R., Mondloch, C.J., Maurer, D., Brent, H.P. (2003) Expert face processing requires visual input to the right hemisphere during infancy. *Nature Neuroscience 6*: 1108-112.

Lewis, J. W., Van Essen, D. C. (2000) Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *Journal of Comparative Neurology 428*: 112−37

Li, L., Miller, K., Desimone. R. (1993) The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of Neurophysiology 69*: 1918-929

Linden, J. F., Grunewald, A., Andersen, R. A. (1999) Responses to auditory stimuli in macaque lateral intraparietal area II. Behavioral modulation. *Journal of Neurophysiology 82*: 343−58.

Linden, D.E., Thornton, K., Kuswanto, C.N., Johnston, S.J., van de Ven, V., Jackson, M.C. (2011) The brain's voices: comparing nonclinical auditory hallucinations and imagery. *Cerebral Cortex 21*: 330-37.

Linke, C.E. (1973) A study of pitch characteristics of female voices and their relationship to vocal effectiveness. *Folia Pho- niatrica 25*: 173-185

Liu, J., Harris, A., Kanwisher, N. (2002) Stages of processing in face perception: An MEG study. *Nature Neuroscience 5*: 910–16.

Logothetis, N.K., Wandell, B.A. (2004) Interpreting the BOLD signal. *Annual Review of Physiology 66*: 735–69.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A. (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature 412*: 150–57.

Love, S.A., Pollick, F.E., Latinus, M. (2011) Cerebral correlates and statistical criteria of cross-modal face and voice integration. *Seeing and Perceiving 24*: 351-67.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., Parra, L. C. (2009) Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PloS one 4*: e4638.

Macaluso, E., Driver, J. (2005) Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends in Neurosciences 28*: 264-71.

Macaluso, E. (2000) Modulation of human visual cortex by crossmodal spatial attention. *Science 289*: 1206–208

Macaluso, E., Driver, J. (2001) Spatial attention and crossmodal interactions between vision and touch. *Neuropsychologia 39*: 1304−316.

Macaluso, E., George, N., Dolan, R., Spence, C., Driver J. (2004) Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage 21*: 725-32.

Mack, A., Rock, I. (1998) *Inattentional Blindness*. Cambridge, MA: MIT Press.

MacSweeney, M., Amaro, E., Calvert, G.A., Campbell, R., David, A.S., McGuire, P., Williams, S.C., Woll, B., Brammer, M.J. (2000) Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport 11*: 1729-733.

Massaro, D.W., Egan, P.B. (1996) Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review 3*: 215–21.

McDonald, A.J. (1998) Cortical pathways to the mammalian amygdala. *Progress in Neurobiology 55*: 257–332.

McDonald, J. J., Teder-Salejarvi, W. A., Di Russo, F., Hillyard, S. A. (2003) Neural substrates of perceptual enhancement by cross-modal spatial attention. *Journal of Cognitive Neuroscience 15*: 10−19.

McCarthy, G., Puce, A., Gore, J.C., Allison, T. Face-Specific Processing in the Human Fusiform Gyrus. *Journal of Cognitive Neuroscience 9*: 605-10

McGurk, H., MacDonald, J. (1976) Hearing lips and seeing voices. *Nature 64*: 746-48.

McKone, E., Kanwisher, N. (2005) Does the human brain process objects of expertise like faces? A review of the evidence. In S. Dehaene, J.R. Duhamel, M. Hauser and G. Rizzolatti (Eds.) *From monkey brain to human brain*. Cambridge: MIT Press.

Meng, M., Cherian, T., Singal, G., Sinha, P. (2012) Lateralization of face processing in the human brain. *Proceedings*. *Biological Sciences/The Royal Society 279*: 2052-061.

Meredith, M.A., Clemo, H.R. (2010) Corticocortical connectivity subserving different forms of multisensory convergence. In M.J. Naumer and J. Kaiser (Eds.) *Multisensory Object Perception in the Primate Brain* (pp 7-20). Netherlands: Springer Science

Meredith, M. A., Stein, B. E. (1983) Interactions among converging sensory inputs in the superior colliculus. *Science 221*: 389–91.

Meredith, M.A, Allman, B.L. (2009) Subthreshold multisensory processing in cat auditory cortex. *NeuroReport 20*: 126–31.

Meredith, M. A., Nemitz, J. W., Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurones: I. Temporal factors. *Journal of Neuroscience 10*: 3215-229.

Meredith, M.A. (2002) On the neuronal basis for multisensory convergence: a brief overview. *Cognitive Brain Research 14*: 31–40.

Mesulam, M.M., Mufson, E.J. (1982) Insula of the old world monkey. III: Efferent cortical output and comments on function. *Journal of Comparative Neurology 212*: 38-52.

Miller, E.K., Desimone, R. (1994) Parallel neuronal mechanisms for short-term memory. *Science 263*: 520-22

Miller, J. (1982) Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology 14*: 247–79.

Miller, J. (1986) Timecourse of coactivation in bimodal divided attention. *Perception and Psychophysics 40*: 331–43.

Miller, L.M., D'Esposito, M. (2005) Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience 25*: 5884 –893.

Miller, M.B., Kingstone, A., Gazzaniga, M.S. (2002) Hemispheric encoding asymmetry is more apparent than real. *Journal of Cognitive Neuroscience 14*: 702-08.

Mitchell, R. L., Elliott, R., Barry, M., Cruttenden, A., Woodruff, P. W. (2003) The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia 41*: 1410–421.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., Foxe, J. J. (2002) Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research 14*: 115-28.

Montgomery, K.L., Haxby, J.V. (2008) Mirror neuron system differentially activated by facial expressions and social hand gestures: a functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience 20*: 1866–877.

Morris, J.S., Friston, K.J., Büchel, C., Frith, C.D., Young, A.W., Calder, A.J., Dolan, R.J. (1998) A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain 121*: 47–57.

Morris, J. S., Scott, S. K., & Dolan, R. J. (1999) Saying it with feeling: Neural responses to emotional vocalizations. *Neuropsychologia 37*: 1155–163.

Morton, J., Johnson, M. H. (1991) CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological Review 98*: 164-81.

Möttönen, R., Krause, C.M., Tiippana, K., Sams, M. (2002) Processing of changes in visual speech in the human auditory cortex. *Brain Researh*. *Cognitive Brain Research 13*: 417-25.

Möttönen, R., Schurmann, M., Sams, M. (2004) Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neuroscience Letters 363*: 112-15.

Mufson, E.J., Mesulam, M.M. (1982) Insula of the old world monkey. II: Afferent cortical input and comments on the claustrum. *Journal of Comparative Neurology 212*: 23-37.

Mufson, E.J., Mesulam, M.M. (1984) Thalamic connections of the insula in the rhesus monkey and comments on the paralimbic connectivity of the medial pulvinar nucleus. *Journal of Comparative Neurology 227*: 109-20.

Müller, V.I,, Habel, U., Derntl, B., Schneider, F., Zilles, K., Turetsky, B.I., Eickhoff, S.B. (2011) Incongruence effects in crossmodal emotional integration. *Neuroimage 54*: 2257-266.

Munhall, K.G., Gribble, P., Sacco, L., Ward, M. (1996) Temporal constraints on the McGurk effect. *Perception & Psychophysics 58*: 351-62.

Munhall, K.G., Buchan, J.N. (2004) Something in the way she moves. *Trends in Cognitive Sciences 8*: 51-3.

Naccache, L., Dehaene, S. (2001) The priming method: imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral Cortex 11*: 966–74

Nath, A.R., Fava, E.E., Beauchamp, M.S. (2011) Neural correlates of interindividual differences in children's audiovisual speech perception. *Journal of Neuroscience 31*: 13963-3971.

Nath, A.R., Beauchamp, M.S. (2012) A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage 59*: 781-87.

Nath, A.R., Beauchamp, M.S. (2011) Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience 31*: 1704-714.

Nonyane, B.A.S., Theobald, C.M. (2007) Design sequences for sensory studies: achieving balance for carry-over and position effects. *British Journal of Mathematical and Statistical Psychology 60*: 339-49.

Ochiai, T., Grimault, S., Scavarda, D., Roch, G., Hori, T., Rivière, D., Mangin, J.F., Régis, J. (2004) Sulcal pattern and morphology of the superior temporal sulcus. *Neuroimage 22*: 706-19.

Ochsner, K.N., Bunge, S.A., Gross, J.J., Gabrieli, J.D. (2002) Rethinking feelings: an FMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience 14*: 1215–229.

Olson, I.R., Gatenby, J.C., Gore, J.C. (2002) A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Brain Research*. *Cognitive Brain Research 14*: 129-38.

Padberg, J., Seltzer, B., Cusick, C.G. (2003) Architectonics and cortical connections of the upper bank of the superior temporal sulcus in the rhesus monkey: an analysis in the tangential plane. *Journal of Comparative Neurology 467*: 418-34.

Patterson, M., Werker, J.F. (2002) Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology 81*: 93–115.

Paulmann, S., Kotz, S.A. (2008) Early emotional prosody perception based on different speaker voices. *Neuroreport 19*: 209–13

Paus, T., Koski, L., Caramanos, Z., Westbury, C. (1998) Regional differences in the effects of task difficulty and motor output on blood flow response in the human anterior cingulate cortex: a review of 107 PET activation studies. *NeuroReport 9*: R37–R47

Paus, T. (2001) Primate anterior cingulate cortex: where motor control, drive and cognition interface. *Nature Reviews Neuroscience 2*: 417–24

Pearson, R.C., Brodal, P., Gatter, K.C., Powell, T.P. (1982) The organization of the connections between the cortex and the claustrum in the monkey. *Brain Research 234*: 435-41.

Pernet, C., Schyns, P.G., Demonet, J.F. (2007) Specific, selective or preferential: comments on category specificity in neuroimaging. *Neuroimage 35*: 991-97

Perrault, T.J. Jr., Vaughan, J.W, Stein, B.E, Wallace, M.T. (2003) Neuron-specific response characteristics predict the magnitude of multisensory integration. *Journal of Neurophysiology 90*: 4022–026

Philiastides, M.G., Sajda, P. (2007) EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making. *Journal of Neuroscience 27*: 13082–3091

Phillips, M.L., Young, A.W., Senior, C., Calder, A.J., Rowland, D., Brammer, M., Bullmore, E.T., Andrew, C., Willimas, S.C.R., Gray, J., David, A.S. (1997) A specific neural substrate for perception of facial expressions of disgust. *Nature 389*: 495-498.

Pick, H.L., Warren, D.H., Hay, J.C. (1969) Sensory conflict in judgments of spatial direction. *Perception & Psychophysics 6*: 203–05

Picard, N., Strick, P.L. (1996) Motor areas of the medial wall: a review of their location and functional activation. *Cerebral Cortex 6*: 342– 53.

Pihan, H., Altenmuller, E., Ackermann, H. (1997) The cortical processing of perceived emotion: a DC-potential study on affective speech prosody. *Neuroreport 8*: 623–27

Pitcher, D., Dilks, D.D., Saxe, R.R., Triantafyllou, C., Kanwisher, N. (2011) Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage 56*: 2356-363.

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

Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., Crommelinck, M. (2000) The time-course of intermodal binding between seeing and hearing affective information. *Neuroreport 11*: 1329–333.

Pourtois, G., Debatisse, D., Despland, P.A., de Gelder, B. (2002) Facial expressions modulate the time course of long latency auditory brain potentials. *Brain Research*. *Cognitive Brain Research 14*: 99–105.

Puce, A., Allison, T., Asgari, M., Gore, J.C., McCarthy, G. (1996) Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of Neuroscience 16*: 5205–215.

Puce, A., Allison, T., Gore, J.C., McCarthy, G. (1995) Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology 74*: 1192–199.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences 24*: 574–90.

Rama, P., Martinkauppi, S., Linnankoski, I., Koivisto, J., Aronen, H. J., Carlson, S. (2001) Working memory of identification of emotional vocal expressions: An fMRI study. *Neuroimage 13*: 1090–101.

Remedios, R., Logothetis, N.K., Kayser, C. (2009) Monkey drumming reveals common networks for perceiving vocal and nonvocal communication sounds. *Proceeding of the National Academy of Sciences of the United States of America 106*: 18010–8015.

Rhodes, G., Byatt, G., Michie, P.T., Puce, A. (2004) Is the fusiform face area specialized for faces, individuation, or expert individuation? *Journal of Cognitive Neuroscience 16*: 189–203

Robins, D.L., Hunyadi, E., Schultz, R.T. (2009) Superior temporal activation in response to dynamic audio-visual emotional cues. *Brain and Cognition 69*: 269–78

Rockland, K.S., Ojima, H. (2003) Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology 50*: 19-26.

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006) Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics 68*: 84-93.

Ross, L.A.., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., Foxe, J. J. (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex 17*: 1147–153.

Rossion, B., Caldara, R., Seghier, M., Schuller, A.M., Lazeyras, F., Mayer, E. (2003) A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain 126*: 2381-395.

Royet, J.P., Zald, D., Versace, R., Costes, N., Lavenne, F., Koenig, O., Gervais, R. (2000) Emotional responses to pleasant and unpleasant olfactory, visual and auditory stimuli: a positron emission tomography study. *Journal of Neuroscience 20*: 7752-759.

Sabatinelli, D., Fortune, E.E., Li, Q., Siddiqui, A., Krafft, C., Oliver, W.T., Beck, S., Jeffries, J. (2011) Emotional perception: Meta-analyses of face and natural scene processing. *Neuroimage 54*: 2524-533

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S-T, Simola, J. (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters 127*: 141-45.

Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., Vuilleumier, P. (2005) Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody. *Neuroimage 28*: 848-58.

Santangelo, V., Spence, C. (2007) Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology-Human Perception and Performance 33*: 1311−321.

Saxe, R., Brett, M., Kanwisher, N. (2006) Divide and conquer: a defense of functional localizers. *Neuroimage 30*: 1088–096

Scherer, K.R. (2003) Vocal communication of emotion: A review of research paradigms. *Speech Communication 40*: 227–56

Schroeder, C. E., Lindsley, R.W., Specht, C., Marcovici, A., Smiley, J. F., Javitt, D. C. (2001) Somatosensory input to auditory association cortex in the macaque monkey. *Journal of Neurophysiology 3*: 1322–327.

Schroeder, C.E., Foxe, J.J. (2002) The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Research*. *Cognitive Brain Research 14*: 187-98.

Schroeder, C.E., Foxe, J.J. (2004) Multisensory Convergence in Early Cortical Processing. In G. Calvert, C. Spence and B.E. Stein (Eds.) *The Handbook of Multisensory Processes* (pp 295-310). Cambidge, Massachusetts: The MIT Press

Schroeder, C.E., Smiley, J., Fu, K.G., McGinnis, T., O'Connell, M.N., Hackett, T.A. (2003) Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology 50*: 5-17.

Schultz, R.T. (2005) Developmental deficits in social perception in autism: The role of the amygdala and fusiform face area. *International Journal of Developmental Neuroscience 23*:125–141.

Schweinberger, S.R., Herholz, A., Sommer, W. (1997) Recognizing famous voices: influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language and Hearing Research 40*: 453-63.

Schweinberger, S.R., Robertson, D., Kaufmann, J.M. (2007) Hearing facial identities. *The Quarterly Journal of Experimental Psychology 60*: 1446 – 456.

Scott, S.K., Young, A.W., Calder, A.J., Hellawell, D.J., Aggleton, J.P., Johnson, M. (1997) Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature 385*: 254- 57

Sekiyama, K., Kanno, I., Miura, S., Sugita, Y. (2003) Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research 47*: 277–87.

Seltzer, B., Pandya, D.N. (1978) Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Research 149*: 1 – 24.

Shah, N.J., Marshall, J.C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H.J., Fink, G.R. (2001) The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. *Brain 124*: 804-15.

Shams, L., Kamitani, Y., Shimojo, S. (2000) Illusions. What you see is what you hear. *Nature 408*: 788

Sheffert, S.M., Olsen, E. (2004) Audiovisual speech facilitates voice learning. *Perception & Psychophysics 66*: 352-62

Skipper, J.I., Nusbaum, H.C., Small, S.L. (2005) Listening to talking faces: motor cortical activation during speech perception. *Neuroimage 25*: 76-89.

Smith, E.L., Grabowecky, M., Suzuki, S. (2007) Auditory-visual crossmodal integration in perception of face gender. *Current Biology 17*: 1680–685.

Smith, M.L., Fries, P., Gosselin, F., Goebel, R., Schyns, P.G (2009) Inverse Mapping the Neuronal Substrates of Face Categorizations. *Cerebral Cortex 19*: 2428-438.

Sobotka, S., Ringo, J.L. (1996) Mnemonic responses of single units recorded from monkey inferotemporal cortex, accessed via transcommissural versus direct pathways: a dissociation between unit activity and behavior. *Journal of Neuroscience 16*: 4222–230.

Sokhi, D.S., Hunter, M.D., Wilkinson, I.D., Woodruff, P.W. (2005) Male and female voices activate distinct regions in the male brain. *Neuroimage 27*: 572-78.

Soken, N.H., Pick, A.D. (1992) Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. *Child Development 63*: 787–95.

Spence, C., Driver, J. (1997) Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics 59*: 1−22.

Spence, C., Driver, J. (2004) *Crossmodal Space and Crossmodal Attention*. Oxford: Oxford University Press.

Spence, C. (2002) Multisensory integration, attention and perception. In D. Roberts (Ed.) Signals and perception: *The fundamentals of human sensation* (pp. 345-354). Basingstoke, UK: Palgrave Macmillan.

Spence, C., Nicholls, M. E. R., Driver, J. (2001) The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics 63*: 330-36.

Spence, C. (2009) Explaining the Colavita visual dominance effect. *Progress in Brain Research 176*: 245-58.

Sreenivas, S., Boehm, S.G., Linden, D.E. (2012) Emotional faces and the default mode network. *Neuroscience Letters 506*: 229-34.

Stanford, T.R, Stein, B.E. (2007) Superadditivity in multisensory integration: Putting the computation in context. *NeuroReport 18*: 787–92

Stein, B.E., Meredith, M.A. (1993) *The merging of the senses*. Cambridge, MA: MIT Press.

Stein, B. E., Magalhaes-Castro, B., Kruger, L. (1976) Relationship between visual and tactile representations in cat superior colliculus. *Journal of Neurophysiology 39*: 410–19.

Stein, B.E., Stanford, T.R. (2008) Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience 9*: 255–66.

Stein, B.E., Stanford, T.R., Ramachandran, R., Perrault, T.J. Jr., Rowland, B.A. (2009) Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research 198*: 113–26.

Stein, B.E., Wallace, M.T. (1996) Comparisons of cross-modality integration in midbrain and cortex. *Progress in Brain Research 112*: 289–99.

Stern, C.E., Corkin, S., González, R.G., Guimaraes, A.R., Baker, J.R., Jennings, P.J., Carr, C.A., Sugiura, R.M., Vedantham, V., Rosen, B.R. (1996) The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. *Proceeding of the National Academy of Sciences of the United States of America 93*: 8660–665

Stevenson, R.A., Geoghegan, M.L., James, T.W. (2007) Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. *Experimental Brain Research 179*: 85–95

Sugase, Y., Yamane, S., Ueno, S., Kawano, K. (1999) Global and fine information coded by single neurons in the temporal visual cortex. *Nature 400*: 869-73.

Sugihara, T., Diltz, M.D., Averbeck, B.B., Romanski, L.M. (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *Journal of Neuroscience 26*: 11138–1147

Sugimura, T. (2006) How accurately do young children and adults discriminate the gender of natural faces? *Perceptual and Motor Skills 102*: 654-64.

Sumby, W. H. and Pollack, I. (1954) Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America 26*: 212–15.

Szycik, G.R., Jansma, H., Münte, T.F. (2009) Audiovisual integration during speech comprehension: an fMRI study comparing ROI-based and whole brain analyses. *Human Brain Mapping 30*: 1990-999.

Tal, N., Amedi, A. (2009) Multisensory visual–tactile object related network in humans: insights gained using a novel crossmodal adaptation approach. *Experimental Brain Research 198*: 165–82.

Talsma, D., Doty, T.J., Woldorff, M.G. (2007) Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cerebral Cortex 17*: 679–90.

Talsma, D., Woldorff, M.G. (2005) Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience 17*: 1098–114

Thielscher, A., Pessoa, L. (2007) Neural correlates of perceptual choice and decision making during fear–disgust discrimination. *Journal of Neuroscience 27*: 2908–917

Thierry, G., Martin, C.D., Downing, P.E., Pegna, A.J. (2007) Is the n170 sensitive to the human face or to several intertwined perceptual and conceptual factors? *Nature Neuroscience 10*: 802-03

Thompson, J.K., Peterson, M.R., Freeman, R.D. (2003) Single-neuron activity and tissue oxygenation in the cerebral cortex. *Science 299*: 1070–072.

Tiddeman, B., Perrett, D. (2001) Moving facial image transformations based on static 2D prototypes. Paper presented at the *9th International conference in Central Europe on Computer Graphics, Visualization and Computer Vision* 2001 (WSCG 2001), Plzen, Czech Republic.

Turner, B.H., Mishkin, M., Knapp, M. (1980) Organization of the amygdalopetal projections from modality-specific cortical association areas in the monkey. *Journal of Comparative Neurology 191*: 515-43.

van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L. (2004) Integration of letters and speech sounds in the human brain. *Neuron 43*: 271-82.

Van Essen, D.C., Dierker, D.L. (2007) Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron 56*: 209–25

van Lancker, D. (1997) Rags to riches: Our increasing appreciation of cognitive and communicative abilities of the human right hemisphere. *Brain and Language 57*: 1-11

van Wassenhove, V., Grant, K.W., Poeppel, D. (2005) Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America 102*: 1181-186

van Wassenhove, V., Grant, K.W., Poeppel, D. (2007) Temporal window of integration in auditory-visual speech perception. *Neuropsychologia 45*: 598–607

Vander Wyk, B.C., Hudac, C.M., Carter, E.J., Sobel, D.M., Pelphrey, K.A. (2009) Action understanding in the superior temporal sulcus region. *Psychological Science 20*: 771-77

Vizioli, L., Smith, F., Muckli, L., Caldara, R. (2010) Face encoding representations are shaped by race. Poster presented at the *16th Annual Meeting of the Organization for Human Brain Mapping*, Barcelona, Spain.

Vohn, R., Fimm, B., Weber, J., Schnitker, R., Thron, A., Spijkers, W., Willmes, K., Sturm, W. (2007) Management of attentional resources in within-modal and cross-modal divided attention tasks: an fMRI study. *Human Brain Mapping 28*: 1267-275.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.L. (2005) Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience 17*: 367-76.

von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, A.L. (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research 17*: 48-55.

von Kriegstein, K., Giraud, A.L. (2006) Implicit multisensory associations influence voice recognition. *PLoS Biology 4*: e326.

von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A.L., Kell, C.A., Grüter, T., Kleinschmidt, A., Kiebel, S.J. (2008) Simulation of talking faces in the human brain improves auditory speech recognition. *Proceeding of the National Academy of Sciences of the United States of America 105*: 6747–752.

von Kriegstein, K, Giraud, A.L. (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage 22*: 948-55.

Vroomen, J., Driver, J., de Gelder, B. (2001) Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective and Behavioural Neurosciences 1*: 382-87

Vogt, B.A., Vogt, L., Laureys, S. (2006) Cytology and functionally correlated circuits of human posterior cingulate areas. *Neuroimage 29*: 452–66.

Vuilleumier, P., Armony, J.L., Driver, J., Dolan, R.J. (2001) Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron 30*: 829–41.

Vuilleumier, P., Richardson, M.P., Armony, J.L., Driver, J., Dolan, R.J. (2004) Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neuroscience 7*: 1271–278.

Walker, A.S. (1982) Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology 33*: 514–35.

Walker-Andrews, A.S., Bahrick, L.E., Raglioni, S.S. Diaz, I. (1991) Infants' bimodal perception of gender. *Ecological Psychology 3*: 55 – 75.

Walker-Andrews, A.S. (1986) Intermodal perception of expressive behaviors: relation of eye and voice? *Developmental Psychology 22*: 373–77.

Walker, S., Bruce, V., O'Malley, C. (1995) Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics 57*: 1124–133.

Wallace, M. T., Wilkinson, L. K., Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *Journal of Neurophysiology 76*: 1246-266.

Wallace, M. T., Meredith, M. A., & Stein, B. E. (1998) Multisensory integration in the superior colliculus of the alert cat. *Journal of Neurophysiology 80*: 1006-010.

Wallraven, C., Breidt, M., Cunningham, D.W., Bülthoff, H. (2008) Evaluating the perceptual realism of animated facial expressions. *ACM TAP 4*: 1–20.

Wendelken, C., Ditterich, J., Bunge, S.A., Carter, C.S. (2009) Stimulus and response conflict processing during perceptual decision making. *Cognitive, Affective and Behavioural Neuroscience 9*: 434-47.

Werner, S., Noppeney, U. (2010) Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex 20*: 1829-842.

Wiethoff, S., Wildgruber, D., Kreifelts, B., Becker, H., Herbert, C., Grodd, W., Ethofer, T (2008) Cerebral processing of emotional prosody–influence of acoustic parameters and arousal. *Neuroimage 39*: 885–93

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., Ackermann, H. (2005) Identification of emotional intonation evaluated by fMRI. *Neuroimage 24*: 1233-241.

Winder, J., Darvann, T.A., McKnight, W., Magee, J.D.M., Ramsay-Baggs, P. (2008) Technical validation of the Di3D stereophotogrammetry surface imaging system. *British Journal of Oral and Maxillofacial Surgery 46*: 33-7

Woodman, G. F., Luck, S. J. (1999) Electrophysiological measurement of rapid shifts of attention during visual search. *Nature 400*: 867−69.

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., McCarthy, G. (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex 13*: 1034-034

Yin, R. K. (1969) Looking at upside-down faces. *Journal of Experimental Psychology 81*: 141-45.