Integrating naturalistic signals from audition and touch

By

Giulio Degano

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Computational Cognitive Neuroimaging Lab
Centre of Human Brain Health
School of Psychology
College of Life Science
University of Birmingham
October 2020

ABSTRACT

In everyday life, humans are exposed to a plethora of sensory inputs that form the so-called multisensory environment. To react appropriately, the brain combines information carried by the different senses. Although a valid line of ecologically-valid experiments has been conducted on the investigation of audio-visual integration, fewer studies have explored other types of cross-modal interactions in the same naturalistic setting.

The present thesis uncovered the neural correlates of audio-tactile pairing during the free-behaving perception of musical compositions. Specifically, fMRI and EEG data were analysed to (1) understand the temporal and spatial signature of audio-tactile binding; (2) assess the neural benefit of tactile stimulation during auditory scene analysis; (3) comprehend the modulation of cross-modal formations across different levels of awareness. To achieve such aims, neural activations in response to multisensory and unisensory conditions were collected during wakefulness and sleep.

The results of the neuroimaging analyses revealed that naturalistic audio-tactile interactions verify the neural criteria of multisensory object formation. Precisely, it is demonstrated that the audio-tactile binding involves low-level sensory areas and occurs at early time windows of integration [0-150ms]. In regard to the auditory scene analysis, the presented findings confirmed that the tactile signal boosts the representation of the congruent auditory stream during naturalistic scenarios. Finally, while this binding is shown to occur during wakefulness, it is suggested that it is modulated by different levels of awareness, with stages of deeper sleep cancelling out the neural multisensory benefit.

ACKNOWLEDGMENTS

Firstly, I want to thank my supervisor Uta Noppeny for the guidance and support that I received during these years of my PhD. I am grateful for the intellectual respect I was given, which was undoubtedly reciprocated. Moreover, I must admit that I enjoyed talking about our shared interests, not only in science, but also outside of it.

Second, I want to thank the reviewers of this thesis Dr Anne Keitel and Dr Hyojin Park. Reviewing a PhD work always takes some precious time so I am grateful that you have considered doing it. I hope you will enjoy the reading.

Thanks to (my) Computational Cognitive Neuroimaging Lab. Being around you made me a better researcher and it has always be fun. Steffen of course you are my honorable mention. A couple of days ago I went to room 3.14 to remove my last belongings and I sat for a while on my chair remembering our discussions. I will not either deny nor admit that a couple of tears were dropped.

Thanks to Ambra, as a coworker and as a friend. I have and always will consider your feedback on anything. You have been incredibly helpful throughout my PhD. Definitely my third unofficial supervisor.

Thanks to Arianna. I have enjoyed so much our conversation on life and science, and I still do. I really hope that at some point we will work together on a small or big project, but mostly I hope that we will live ”physically” closer to each other again.

Thank to Marco and Fede for your friendship. I am going to miss our time together a lot. The memories that I have of you two are priceless.

Thanks to Kate, Son, Affi and Tane. What an amazing family. You opened the door of

your house every time I needed it. You are definitely the main reason why I will come back to Birmingham in the next years.

Thanks to my climbing buddies, Blake and Stu. Sharing such an important part of my non-work time with you on rocks and plastic holds made my experience in the UK way sweeter than what I expected.

Thanks to my Italian friends in Birmingham. Bringing here the warmth of our culture was an incredible gift to survive the numerous moment of homesickness.

Thanks to Lorenzo. My brother from a different mother. Not much to add there, you know how important our bound is.

A massive thank you to my parents. I know that I was away most of the time, but our weekly Skype sessions were fundamental also for this project. It's incredible that, even if I am 29 year old, I still feel safe just by talking to you. I hope you are proud of this last work.

Finally, thank you to my soulmate Isotta. You are the bravest person I know. Not only you followed me on this island but you were able to accomplish more in three years than what humans do in ten. Our relationship resisted the distance and even a pandemic. I cannot get enough of you and I look forward for our life together, especially now that I have tricked you into saying "yes". I am lucky and I know it. (I don't deny that I might reuse some of this for my vows. I expect you to do the same.)

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

In every day life, humans are surrounded by multisensory inputs [Soto-Faraco et al., 2019]. Even when PhD candidates stare at their screens thinking about the introduction of a thesis on cross-modal integration, they are actually using the benefits of multisensory interactions while typing. In fact, when writing few lines, they absorb a wide spectrum of information ranging from the light touch of the keyboard to the visual representation of letters standing out on the blank pages. Thus, while it is of course important that each sense is studied by itself, it is undeniable that a multisensory perspective can offer a more comprehensive view on human perception. When an event or object is experienced by our brain, there is almost always some type of interaction between senses even when there is no clear awareness of one sensory modality[Mudrik et al., 2014].

Multisensory integration and its neural mechanisms have been vastly studied in the past 30 years (for a comprehensive overview [Calvert et al., 2004]), but, despite its original appeal as an ecologically-valid phenomenon, the majority of multisensory research have been conducted with controlled laboratory paradigms [Soto-Faraco et al., 2019]. One should in fact expect some limitations if the experience is not let free-behave: (1) it is likely to run into the risk of isolating one aspect of integration in the fear of confounding elements, while it is possible that these element are also naturally part of more complex interactions and (2) it simply does not represent the way humans explore the real world [Maguire, 2012, Matusz et al., 2019]. A naturalistic point-of-view becomes even more important when attentional processes are taken into account, since they are inherently part of complex scene analysis [Peelen and Kastner, 2014] and cross-modal perception [Macaluso et al., 2016] in everyday life. In this framework, music is an interesting example of an ecologically valid stimulus as it offers the possibility to investigate at the same time auditory scene analyses [Disbergen et al., 2018] and multisensory processes [Chuen and Schutz, 2016]. Hence, the focus of the present thesis will be on the neural correlates involved in the formation of multisensory interactions during music perception

and their relationship with attention and awareness. To this aim, audio-tactile integration in music will be investigated in the context of cocktail-party scenarios as well as its modulation during different level of awareness (i.e. wakefulness and sleep). The reason behind this choice was justified by the fact that audition and touch offer redundant information about loudness and frequency content of vibratory events, thus favouring the characterization of the musical content and promoting multisensory integration [Soto-Faraco and Deco, 2009].

1.1 The case of music

An important aspect of the current work is represented by the choice of music as a naturalistic stimulation to assess the integration of tactile and auditory information. Whereas speech has been extensively used for multisensory research, music has been rarely investigated despite the commonalities between these two. Indeed, a considerate amount of speech and music information is carried by slow temporal modulation of the sound in a frequency range that is almost comparable. For example, these slow fluctuations - that evolve over time - do not only carry information about beat and meter [Gordon, 1987, Scheirer, 1998, Large and Palmer, 2002] but also support speech intelligibility and comprehension [Wilsch et al., 2018, Shannon et al., 1995]. From a neural point of view, the sound intensity variation (more classically described by the *envelope*) was shown to ”entrain” or couple with the slow oscillatory activity in the brain [Nozaradan et al., 2011, Giraud and Poeppel, 2012, Doelling and Poeppel, 2015, Schroeder et al., 2008, Santoro et al., 2014, Harding et al., 2019]. More specifically, the coherence in phase, quantified by the amount of angular distance [1] between the oscillatory behaviour of the stimuli and neural activity, was suggested to be a fundamental element for the

[1] The smaller this angle is, the greater the brain activity and the stimuli are coupled together

correct perception of a target-stream [Shamma et al., 2011] and, more generally, an optimization mechanism behind listening behaviors [Henry and Obleser, 2012, Asaridou and McQueen, 2013, Ding et al., 2017]. In the context of multisensory integration, music and speech can be used to assess the increase in the coupling between cortical activity and envelope information, given by cross-modal interactions. Thus, in the following work, a set of two contrapuntal acoustic streams, one of which was matched with a vibrotactile stimulus, were used to assess whether neural and perceptual benefits could arise as a results of multisensory integration.

The cross-modal neural benefit in challenging scenarios represents the recurring theme of this thesis, but, before highlighting the aim of the next chapters, a more thorough overview is needed. In the following sections, the notions of the cocktail party problem, multisensory integration as well as the commonalities in hearing and somatosensory inputs will be reviewed. Then, an outline of how perception is modulated by awareness will be given. Finally, a summary of the two empirical studies and their scopes can be found at the end of the chapter.

1.2 Principles of multisensory integration

Let's imagine for a second the incredible number of multimodal stimuli to which a person is exposed while simply crossing a busy street at rush hour. It is undeniably very difficult to coherently organize this amount of information. In these types of context, multisensory integration reshapes neural and behavioural responses to improve what could be understood if each sensory input was processed by itself [Stein, 2012]. From a neural point of view, any region that responds to different types of sensory stimulations or creates disruption in the perception of more than one sense if lesioned or impeded, is considered as *multisensory* [Calvert et al., 2004, Schroeder et al., 2003, Bolognini

et al., 2013, Ghazanfar and Schroeder, 2006]. Traditionally, associational or *hetero-modal* areas, where information incoming from different sensory regions is integrated, are the superior temporal sulcus (STS), intraparietal sulcus (IPS), frontal cortex and parieto-occipital areas, as well as numerous subcortical regions [Calvert, 2001, Calvert et al., 2004, Driver and Noesselt, 2008]. Crucially, recent findings demonstrated that even early sensory areas, often associated with the processing of one modality, were modulated across senses [Macaluso, 2006, Driver and Noesselt, 2008, Schroeder et al., 2003]. Interplays between activation and inhibition of auditory, visual or somatosensory regions were elicited by non-specific sensory stimuli [Kayser et al., 2005, Foxe et al., 2000, Beauchamp and Ro, 2008, Caetano and Jousmäki, 2006, Schürmann et al., 2006, Lakatos et al., 2007, Campbell, 2008, Laurienti et al., 2002, Werner and Noppeney, 2010] showing multisensory interactions especially in deeper layers [Gau et al., 2020] to the point that the whole neocortex can be generally conceived as multisensory [Ghazanfar and Schroeder, 2006].

It has been established that true multisensory integration phenomena need to show non-linearity, such as superadditivity or subadditivity [Murray and Wallace, 2012]. In other words, when a cross-modal stimulus is processed, neurons must activate in a non-linear fashion to truly integrate different sensory modalities. This is a fundamental criterion under which neural responses to multimodal information are not just mere juxtaposition of unisensory activations. [Noppeney, 2012, Stein and Stanford, 2008, Pourtois et al., 2005]. This was in fact very clear from pioneering studies of the superior colliculus of the cat; when our sensory systems carry temporal and spatial information to the brain, excitatory or inhibitory responses are proportional to the amount of coherence shared across modalities [Meredith and Stein, 1983, Meredith and Stein, 1984, Meredith, 2002]. This non-linear interaction between unisensory responses in the brain plays a central role for cross-modal binding. Indeed, since the goal of this thesis is to investi-

gate audio-tactile objects formations, the responses -to audio-tactile stimuli- measured via neuroimaging techniques to audio-tactile stimuli should reflect super (or sub) additive interactions between somatosensory and auditory inputs in both early time windows [0-150 ms] and primary sensory areas (e.g. primary auditory cortex).

Although this thesis focuses on the neural correlates of multisensory integration, it is worth mentioning that the behavioural benefit of cross-modal interactions is very much present in everyday life and it has been corroborated in laboratory experiments since the beginning of the last century [Todd, 1912]. Similar to what has been shown on a neural level, evidence of multisensory benefits arises when signals from multiple sensory modalities are temporally, spatially or semantically congruent [Vroomen and de Gelder, 2000, Frassinetti et al., 2002, Yau et al., 2009]. Moreover, the principle of inverse effectiveness states that perceptual benefits of multisensory interactions are stronger when weak audio, tactile or visual signals are integrated together [Stein and Meredith, 1993]. For example, during target detection tasks, multisensory information can improve accuracy results especially when the reliability of unisensory signals are low [Odgaard et al., 2004, Gillmeister and Eimer, 2007, Wilson et al., 2010].

Taken together, this corpus of research shows that multisensory processes have functional relevance, since they facilitate the perception of congruent multisensory inputs in the context of environmental noise. In the next section, neural and behavioural multisensory effects will be discussed in relation to a classical scenario where our auditory perceptual abilities are challenged: the cocktail-party problem.

1.3 Auditory scene analysis and multisensory benefit

A particularly interesting scenario that impairs human hearing potential in everyday life is the cocktail-party problem [Cherry, 1953]. In a study from 1953, Cherry tested the ability of participants to filter two different speech streams that were recorded by the

same voice, which were presented simultaneously either to the left and right ears or played by the same sound source. He found out that participants were able to focus easily only in the condition were the two voices were spatially separable but not when they arose from the same loudspeaker. From these results it was evident that the brain was able to filter background noise if a separable parameter was found between acoustic streams (for example high/low pitch, left/right hear). It is in fact relatively easy to picture how our brain can be overloaded by the presence of multiple voices and how a lot of effort can be required to group or segregate their information.

This effect is not specific to speech only, but it involves also other types of naturalistic events. As mentioned in the previous paragraph, when listening to an orchestra or a rock song, our brain is able to focus on different instruments aided by the natural differences in timbre but also *captured* by other salient acoustic features that evolve in time. Indeed, the analysis of an acoustic scene (ASA, [Bregman and McAdams, 1990]) where multiple sources are overlapping, can be driven by selective-attention strategies (top-down focus on one specific melody or instrument) *or* by attention-grabbing events or stimuli [Bregman and McAdams, 1990]. In this work, for example, the segregation of a polyphonic composition (composed by two piano voices) will be aided by matching a tactile signal to one of the two acoustic stream, thus increasing the saliency of one voice over the other.

The importance of Bregman's seminal work on ASA is given by the understanding that our hearing system is able to structure and organize incoming sounds in order to process and detect statistical regularities on both long and short timescales. A large body of literature suggests that spectro-temporal regularities such as frequency [van Noorden, 1975], timbre [Singh, 1987], spatial location [Maddox and Shinn-Cunningham, 2012] or amplitude [Grimault et al., 2002] are drivers of those attentional mechanisms that are fundamental for our brain to resolve an auditory scene. Crucially, the features

belonging to the same sound cannot evolve independently, but must merge together in order to produce a pool of coherent unit elements on which selective attention can operate, thus forming what has been defined as an auditory *object* [Shinn-Cunningham, 2008]. This definition, inherited from the elegant work of Anne Treisman on vision, implies that an event or element is defined as an object when all its different characteristics (e.g. shape, colour and movement for vision but also envelope, pitch and timbre for audition) are jointly grouped automatically, even when only one of these features is attended to [Treisman and Gelade, 1980]. This creates a strong bond that preattentively allows the detection of auditory objects in a stream of information [Shinn-Cunningham, 2008]. Once identified, the auditory objects must be then maintained in time for the signal of interest to be tracked and segregated. Subsequently, selective attentional networks must intervene in order to disregard competing sounds reaching the auditory system, using mechanisms that are likely an interplay between top-down and stimuli-driven control [Middlebrooks et al., 2017]. Indeed, while numerous studies show the endogenous ability of humans to focus on speakers' spectro-temporal characteristics [Darwin, 2008, Maddox and Shinn-Cunningham, 2012], inherently salient stimuli such as the listener's own name can *grab* attention in a very fast and automatic way [Moray, 1959].

In addition to hearing, other senses can come together and increase the saliency of specific messages in a stream of information. Let's imagine for example trying to understand a single voice among a crowd; although auditory objects might be useful to segregate a particular signal, we are almost naturally driven to integrate information coming from the face of the speaker. It has been known for more than 50 years that humans have the ability to combine lipreading and speech sounds in order to facilitate communication [Sumby and Pollack, 1954]. In fact, linking cross-modal signals that originate from the same source is often of great benefit, especially when solving complex

scene analyses [Grant, 2001, Bernstein et al., 2004, Reisberg et al., 1987, Summerfield, 1992, Crosse et al., 2016b, O'Sullivan et al., 2019, Campbell, 2008]. It has been hypothesized that the integration of senses is indeed facilitated - if not even caused - by the statistics shared between the temporal envelope of sounds and the articulatory movements of the face [Chandrasekaran et al., 2009]. Indeed, due to these commonalities, lipreading can enhance neural activations coherent with the attended stimuli in cocktail party scenarios [Zion Golumbic et al., 2013]. Based on this large corpus of evidence, the extension of object-based attention theories to a more comprehensive multisensory perspective follows naturally, as Bizley et al. proposed in recent years [Bizley et al., 2016].

1.4 The multisensory object

In an elegant perspective from 2016, Bizley et al. provide the following definition of cross-modal object:

> "a perceptual construct which occurs when a constellation of stimulus features are bound within the brain"

This description echoes the theory of auditory objects discussed in the previous section and brings forward two main aspects of multisensory biding: (1) all characteristics of a cross-modal object must be coherent (2) attention to one of the dimensions of the object results in the automatic enhancement of all the other features of which it is composed.

Following their own definition, Bizley et al. proceeded to tackle the fundamental issue of cross-modal binding functioning. Among the theories proposed to address such question, the most promising is the one evolving around the assumption that features of an object have to maintain *temporal coherence* [Shamma et al., 2011]. This central criterion defines object identity and supports its segregation over time and, if violated,

creates a challenging setting for the brain to efficiently disentangle multiple auditory sources. Moreover, temporal coherence must occur also from a neural point of view: in order to perform stream formation, neurons need to enhance and maintain phase coherence with the stream that need-to-be segregated. Evidence of this mechanism has been shown in experiments where participants were asked to endogenously attend to specific auditory, visual and somatosensory streams [Steinmetz et al., 2000, Bidet-Caulet et al., 2007, Joon Kim et al., 2007].

Drawing upon this theory, Bizeley et al. suggested that the brain must enhance (or inhibit) the neural populations that are temporally coherent (incoherent) with the multisensory signal for the relevant message to be segregated from the background. They proceeded to demonstrate this framework in a recent study on audiovisual scene analysis in ferrets [Atilgan et al., 2018]. The animals were presented with a cocktail-party scenario where the envelope one of the auditory streams was matched over time with the luminance of a continuous visual stimulus. First, the results on concurrent cross-modal stimulation highlighted that neural populations were enhanced earlier when a congruent visual stimulus was applied. Second, the representations of the other sound features -belonging to the same audio-visual object- were also boosted even if they were not physically paired with the visual input. Crucially, these findings promoted the idea that the bottom-up (stimuli-driven) mechanisms determine audio-visual object formation and represent a cornerstone for complex scene analysis.

While this corpus of work focused mainly on speech processing and lipreading, hence providing evidence of audio-visual binding, the same can be easily extended toward other types of multisensory interactions. With this in mind, the evaluation of cross-modal binding during naturalistic auditory-scene analysis will be performed via shifting the spotlight to a less investigated pairing: the audio-tactile integration.

1.5 The lesser known cousin: Audio-tactile integration

Naturally, our hearing interacts with the other sensory systems to process the energy released during the occurrence of different events (e.g. loud speakers playing music in a pub) [Meredith, 2002]. Tactile and auditory signals are curiously both carried by the same physical phenomenon: oscillations of mechanical pressure, i.e. vibrations. Furthermore, the ranges of sensitivity to vibratory events overlap over a frequency window [Gescheider, 1997] that offers a redundant scenario, optimal for merging senses in one single multisensory formation [Ernst and Bülthoff, 2004].

Emerging evidence supporting interaction of audio-tactile (AT) signals has shown perceptual effects in both directions: on one side, auditory information has the ability to modulate the perception of touch roughness and frequency [Yau et al., 2009, Ro et al., 2009, Jousmäki and Hari, 1998, Guest et al., 2002, Zampini et al., 2007]; on the other side, somatosensations can influence the perception of loudness of the attended sounds [Yau et al., 2010, Murray et al., 2005]. Neuroimaging studies on AT integration have shown that hearing and touch are functionally linked to classical association areas, including the superior temporal and intraparietal sulcus [Leonardelli et al., 2015, Beauchamp and Ro, 2008, Kassuba et al., 2013, Schroeder and Foxe, 2002]. Moreover, low-level auditory and tactile areas have shown responses to multisensory interactions that are consistent with those occurring in response to other types of cross-modal interactions [Macaluso, 2006, Driver and Noesselt, 2008]. Electrophysiological studies in animals have shown that neurons respond to somatosensory stimuli in the auditory caudal belt [Schroeder et al., 2001, Schroeder and Foxe, 2002, Fu et al., 2003] and even in the primary auditory area [Lakatos et al., 2007, Kayser et al., 2005]. Similar results were also found in human neuroimaging studies where AT interactions were detected in auditory regions as well as second somatosensory areas (SII) [Kassuba et al.,

2013, Murray et al., 2005, Hoefer et al., 2013, Foxe et al., 2002, Beauchamp and Ro, 2008, Gobbelé et al., 2003, Caetano and Jousmäki, 2006, Butler et al., 2012, Foxe et al., 2000, Pérez-Bellido et al., 2018]. Crucially for the formation of cross-modal objects, recent results highlighted the existence of a preattentive coupling mechanism necessary for the representation of AT information [Butler et al., 2012].

The anatomical connections between the auditory and the somatosensory cortex have also been investigated. In non-human subjects, AT projections were found between the SII and the belt of the auditory cortex [Schroeder et al., 2001, Smiley et al., 2007, Cappe and Barone, 2005], while lesions in the somatosensory cortex were related to alterations in the responses to sounds observed in auditory areas [Higgins et al., 2008]. In humans, recent evidence has shown numerous ipsilateral connections between auditory and both primary and secondary somatosensory cortex [Ro et al., 2013]. Interestingly, the connections between SII and auditory regions were magnified in a case study of acquired AT synesthesia [Ro et al., 2013, Beauchamp and Ro, 2008]. Thus, it is likely that the anatomical configurations of these regions, which promotes short wiring length and fast conduction [Van Essen, 1997], also favours brain efficiency in perception of AT events [Bullmore and Sporns, 2012].

Finally, while it is clear that audition is very reliable in processing spectro-temporal information, it is also worth mentioning that it does not represent the only sensory system sensitive to rhythm. In a recent experiment, Tranchant et al. recorded body movements while dance music was played via speakers and vibrating platforms to groups of deaf and normal-hearing participants. Interestingly, they found that both groups were able to bounce at vibrations even without the experience of sounds [Tranchant et al., 2017]. These results suggest that, because of their ecological link to music, vibrotactile stimuli have the ability to carry rhythm information that can be useful for beat perception [Tranchant et al., 2017, Ammirante et al., 2016, Brochard et al., 2008].

Driven by the fact that the temporal and spectral dynamics shared by audio and tactile signals seem to offer a great case for stream segregation in auditory scene analysis, the first aim of this thesis is to assess the validity of audio-tactile integration as a candidate for cross-modal object formation.
Secondly, the same validity will be quantified in a context of absence of awareness. Specifically, because auditory and tactile signals are not physically affected when our eyes are closed -differently from the richness of visual information- the second aim of this thesis will evolve around the influence of awareness on AT binding in the context human sleep.

1.6 Sensory perception during sleep

Sleep is a well known state of behavioural unresponsiveness that is fundamental for all humans [Cirelli and Tononi, 2008]. Although this condition of sensory isolation poses an environmental risk because of no conscious control over potential threats, it offers an opportunity for the brain to undergo a significant level of recovery. Moreover, important cognitive functions such as memory consolidation and replay have been shown to occur systematically during sleep [Rasch and Born, 2013].

Sleep is characterized by two main phases: non-rapid eye movement (NREM) and rapid eye movement (REM) [Loomis et al., 1935]. The former is considered the deepest stage as the body shuts down and sensory isolation reaches its peak, the latter is instead marked by greater body movements and human dreaming. The main marker of NREM stage, which is detectable from the EEG signals even by naked eye, is the presence of slow-waves (SW). SWs are characterized by two oscillatory phases that alternate in frequency between 0.5 and 4 Hz and originate in the prefrontal cortex [Vya-

zovskiy and Harris, 2013]. The first phase is identified by membrane depolarization and increased neuronal firing, both of which drive the generation of spindles in the thalamus, and is called *upstate*. The second one is instead characterized by a general membrane hyperpolarization and neuronal silencing and is referred to as *downstate* [Steriade, 2003]. Specifically, the latter phase provides an inhibitory signal that is transmitted through the cortex and synchronizes spindles and hippocampal ripples [Oyanedel et al., 2020]. A subgroup of SWs is represented by the K-complexes, neural markers associated with the transition from light sleep (NREM1) to deeper unresponsive states (NREM2). K-complexes act in a similar fashion to SWs, with the sole difference that they induce only down-states of neuronal silencing [Cash et al., 2009]. Another classic marker observed during deep sleep stages is constituted by sleep spindles. These fast oscillations, which occur at frequencies ranging from 10 to 16 Hz, have been related to memory consolidation [De Gennaro and Ferrara, 2003] and firing modulation of neurons during exogenous stimulations [Elton et al., 1997]. Indeed, it has been shown that SW and sleep spindles are also generated as a consequence of sensory stimulation during animal sleep, further confirming their gating role in inhibiting and protecting the cortex during rest [Mccormick, 1994]. With the intent of understanding the role played by sleep oscillations, McCormick and Bal reached the conclusion that sensory isolation must be modulated by SWs and sleep spindles. Their Thalamic Gating Hypothesis does in fact suggest that the thalamus induces unresponsiveness by closing its relays to external stimuli and synchronizing to SWs and spindles, thus providing sensory protection to the brain [Mccormick, 1994].

Several aspects that the Thalamic Gating Hypothesis fails to integrate are (1) the fact that unresponsiveness is a state that can start already before sleep hallmarks [Ogilvie, 2001] and (2) what happens to exogenous signals during time windows when SWs and spindles are not present [Andrillon and Kouider, 2020]. Crucially, recent results

highlighted the access to external sensory information observed during sleep, suggesting that unresponsiveness does not necessarily correspond to sensory isolation: evidence showed processing of acoustic features [Portas et al., 2000, Atienza et al., 2001], mismatch negativity [Ruby et al., 2008] and even semantic violations [Bastuji et al., 2002, Ibáñez et al., 2006] during NREM stages. Despite these findings, it is clear that the windows of integration of sensory signals is limited during sleep [Ruby et al., 2008, Sharon et al., 2017] and that the brain seems to partly lose its predictive abilities on incoming information [Strauss et al., 2015]. For example, in a recent study on syntax, researches found out that while the auditory steady state signal evoked at the syllable level was comparable between wakefulness and NREM2, the neural tracking of longer structures -like phrases and sentences- was completely degraded during sleep [Sharon et al., 2017]. It is likely that these disruptions on more costly high-level computations are due to the inactivation of task-related regions during NREM stages [Muzur et al., 2002]. Complementary to these results, a recent perspective suggested that if the brain is able to bypass prefrontal areas via previous training, an *automatic* goal-oriented state could be reached even during unresponsiveness [Andrillon and Kouider, 2020]. In other words, since prefrontal areas are assumed to be involved in high cognitive functions that can be disrupted during sleep, it might be possible via mean of training, to automatize specific tasks that would diminish the role played by these regions. To demonstrate this idea, Legendre et al. presented participants with Jabberwocky and normal naturalistic stories in a cocktail-party scenario during different stages of awareness [Legendre et al., 2019]. During the wakefulness phase (training), participants were asked to selectively attend to the semantically meaningful speech, ignoring the Jabberwocky stream. During sleep, which constituted the actual testing phase, participants were presented with the same auditory streams, which were later decoded from the EEG signals. Interestingly, they were able to accurately decode the two speech signals during both light sleep and

the first stages of NREM.

Since multisensory binding supposedly evokes preattentive and automatic formations that are purely stimulus-driven [Bizley et al., 2016], it would become less costly for the brain to disentangle an auditory scene if audio and tactile signals are merged in a cross-modal object. Under this hypothesis, bottom-up processes that are driven by salient multisensory stimuli would bypass prefrontal areas, thus favoring selective attentional processes even during unawareness [Andrillon and Kouider, 2020]. The second part of the thesis will test this hypothesis by assessing the neural encoding of cross-modal music streams during sleep in a cocktail-party scenario.

1.7 Aim of the thesis

Considering the evidence discussed in the previous sections, the following empirical chapters will address the neural correlates of naturalistic audio-tactile integration in a cocktail-party scenario and its modulation by awareness.

Aim 1 In the first study, EEG and fMRI data were collected to address the temporal and spatial neural characteristics of audio-tactile object formation. More specifically, audio and tactile music signals were used to: (1) identify the time window for the occurrence of integration between the two sensory signals; (2) quantify the effect of temporal coherence in the features of multisensory stimuli; (3) determine the networks employed to sustain the formation of the AT object and importantly (4) understand whether a tactile signal is able to enhance the activation of coherent neural populations when more than one auditory source is present.

In the EEG experiment, neural tracking of music pieces was assessed with multivariate decoding. This technique has been shown to be a reliable reflection of neural activations that are temporally coherent with concurrent acoustic streams especially for decoding speakers in cocktail-party scenarios [Crosse et al., 2016a, O'Sullivan et al., 2015, Fiedler et al., 2017]. Moreover, linear decoders can be used to efficiently assess multisensory responses across different time windows [Crosse et al., 2015, Crosse et al., 2016b, Riecke et al., 2019].

In the fMRI, superadditivity effects were investigated to determine the location of non-linear cross-modal responses [Noppeney, 2012] during AT stimulations in and outside auditory scene analysis. Moreover, whole-brain analysis of the encoding of music envelopes was examined to uncover the role of congruency in audio-tactile object formation.

Aim 2 In the second project, the enhancement of neural tracking via tactile stimuli was assessed during cocktail party scenario at different level of awareness. This design allowed for the testing of multisensory saliency boosting during sleep.

Overnight EEG recordings were used to quantify the neural tracking of melodies and the decodability of multisensory streams in auditory scene analysis. Firstly, reconstructions of cross modal conditions were compared to those obtained from unisensory ones. Secondly, the temporal coherence between neural activations and the AT congruent stream was analyzed against the temporal coherence between the same neural activations and the incongruent competing stream, during both NREM1 and NREM2 sleep.

Chapter 2

Methods

This chapter will provide an overview of the different methods used in the present thesis, with a particular focus on the neuroimaging techniques and the multivariate approaches used to investigate multisensory integration from neural data.

The choice of these methods has been shown to be ideal in the context of processing naturalistic information [Naselaris et al., 2011, Alday, 2018].

2.1 Signal detection theory

Behavioural measures, although not the focus of the present thesis, were used for screening volunteers before the neuroimaging experiments. More specifically, signal detection theory was employed to assess participants' ability to differentiate between congruent and incongruent cross-modal stimuli. In a yes-no response task to the question "Are the audio and tactile stimuli congruent or not?", conditions were generated from two possible distributions: one defining the multisensory signal as produced by a common source and the other representing it as coming from two separate sources, respectively (see Fig.2.1).

Since both conditions (congruent or incongruent) had two possible responses (yes or no), trials could be grouped in four different categories: (1) "yes" when trials were congruent (*hit*) (2) "yes" when trials where incongruent (*false-alarm*), (3) "no" when trials were congruent (*miss*) and (4) "no" when trials were incongruent (*correct rejection*).

Response vs Condition	Congruent	Incongruent
Yes	Hit	False alarm
No	Miss	Correct rejection

The sensitivity index (d') of our screening test coincides with the ability of participants to discriminate between congruent and incongruent trials (separation between

distribution, see Fig. 2.1). The d' is computed as [Wickens, 2001]:

$$\mathrm{d}' = \mathrm{Z}(\mathrm{h}) - \mathrm{Z}(\mathrm{f})$$

where Z(h) and Z(f) is the z-normalization of the hit and false rate, respectively, and are defined as:

$$\mathrm{h} = \frac{\text{Numb. of hits}}{\text{Total numb. of congruent trial}} \quad \text{and} \quad \mathrm{f} = \frac{\text{Numb. of false alarm}}{\text{Total numb. of incongruent trial}}$$

A second measure employed to characterise participants' congruency judgements is the *criterion*$_{central}$. While the criterion (C) represents decision boundary between yes and no responses, *criterion*$_{central}$ gives the distance between C and the mid-point between the two distributions and hence quantifies decision biases:

$$\textit{criterion}_{central} = -\frac{Z(h) + Z(f)}{2}$$

It follows that:

- The greater the d', the greater is the perceptual ability of participants to define congruent and incongruent trials;
- *criterion*$_{central} = 0$ means that no perceptual bias is detected;
- *criterion*$_{central} > 0$ indicates that participants tend to perceive the multisensory stimulation as congruent (leaning towards "yes");
- *criterion*$_{central} < 0$ suggests that participants tend to perceive the multisensory stimulation as incongruent (leaning towards "no").

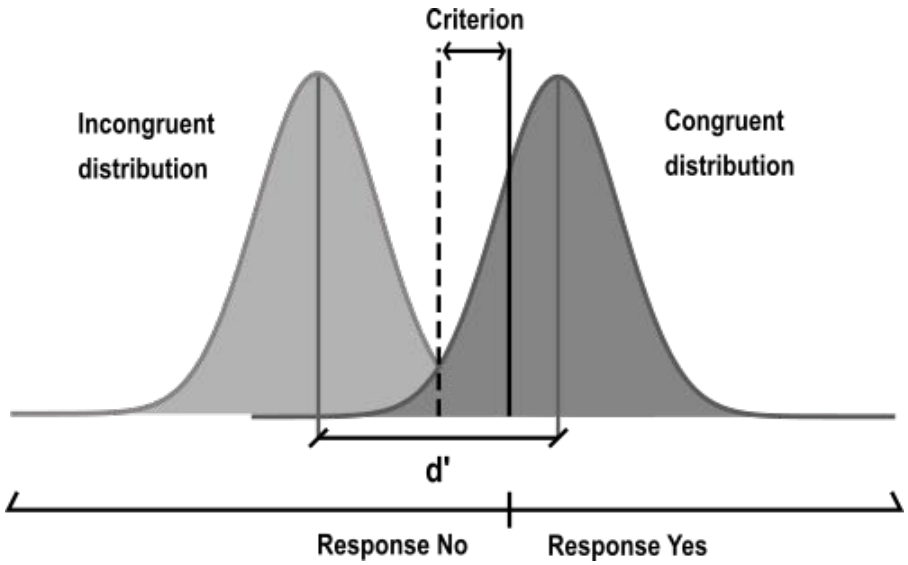

Figure 2.1: Congruent and incongruent signal distributions. The two indices used in this signal detection theory analysis are: (1) *d'* (also called sensitivity index), which represents the amount of separation between the two distributions; (2) $criterion_{central} > 0$ (or response bias), which represents the amount of deviation in response probability from the mid-point between the two signal distributions.

2.2 fMRI

2.2.1 Introduction to MRI imaging

It has been known for more than a century that brain regions activated in response to a specific task require an enhanced oxygenation, enabled by an increase in blood flow. This demand leads to the creation of local inhomogeneities that depart from equilibrium state, where the concentration of molecules of oxyhemoglobin and deoxyhemoglobin in brain tissues are normally similar. Due to this local exceedance of oxygen, scientists started to develop different imaging techniques in order to detect and exploit these variations in magnetic properties of hemoglobin.

Due to its ability to detect changes in the aforementioned properties of brain tissues, magnetic resonance imaging (MRI) became one of the most employed technologies to investigate cognitive functions. The basic principle on which MRI is based is the

property of dipoles to orient themselves when subjected to an external magnetic field. Specifically, the dipoles of hydrogen atoms consist of single protons spinning around their centers and create magnetic fields that, at equilibrium, are randomly oriented in our tissues. When these dipoles are placed in a strong magnetic field, usually labelled with notation B_0, they react by aligning to its direction and result in their spinning moments being parallel to each other.

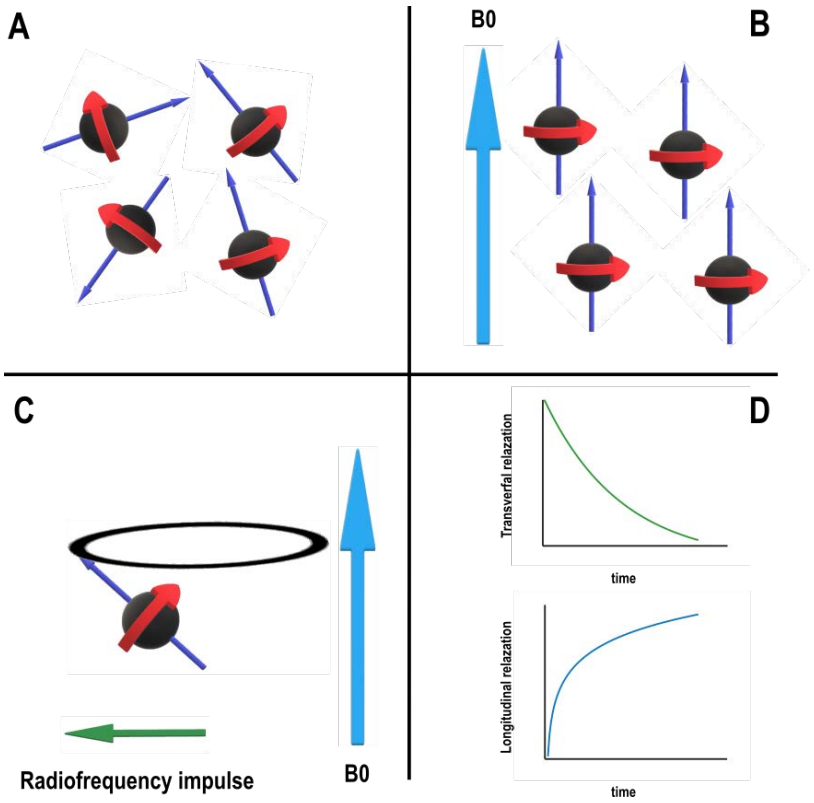

Figure 2.2: A) shows the random orientation of protons in tissues at equilibrium (red arrow is the direction of the spin, blue arrow the magnetic field). B) the protons align their magnetic field to the external magnetic filed B_0. C) When a radio-frequency impulse is applied (green arrow) the dipoles first align themselves to this orthogonal impulse and subsequently start spinning back (black circle). D) The blue plot represents the longitudinal realization (T1) when the protons move back to the direction parallel to B_0 after being subjected to the second impulse; the green one depicts the transversal relaxation.

During MRI acquisitions, an external B_0 is generated; once the dipoles are oriented

in parallel to B_0, a strong orthogonal impulse is applied to the dipoles. The latter firstly shift their magnetic fields to the direction imposed by the second impulse and then move back to their original positions, in a time frame named ”relaxation”. During the latter phase, protons spin in a non-synchronised fashion and give origin to a phenomenon known in physic as Larmor precession. This is described by two different constants T_1 and T_2 that represent the amount of relaxation time of dipoles in both the longitudinal and transversal planes (Fig,2.2).

Tissues, cerebrospinal fluid, bones, gray and white matter have different relaxation rates (T_1 and T_2) because of their characteristical distributions and structure of molecules. By acquiring MRI images at specific times within the relaxation phase, it is possible to capture the differences in response time between tissues and therefore obtain an image of the structure of interest (Fig2.3). The functional magnetic imaging (fMRI) exploits the dynamic differences in T2 relaxations between oxyhemoglobin and deoxyhemoglobin to obtain an image that reflects how brain regions recall oxygen while performing cognitive functions over time. In other words, fMRI acquisitions allow for the tracking of what is known as blood oxygenated level dependent (BOLD) signal.

In the present thesis the analysis of the functional data was conducted with a combination of open source software and custom code. Preprocessing and general linear model (GLM) were computed in Matlab (MathWorks Inc.) by using the common Statistical Parmetrical Mapping Toolbox (SPM, `https://www.fil.ion.ucl.ac.uk/spm/` [Friston et al., 2007]).

The brain atlases used to perform the segmentation of the brain were also open source and obtained from the Brainettome Atlas [Fan et al., 2016].

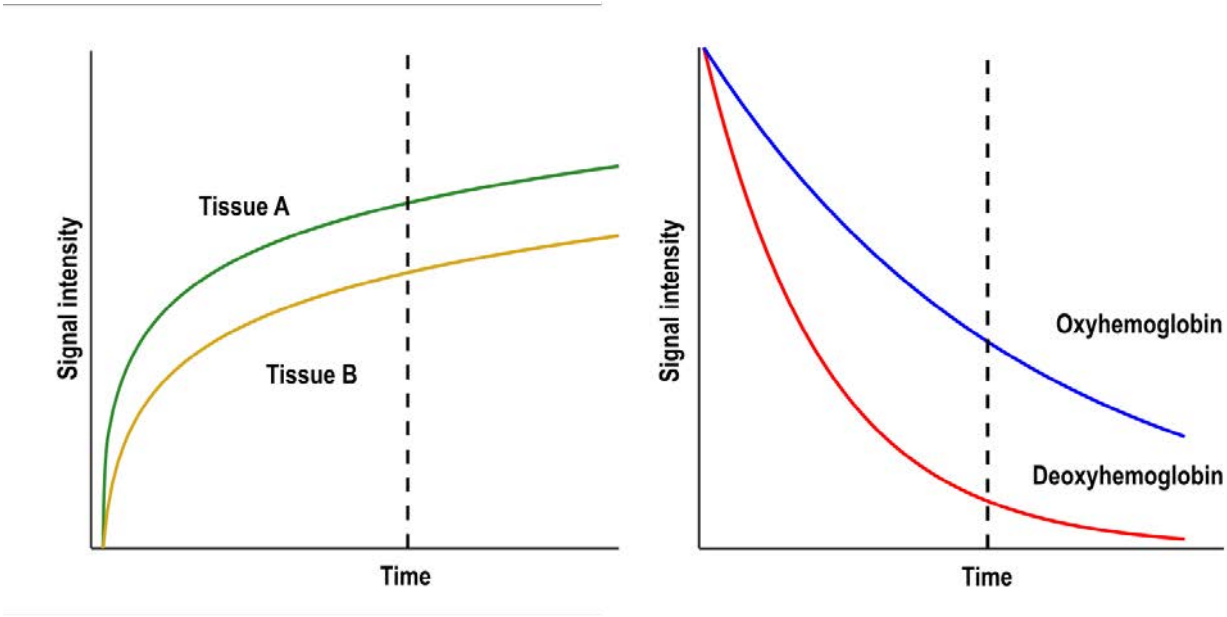

Figure 2.3: The timing at which the MRI images are taken should ideally maximise differences between the tissue properties. The left plot shows the difference in terms of T1 relaxation between two different tissues. The right one shows the difference in terms of T2 relaxation between deoxyhemoglobin and oxyhemoglobin.

2.2.2 fMRI Analysis: Preprocessing

The preprocessing of fMRI data is divided in two steps: temporal preprocessing and spatial preprocessing. The first is employed to balance differences in time acquisition between brain imaging slices while the second is used to rotate, smooth and standardize the brain volume.

Neural data obtained from fMRI imaging are organized in three-dimensional (3D) structures. These volumes are composed by slices acquired at different time points that sum up to the total repetition time (TR) of the fMRI sequence. In Chapter 3 *slice time correction* was used to obtain brain volumes with a time stamp interpolated to the central slice of the 3D images.

Spatial realignment and unwarping was performed to correct for head movements

and distortions present in the image. Since the head volume and shape do not change between acquisitions, the correction is based on rigid body transformations. The transformation matrix T is built as :

$$\mathbf{T} = \begin{bmatrix} \mathbf{cos}\alpha\mathbf{cos}\beta & \mathbf{cos}\alpha\mathbf{sin}\beta\mathbf{sin}\gamma - \mathbf{sin}\alpha\mathbf{cos}\gamma & \mathbf{cos}\alpha\mathbf{sin}\beta\mathbf{cos}\gamma + \mathbf{sin}\alpha\mathbf{sin}\gamma & \mathbf{X} \\ \mathbf{sin}\alpha\mathbf{cos}\beta & \mathbf{sin}\alpha\mathbf{sin}\beta\mathbf{sin}\gamma + \mathbf{cos}\alpha\mathbf{cos}\gamma & \mathbf{sin}\alpha\mathbf{sin}\beta\mathbf{cos}\gamma - \mathbf{cos}\alpha\mathbf{sin}\gamma & \mathbf{Y} \\ -\mathbf{sin}\beta & \mathbf{cos}\beta\mathbf{sin}\gamma & \mathbf{cos}\beta\mathbf{cos}\gamma & \mathbf{Z} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where α, β and γ are the rotation angles usually defined as yaw, pitch and roll and X, Y, Z the translation vectors.

Subsequently, the brain is *segmented* in three different tissue classes: white matter (WM), gray matter (GM) and cerebro-spinal fluid (CSF). Importantly, by separating these clusters, it is possible to apply tissue-specific deformations matrices that, in the following step, are being used to normalize voxels to the group-subject space. WM, GM and CSF prior maps are used to inform a Gaussian mixture model (GMM) and fitted to each subject data using the expectation-maximization algorithm. The final warping and *normalization* to the reference space (conventionally the Montreal National Institute space or MNI) is then computed using the clusters obtain from the GMMs.

Finally, *spatial smoothing* is applied to boost the signal-to-noise ratio and increase the normality of the errors, thus providing a better estimation of the parameters. To this end, a low-pass Gaussian kernel with user-selected full-width-at-half-maximum is convolved with the neural data.

2.2.3 fMRI Analysis: The General linear model

A GLM is the linear model used in Chapter 3 to characterize the relationship between each preprocessed voxel and the experimental paradigm. The main hypothesis at the core of the GLM is that the BOLD signal can be considered a linear-time invariant (LTI) system and, as such, is characterised with the following properties:

- it represents a linear relationship between the input **X** (the design matrix of the experiment) and the output **Y** (BOLD signal measures):

$$y(t) = \sum_k \beta_k x_k(t)$$

  where $y(t)$ corresponds to the response of each voxel at time t, $x_k(t)$ the k regressors at time t and β_k the coefficients that quantifies the magnitude of the linear relationship.

- it is time-invariant as the system behaves in the same fashion even if the input is shifted in time, so that if $x(t) \rightarrow y(t)$ then $x(t \pm T) \rightarrow y(t \pm T)$ for an arbitrary shift T.

The regressors inserted in the design matrix were represented by the different experimental conditions used in Chapter 3. The estimation of the β weights is computed by minimizing the cost function:

$$argmin_\beta[(y - \hat{y})^T(y - \hat{y})]$$

where $\hat{y}$ is the estimate obtained from βX.
Solving the equation in beta for each voxel (mass-univariate):

$$\beta = (X^T X)^{-1} X^T y$$

where X is the design matrix and y the time course of the voxel.

The statistical analyses were performed using a hierarchical summary statistic approach. Firstly, the GLM was estimated for each subject in a first level analysis; secondly, the statistical maps (e.g. t-test) were entered in a second level GLM as dependent variables. Contrasts on the latter GLM were evaluated to asses significant effects at the population level. Results were corrected at the cluster level (random field) [Friston et al., 1994a, Friston et al., 1994b].

2.3 EEG

2.3.1 Introduction to EEG imaging

While fMRI is characterized by a great spatial resolution that enables the localization of task-specific brain activations, it does lack temporal resolution. In fact, the majority of cognitive processes occur over a time span that ranges between milliseconds to few seconds and their early dynamics are impossible to track by the sole employment of hemodynamic response functions. Electroencephalography (EEG) is instead a great tool to address these fast dynamics because it offers a bigger temporal resolution [Cohen, 2011].

Another advantage of using EEG is that, while it measures a population of thousands of neurons at the same time, it is also one of the few non-invasive techniques to directly quantify neural activity. It has already been established that population-level recordings like EEG signals can be modelled quite accurately [Buzśaki and Wang, 2012].

The *magnitude* of electric dipoles, recorded by EEG, is a reflection of the postsynaptic potentials of pyramidal neurons in the cortex, while their *direction* depends on the type of synapses that can be either inhibitory or excitatory. However, it is impos-

sible to distinguish between these types of flow since the overall signal measured is a summation of multiple postsynaptic currents.

Although EEG signals are susceptible to local inhomogeneities in the electric field cause by the presence of different biological materials (e.g. skull, skin), electroencephalography is sensible to dipoles that are positioned tangentially and radially with respect to the electrodes. These characteristics give EEG systems a great advantage to measure deep and early activations, such as the auditory-evoked brainstem responses.

Because of the aforementioned reasons, EEG was recorded to assess the temporal dynamics of multisensory integration in wakefulness and sleep in the empirical chapters of this thesis. Since these signals are also affected by non-brain activity such as muscular and eye movements and even electric artifacts from house-lines, it is important to thoroughly preprocess the data before analyzing neuronal responses. The processing of the data was done following previous literature and code from our research group (see e.g. [Aller and Noppeney, 2019, Zumer et al., 2020]).

2.3.2 EEG preprocessing

The following steps were followed for the preparation of the EEG data:

- stop-band filters were applied at 50 Hz ($\pm$ 2Hz) and harmonics to the raw signal. These frequencies correspond to the electric line frequency range present in UK households;
- eye movement removal was performed via independent component analysis; fastICA from fieldtrip toolbox (visual inspection of 20 best components).
- visual inspection was carried out to identify trials corrupted by muscle artifacts: such trials were excluded from the successive analyses;
- visual inspection of retained trials was performed to locate channels with a low

signal-to-noise ratio: such channels were replaced with the linear interpolation of the signals from the neighbouring electrodes;

- the clean EEG signals were finally down-sampled to 100Hz.

2.3.3 EEG multivariate analysis

In the following section, multivariate regression analysis for EEG will be discussed, with details on feature selection, statistical learning and cross-validation procedures.

2.3.4 Selection of the feature of interest: Music Envelope

As mentioned in the introduction, one of the fundamental aspect of cross-modal binding is the temporal coherence within multisensory objects as well as between stimuli and neural populations [Bizley et al., 2016]. A straightforward way to characterize the evolution over time of these features was to employ music as a mean to facilitate cross-modal object formations. It is well established that the envelope of a sound stream is strictly connected to the onset and offset of the notes composing it, carrying information about its meter and beat [Gordon, 1987, Scheirer, 1998, Large and Palmer, 2002]. Indeed, neural slow oscillations coherent with sound amplitude fluctuations over time are thought to reflect an optimization mechanism behind listening behaviours [Henry and Obleser, 2012, Asaridou and McQueen, 2013, Ding et al., 2017] and have been considered relevant for both music and speech perception [Nozaradan et al., 2011, Giraud and Poeppel, 2012, Doelling and Poeppel, 2015, Schroeder et al., 2008, Santoro et al., 2014].

In the next empirical chapters, the envelope of the sounds will be the extracted with a three phase procedure [Yang et al., 1992]:

1. the music signal was filtered with a band pass filter-bank between 100-8000Hz,

logaritmically spaced in eight non-overlapping bands;

2. the output of each band-pass filter was Hilbert-transformed[1];

3. the final envelope is obtain by averaging the eight absolute values of the transformed signals.

2.3.5 Decoding the sound

A multivariate backward approach was used to asses multisensory integration that occurred in the brain. This solution is optimal when dealing with multicollinearity and low signal-to-noise ratio [Haufe et al., 2014]. In particular, it is very advantageous for detecting differences between correlated EEG channels, producing a better estimation of the amount of information encoded in the brain.

The main assumption in multivariate analyses is the modelling of the brain responses as linear systems:

$$\hat{\mathrm{S}}(t) = \sum_{\mathrm{ch}=1}^{\mathrm{N_{channels}}} \sum_{\tau=-\mathrm{K}^{K}} \mathrm{r}(\mathrm{t}+\tau, \mathrm{ch})\mathrm{g}(\tau, \mathrm{ch})$$

where ch represents the EEG channel, τ is the time lag of the brain response measured at each channel and $g(\tau, ch)$ depicts the modelled impulse response function of the brain.

The previous equation can be rewritten in a matricial form:

$$\mathbf{S} = \mathbf{gR}$$

The cost function to minimize is a least squared estimation with constraints on **g**:

$$argmin_g \left\|\mathbf{gR} - \mathbf{S}\right\|^2 + \lambda \left\|\mathbf{g}\right\|^2$$

[1]The Hilbert transform gives a frequency transformation of a real signal. By shifting by 90 degrees the phase, it eliminates the carrier frequency leaving the modulator slow signal

The reason for these constraints is driven by the fact that the regularization term λ deals with collinearity in the EEG data in a more efficient way. Solving in **g**:

$$\mathbf{g} = (\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I})^{-1}\mathbf{R}^T\mathbf{S} \tag{2.1}$$

where λ is the Tikhonov regularization or Lagrange multiplier, **I** is is the identity matrix and **R** is the lag matrix. The latter is defined as:

$$\begin{bmatrix}
r_1(\tau_{max}+1) & \dots & r_N(\tau_{max}+1) & r_1(\tau_{max}) & \dots & r_N(\tau_{max}) & \dots & r_1(1) & \dots & r_N(1) \\
\vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots \\
r_1(T) & \dots & r_N(T) & r_1(T-1) & \dots & r_N(T-1) & \dots & r_1(T-\tau_{max}) & \dots & r_N(T-\tau_{max}) \\
0 & \dots & 0 & r_1(T) & \dots & r_N(T) & \dots & r_1(T-\tau_{max}+1) & \dots & r_N(T-\tau_{max}+1) \\
0 & \dots & 0 & 0 & \dots & 0 & \dots & r_1(T-\tau_{max}+2) & \dots & r_N(\tau_{max}+2) \\
\vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & 0 & \dots & \vdots \\
0 & \dots & 0 & \dots & \dots & 0 & \dots & 0 & \dots & 0
\end{bmatrix}$$

where T is the length of the EEG signal, N the number of channels and τ_{max} the max number of lags considered in the estimation. The Matlab functions use to compute the temporal responses were taken from the mTRF toolbox `https://github.com/mickcrosse/mTRF-Toolbox.git`.

The brain response of each channel and time lag is then quantified with Equation 2.1. Finally, **g** is used to estimate the reconstruction accuracy score:

$$\text{corr} = \frac{\sum_{i=1}^{N}(s_i-\bar{s})(\hat{s}_i-\bar{\hat{s}}_i)}{\sqrt{\sum_{i=1}^{N}(s_i-\bar{s})^2}\sqrt{\sum_{i=1}^{N}(\hat{s}_i-\bar{\hat{s}})^2}}$$

which value has been shown to reflect the level of attention, tracking and processing of sensory information over time [Mesgarani and Chang, 2012, Ding and Simon, 2012, Ding and Simon, 2014, Crosse et al., 2016a].

2.3.6 Cross validation

The tuning of the ridge parameter λ and the testing of the linear kernel were done via a nested cross-validation procedure. This allowed the statistical learning of multivariate parameters to avoid overfitting of the EEG data, thus favoring greater generalizability of the results. An 8-fold cross-validation was used in study 1 and 2 to avoid anticorrelation effects that commonly occur in leave-one-out methods [Poldrack et al., 2020]. To do that, the following algorithm was used:

Algorithm 1: Algorithm Nested Cross-validation

Result: Accuracy for each outer fold

for $i \leftarrow 1$ **to** 8 **do**

  select OuterTest set i;

  select OuterTraining set i;

  for $k \leftarrow 1$ **to** N_{lambda} **do**

    initialize λ(k);

    for $j \leftarrow 1$ **to** 8 **do**

      select InnerTest set j;

      select InnerTraining set j;

      learn kernel(λ) on InnerTraining;

      assign accuracyIn(j) from InnerTest;

    end

    compute mean(accurancyIn);

    assign $\lambda_{avg}(k)$ = mean(accuracy);

  end

  compute LAMBDA = max(λ_{avg});

  learn kernel(LAMBDA) on OuterTraining;

  assign accuracyOut(i) from OuterTest;

end

Where N_{lambda} correspond to the max value of lambda used in the grid search.

Chapter 3

Audio-tactile integration in music: an EEG and FMRI study

Giulio Degano, Ambra Ferrari, Uta Noppeney

Computational Cognitive Neuroimaging lab, Centre of Human Brain Health, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, Birmingham, UK

Citation:

Degano, G., Ferrari, A., Noppeney, N. (in preparation). Audio-tactile integration in music: an EEG and FMRI study.

Authors contributions:

Experiment conceptualisation and design: Giulio Degano, Ambra Ferrari, Uta Noppeney.

Data collection: Ambra Ferrari, Giulio Degano.

Data analysis: Giulio Degano (supervised by Uta Noppeney).

Writing: Giulio Degano (supervised by Uta Noppeney).

3.1 Abstract

The term cross-modal binding indicates the automatic process for the representation of multisensory information in the brain. This phenomenon has been inherited from the auditory object theory used to describe the foundations on which attentional mechanisms are employed to solve complex scene analysis. Indeed, fast stimulus-driven attentional mechanisms evoked by audio-visual signals have been shown to facilitate the segregation of a cocktail party problem. Yet, the generalization of audio-visual objects to other type of cross-modal interaction is still unexplored. In this experiment, the extent of audio-tactile (AT) binding was investigated during the processing of naturalistic musical compositions. Next, the tactile enhancement of neural activations was decoded during cocktail-party conditions.

Temporal and spatial components of the audio-tactile object were assessed using a combination of EEG and fMRI analyses. Using multivariate decoding methods, the neural tracking of music information was quantified for unisensory and multisensory conditions, within and outside the auditory scene analysis. The amount of envelope information encoded in the brain was investigated using a GLM of the fMRI data. Results from the neuroimaging analysis showed that AT binding formations occurred at early timescales of the sensory processing [0-150 ms] and that multisensory interaction modulated activity in the early auditory areas. Crucially, the tactile signal evoked an enhancement of those neural activations that were temporally coherent with the AT stream during the cocktail-party condition.

These results show that the neural criteria of cross-modal binding can be extended to audio-tactile signals, uncovering a new scenario for the investigation of the multisensory benefits in auditory scene analysis.

3.2 Introduction

One of the most fundamental abilities of the human brain is to select and organize the mixture of sensory inputs it encounters in real life. A classic example of how our attention is shifted to relevant information in a complex auditory scene is the cocktail-party scenario [Cherry, 1953, Bregman and McAdams, 1994]. It is solidly established that humans can enhance their ability to select a specific stream of information by linking cross-modal signals, especially when processing naturalistic stimuli like speech [Sumby and Pollack, 1954, Grant, 2001, Bernstein et al., 2004, Reisberg et al., 1987, Summerfield, 1992, Crosse et al., 2016b, O'Sullivan et al., 2019, Park et al., 2016]. In fact, it has been suggested that lipreading and the temporal envelope of sounds enhance the neural tracking of speech units due to shared temporal statistics across sensory modalities [Chandrasekaran et al., 2009, Zion Golumbic et al., 2013]. In a broader sense, "redundancy" of information is a key aspect for the theory of cross-modal binding [Busse et al., 2005, Bizley et al., 2016]. In this framework, the brain does not simply integrate together coherent signals but has also the ability to group different statistical features across a pool of sensory inputs, hence creating a multisensory link that favours stronger perceptual benefits across all the characteristics of an object [Bizley et al., 2016].

It is almost indisputable that, at the neural level, multisensory integration occurs already at the bottom of the auditory cortical hierarchy [Fishman and Michael, 1973, Driver and Noesselt, 2008, Molholm et al., 2002, Kayser et al., 2005, Kayser et al., 2007, Macaluso et al., 2000, Noppeney and Lee, 2018, Calvert et al., 1999, Martuzzi et al., 2007] and at early temporal latencies [Crosse et al., 2015, Crosse et al., 2016a, Riecke et al., 2019, Luo et al., 2010]. Converging evidence in hearing research support the idea of cross-modal integration happening in early sensory areas. In particular, there appears to be: (1) anatomical connections between afferent non-auditory re-

gions and the auditory cortex; (2) activations of classical auditory neural populations by non-specific modalities (e.g. auditory cortex activation due to lipreading) (see [Musacchia and Schroeder, 2009] for a review). These feedforward streams can subsequently help form multisensory objects and orient the spotlight of attention on coherent salient stimuli, hence favouring higher level cross-sensory associations [Foxe and Schroeder, 2005, Macaluso et al., 2016, Noppeney and Lee, 2018]. In an attempt to combine these ideas, Atilgan el al. elucidated the formation of multisensory binding in auditory areas and its benefits [Atilgan et al., 2018]. In their interesting study on ferrets, it was demonstrated that a temporally coherent audio-visual stream (envelope and luminance fluctuating together as a function of time) can enhance stimulus representation on early cortical areas. Thus, these results might be interpreted as a supporting mechanism for the segregation of a complex mixture of sounds.

While a lot of recent literature has focused on audio-visual object formation -guided also by a logical interest towards speech perception and comprehension- audio-(vibro)tactile integration has been less investigated although both sensory systems share a substantial number of similarities [Soto-Faraco and Deco, 2009]. Firstly, vibrotactile and auditory perception intersect in terms of frequency range [Gescheider, 1997]. Secondly, it can be argued that both sensory inputs are an epiphenomenon of the same physical occurrences, namely oscillations of mechanical pressure. Functionally, both hearing and touch modulate neural activity in the reciprocal sensory areas [Kayser et al., 2005, Foxe et al., 2000, Beauchamp and Ro, 2008, Caetano and Jousmäki, 2006, Schürmann et al., 2006, Lakatos et al., 2007] and, at the same time, interfere -or enhance- the perception of the other modality [Wilson et al., 2010, Yau et al., 2009, Jousmäki and Hari, 1998]. Furthermore, there is evidence suggesting the existence of cross-modal connectivity between somatosensory areas and primary auditory cortex in both non-human [Schroeder et al., 2001, Cappe et al., 2009] and human [Ro et al., 2013] individuals. Moreover, it

has been proposed that vibrotactile stimuli, due to their ecological link to music, carry information that are useful for beat and rhythm perception [Tranchant et al., 2017, Ammirante et al., 2016, Brochard et al., 2008].

Taken together, this evidence led to the development of the present study, which examines audio-tactile integration during perception of naturalistic music stimuli. The reason behind the experimental design is twofold. In the first place, temporally coherent audio-tactile stimulation can favour the formation of multisensory objects without linguistic confounds that are usually present in speech research [Davis and Johnsrude, 2003, Hickok and Poeppel, 2007, Broderick et al., 2018]. Secondly, uncontrolled but ecologically valid experiments have shown promising results to reliably asses the encoding of music features in the brain [Hausfeld et al., 2018, Burunat et al., 2015, Hoefle et al., 2018, Di Liberto et al., 2020].

Electroencephalography (EEG) and functional magnetic resonance (fMRI) were used to address the questions of (1) whether audio-tactile stimuli can aid the formation of strong cross-modal binding and (2) which networks would be involved in favouring stimulus-induced attentional effects in cocktail-party scenarios. These two techniques were chosen because of their complementary strengths since, taken together, they offer solid temporal and spatial resolution. In the EEG analyses, neural tracking of musical pieces was quantified via multivariate reconstruction of the envelope of each melody, in line with previous research [Riecke et al., 1995, Mesgarani et al., 2009, Crosse et al., 2016a]. Crucially, to assess the temporal dynamics of multisensory integration, the decoding performances of superadditive linear models were compared with the congruent audio-tactile reconstruction of the envelope across different time windows [Crosse et al., 2015, Crosse et al., 2016b, Riecke et al., 2019]. In the fMRI analysis, non-linear additivity in brain activations was examined as it offers a strict marker of multisensory integration at the neural level [Noppeney, 2012]. Moreover, as the goal of this study

revolved around the assessment of the coherence between the sensory signals and its influence on the encoding of temporal information carried by each melody [Hoefle et al., 2018], the linear mapping of the envelope in each voxel was investigated in congruent and incongruent trials.

3.3 Materials and Methods

3.3.1 Participants

After giving written consent, 12 volunteers participated in the both the EEG and fMRI study (9 females, age mean = 27.75 SD = 4.08). The sample size was consistent with the number of participants reported in previous research in music perception [Alluri et al., 2012, Burunat et al., 2015, Hoefle et al., 2018, Hoefle et al., 2018, Nozaradan et al., 2011]. None of the subjects reported any history of neurological or psychiatric conditions. All volunteers were right handed based on the Edinburgh Handedness Inventory [Oldfield, 1971] (mean Laterality Quotient ($L.Q.$) = 85, with $L.Q. \in [-100, 100]$). The volunteers were reimbursed for taking part in the experiment based on the amount of hours spent in the laboratory (8£ per hour). The study was approved by the University of Birmingham Ethics Committee.

Participant inclusion and exclusion criteria

In order to evaluate the level of musical training, the MusicUSE (MUSE) questionnaire was used to score the amount of hours spent by each participant listening, learning and practicing music. The Index of Music Instrument Playing (IMIP) for all subjects resulted $< .4$, satisfying the exclusion criteria for non-trained musician.

Because audio-tactile congruency was a factor of interest in our study, participants performed in a congruency judgment task to assess their suitability for the experiment.

See Section 3.3.11 for more information about perceptual inclusion criteria. 12 participants were screened and all were included in the experiment.

3.3.2 Stimulation

Twenty-four different counterpoint Musical Instrument Digital Interface (MIDI) files were custom-made by a professional composer. From each MIDI file, two mono voices (high/low pitch melody of the composition) were synthesized at 44100 Hz using Linux MultiMedia Studio 1.1.3 (LMMS 2004, `github.com/LMMS/lmms`) with a Yamaha YDP piano sound font. The tempo at which the stimulus was generated was fixed at 60 bpm to facilitate non-musicians in the detection of temporal modulations in the piano tracks. From the same MIDI score, two mono tracks were synthesized using the default options on the Triple Oscillator of LMMS. These two latter files were later used to drive the tactile stimulator. The Triple Oscillator was set one octave lower than the original score in order to respect the vibratory limits of the piezoelectric simulator that was used for the experiment.

To create a perfect match between the audio and tactile signals, the envelope of the piano melodies was first extracted by computing the root-mean-square amplitude on each piano sound. Secondly, a moving average filter with windows of 20ms was applied to the envelopes to avoid impulsive signal content. Finally, the processed signal was normalized to the maximum of its envelope and used to modulate the tactile carrier frequencies (Fig.3.1).

Incongruent monophonic trials were created by minimizing the correlation between audio envelopes. To do so, the correlation between each tactile and audio track was computed for 1128 paired permutations of melodies. The combination that gave the lowest overall correlation score was selected as the candidate for the incongruent trials.

The envelope extraction, normalization and permutation were computed with a com-

bination of custom MATLAB code (MathWorks, 2019a) and Audacity (The Audacity Team, 2019).

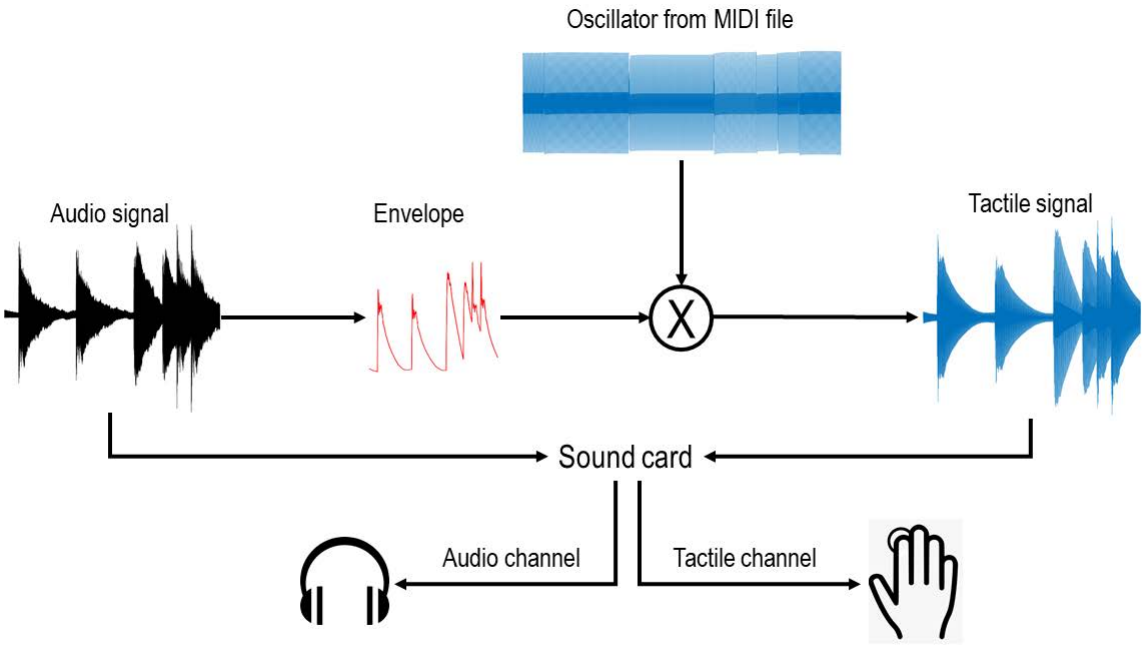

Figure 3.1: Two different kinds of stimulation resulted from the MIDI files. The audio envelope was extracted to modulate the amplitude of the tactile signal. The same sound card was used to control in parallel the sounds played diotically on the headphones and the vibration of the piezoelectric stimulator. Participants placed the index finger of each hand on one piezoelectric actuator to perceive vibrations via fingertips.

3.3.3 Experimental design and procedure

Participants perceived 6 different stimuli, consisting of 4 monophonic and 2 cocktail party conditions. The former conditions consisted of Audio (A), Tactile (T), Audio-Tactile congruent (ATc) and Audio-Tactile incongruent (ATi) trials. The latter conditions consisted of unisensory Audio cocktail-party (Acp) trials, where two concurrent monophonic pieces were presented, and multisensory Audio-Tactile cocktail-party (ATcp) trials, where a tac-

tile monophonic piece matched one of the two monophonic tracks presented in the auditory modality (Fig.3.2).

The stimuli had a duration of 28 s and were presented with a fixed inter-trial interval of 6s for the fMRI and 2s for the EEG. In order to avoid habituation or prediction effects, the order in which conditions were presented was counterbalanced between runs for each participant.

To control for vigilance and to avoid any task-modulatory effects during stimulation, participants responded to random interspersed full-screen flashes (luminance: 85 cd/m2; duration: 50 ms) by pressing a pedal positioned under their right foot. In each run, 7 flashes were presented across 5 different blocks (2 blocks contained 2 flashes, 3 block only one). Flashes were presented with two constraints : first, no flash was presented in the first or last 2 seconds of the stimulation block; second, the minimum time gap between two flashes was set to 2 seconds.

During the fMRI study, each condition was presented 48 times across 16 different runs (2 days of recording, 8 runs per day). In the EEG study, each condition was presented 48 times, but across 12 different runs (2 days of recording, 6 runs per day). The response to the visual input was considered a hit if the participants pressed the pedal in the range of $[100ms - 2000ms]$ after flash onset (group average accuracy score: 0.893 ± 0.002).

3.3.4 Experimental setup

Stimuli were presented using Psychtoolbox version 3.0.15 [Brainard, 1997] (`http://psychtoolbox.org/`) under MATLAB R2018b (MathWorks) on a laptop running Linux Ubuntu.

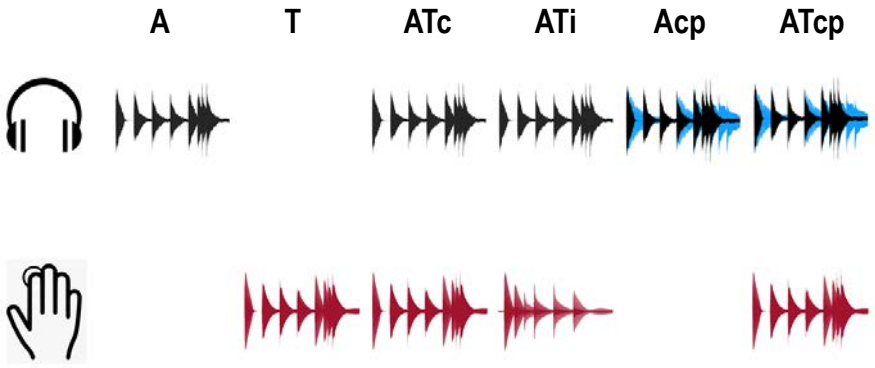

Figure 3.2: Six conditions were presented to the participants: two conditions consisted of monophonic unisensory Audio (A) and Tactile (T) trials; two conditions consisted of monophonic multisensory congruent (ATc) and incongruent (ATi) trials; two conditions consisted of cocktail-party trials. The latter were either unisensory Audio (Acp) or multisensory Audio-Tactile, where the tactile signal matched one of the two auditory streams (ATcp).

EEG setup Visual flashes were presented via a 30in LCD monitor with a resolution of 2560 x 1600 pixels at a frame rate of 60 Hz. Auditory stimuli were presented diotically at a sampling rate of 44.1 kHz via EEG compatible earplugs (EARTONE, Insert Earphone 3A) and an ASUS Xonar DSX sound card. The same sound card was used to drive the tactile vibrations (piezoelectric system (PTS-C2, Dancer Design, UK)). The stimulation was applied to both hands to the fingertip of each index finger. Participants rested their head on a chin rest at a distance of 600 mm from the monitor and at a height that matched participants ears to the horizontal midline of the monitor. Participants responded by pressing a pedal with their right foot (SODIAL, Shenzhen IMC Digital Technology Co.). Background white noise was additionally played through external speakers (65 dB sound pressure level) to mask eventual sounds from the tactile vibrations.

fMRI setup The same ASUS sound card was used to drive the auditory stimuli played through an MR-compatible system (SOUNDPixx MRI pneumatic transducer and amplifier VPX-ACC-8100, QC Canada; MRIaudio in-ear headphones, USA) and the tactile

one reproduced by the piezoelectric stimulator.

Visual flashes were back-projected onto a Plexiglas screen using a Barco Present-C F-Series projector (F35 WUXGA, UK; 1920 x 1024 pixels resolution; 60 Hz frame rate) and they were visible to the participants via a mirror mounted on the MR head-coil (horizontal visual field of $\sim 40^{\circ}$ visual angle at a viewing distance of ~ 68 cm).

Participants responded by pressing any keys of an MR-compatible keypad (NATA LXPAD 1x5-10M, BC Canada) attached to the right foot with elastic cohesive bandage and secured via foam supports.

3.3.5 Feature extraction

Spectral and temporal information were extracted from each melody separately. The envelope was computed by first band-pass filtering each musical piece with a filter bank of 8 logarithmically-spaced filters between 100-8000 Hz. The Hilbert transform of each signal obtained from the filter bank was then calculated. The final envelope was obtained by averaging the 8 analytic signals together [Yang et al., 1992].

3.3.6 EEG acquisition

Continuous EEG signals were recorded from 64 channels using AgAgCl active electrodes arranged in a 10/20 layout (ActiCapSlim, Brain Products GmbH, Gilching, Germany). Signals were digitised at 5000 Hz with an anti-aliasing filter at 1000 Hz and down-sampled to 1000 Hz. Subsequently, the data was high-pass filtered at 0.1 Hz and low-pass filtered at 500 Hz. Electrode impedances were kept below 20 kOhm. Triggers from the stimulus-control computer were sent via LabJack to the EEG acquisition computer.

3.3.7 EEG preprocessing

Preprocessing was performed with the FieldTrip toolbox [Oostenveld et al., 2011] (`http://www.fieldtriptoolbox.org/`). Raw data were high-pass filtered at 0.3 Hz, low-pass filtered at 150 Hz and band-stop filtered around the line noise and its harmonics (49-51 Hz, 99-101 Hz, and 149-151 Hz), and epoched for each trial. The epoch length was from -1 s to 28 s. Trials were subsequently visually inspected and independent component analysis was used to remove artefacts due to eye movement. The EEG recording was segmented into trials and rereferenced using the two mastoids (TP9 and TP10).

3.3.8 EEG analysis

The neural tracking of the acoustic envelope was assessed by computing the accuracy of the reconstruction of the stimulus $s(t)$ from the EEG activity $r(t)$. Envelope features were predicted with the following linear model:

$$\mathrm{S}\widehat{(t)} = \sum_{ch=1}^{64} \sum_{\tau=-150\mathrm{ms}}^{500\mathrm{ms}} \mathrm{r(t+}\tau\mathrm{, ch)g}(\tau\mathrm{, ch})$$

where ch- in range [1-64]- represents the EEG channel, τ is the time lag considered in the model and $g(\tau, ch)$ the modelled impulse response function of the brain.

The condition-specific kernel $g(\tau, ch)$ was estimated using ridge regression with an 8-fold (2 runs per fold) nested cross validation [Crosse et al., 2016a]. A least square estimation was computed as following:

$$\mathbf{g} = (\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I})^{-1}\mathbf{R}^T\mathbf{S}$$

where λ is the Tikhonov regularization of Lagrange multiplicator, **I** is the identity matrix and **R** is the lag matrix. The latter is defined as:

$$\begin{bmatrix}
r_1(\tau_{500}+1) & ... & r_{64}(\tau_{500}+1) & r_1(\tau_{500}) & ... & r_{64}(\tau_{500}) & ... & r_1(1) & ... & r_{64}(1) \\
\vdots & ... & \vdots & \vdots & ... & \vdots & ... & \vdots & ... & \vdots \\
r_1(T) & ... & r_{64}(T) & r_1(T-1) & ... & r_{64}(T-1) & ... & r_1(T-\tau_{500}) & ... & r_{64}(T-\tau_{500}) \\
0 & ... & 0 & r_1(T) & ... & r_{64}(T) & ... & r_1(T-\tau_{500}+1) & ... & r_{64}(T-\tau_{500}+1) \\
0 & ... & 0 & 0 & ... & 0 & ... & r_1(T-\tau_{500}+2) & ... & r_{64}(\tau_{500}+2) \\
\vdots & ... & \vdots & \vdots & ... & \vdots & ... & 0 & ... & \vdots \\
0 & ... & 0 & ... & ... & 0 & ... & 0 & ... & 0
\end{bmatrix}$$

where T is the length of the EEG signal, N=64 the number of channels and τ_{500} is the max number of lags considered in the estimation.

The multisensory integration (MSI) was evaluated by comparing the results of the multisensory decoder with those of the additive one [Besle et al., 2004, Crosse et al., 2015, Crosse et al., 2016b].

$$\text{MSI} = \text{corr}[\text{S(t)},\hat{\text{S}}_{\text{AT}}(\text{t})] - \text{corr}[\text{S(t)},\hat{\text{S}}_{\text{A+T}}(\text{t})]$$

$\hat{S}_{AT}(t)$ is the predicted stimulus for the AT condition and $\hat{S}_{A+T}(t)$ is the estimation of the additive unisensory model [Crosse et al., 2015, Crosse et al., 2016b].

3.3.9 fMRI acquisition

A 3T Siemens Prisma MR scanner (Simens Medical) was used to acquire both T1 structural volume images (TR = $2000ms$; TE = $2,03ms$; flip angle = $8°$; FOV = $256mm \times 256mm$; 208 axial slices; $1 \times 1 \times 1mmm^3$) and T2*-weighted axial echo-planar images with blood oxygenation level-dependent (BOLD) contrast (gradient echo multi band with factor 2; TR = $1550ms$; TE = $35ms$; flip angle = $71°$; FOV = 256×256; 60 axial slices;

spatial resolution $2.5 \times 2.5 \times 2.5mm^3$, no interslice gap). In total, 16 sessions were recorded per participant, with 400 volume images per session. The first four volumes of each run were discarded from the analysis to allow for T1 equilibration effects. The Matlab functions use to compute the temporal responses were taken from the mTRF toolbox `https://github.com/mickcrosse/mTRF-Toolbox.git`.

3.3.10 fMRI analysis

The fMRI data were analysed with SPM12 (`https://www.fil.ion.ucl.ac.uk/spm/`) [Friston et al., 2007]. Scans from each participant were time- corrected via slice timing, realigned and unwarped to correct for head motion, spatially normalized into MNI standard space using parameters from segmentation of the T1 structural image [Ashburner and Friston, 2005], resampled at $2 \times 2 \times 2mm^3$ and spatially smoothed with a Gaussian kernel of 8 mm FWHM. The time series of each voxel were high-pass filtered at 1/128 Hz.

General Linear Model (GLM)

The fMRI experiment was modelled in a blocked fashion with each regressor entered in the design matrix after being convolved with the canonical hemodynamic response function. On top of modelling the 6 conditions of our experiment (A, T, AT congruent, AT incongruent, Acp and ATcp), the statistical model included parametric modulators (PM) that modelled the envelope of each block. Pitch information was also included in the GLM, but it did not show any significant multisensory interactions at the second level analysis. The cocktail party conditions (Acp and ATcp) included two PMs for each melody that was present in the composition, for a total of four parametric modulators. Realignment parameters were included as nuisance covariates to account for any residual motion effect. Condition specific effects were estimated according to the GLM and passed to a second level analysis as contrasts. This provided the generation of 13 con-

trast images: 8 contrasts for the A, T, AT congruent and AT incongruent conditions (4 conditions x block and 4 conditions x envelope); 2 contrasts for the Acp condition (the block regressor plus PM); 3 contrasts for the ATcp condition (the block regressor plus the 2 PMs). All the summary statistics were summed over the 16 sessions for each subject and entered in a second-level ANOVA. The second-level ANOVA modelled 6 block conditions (A, T, ATc, ATi, Acp, ATcp) and 7 envelope conditions (A, T, ATc, ATi, Acp, 2xATcp).

Contrasts

To asses cross-modal interactions, the superadditivity criterion was used to test differences in BOLD response profiles as: $(AT-fixation) \neq (A-fixation)+(T-fixation)$. More specifically, the criterion defines that a pure multisensory interaction needs to reflect a non-linear combination of unisensory responses [Laurienti et al., 2005, Noppeney, 2012].

Importantly, since a lot of attention was put into understanding the relevance of temporal coherence for the formation of multisensory objects, the encoding of the envelope of each melody was also tested by comparing the differences in neural activation between congruent and incongruent trials.

All the analyses were computed using SPM toolbox, including the second level statistics. Activations at $p < 0.05$ family-wise-error cluster-corrected were reported with auxiliary uncorrected peak-level threshold of p < 0.001.

3.3.11 Screening

To assess whether participants were able to differentiate between congruent and incongruent audio-tactile stimuli, a screening test was performed before the neuroimaging acquisition. In a yes-no congruency judgment task, participants were presented with different audio-tactile melodies that were matched (or not) across sensory modalities.

Audio and tactile signals were presented similarly as described in Section 4.3.6. Participants rested their head on a chin rest and responded by pressing a pedal positioned under their feet (SODIAL, Shenzhen IMC Digital Technology Co.). Background white noise was additionally played (65 dB sound pressure level) to mask eventual sound from the tactile vibrations. Participants completed 2 separate runs, each of which consisted of 15 trials of congruent and incongruent conditions (2 x 15 x (cong,incong) = 60 trials total). 'Yes' and 'No' responses were respectively categorized as hit and miss in the congruent trials, whereas considered false alarms and correct rejections in the incongruent trials [Wickens, 2001]. D': $[Z(\text{hit rate}) - Z(\text{false alarm})]$, criterion: $[(-\frac{Z(\text{hit rate})+Z(\text{false alarm})}{2})]$ and the proportion of correct responses were computed. Participants that showed $D' > 2.8$ were included in the experiment (12 subjects, D' group mean and SEM = 5.409 ± 0.533; criterion = 0.093 ± 0.206; proportion of correct responses = 0.966 ± 0.011). All of the 12 subjects were included in the experiment.

3.4 Results

3.4.1 EEG

Audio tactile congruency effects were first investigated for monophonic musical pieces. As hypothesized, we found higher correlation between the envelope of the stimulus and its estimate for audio-tactile congruent signals relative to incongruent ones (Wilcoxon signed rank test; ATc Pearson correlation (ρ) 0.244 ± 0.05 ; A ρ 0.206 ± 0.07; ATi ρ 0.227 ± 0.06; T 0.08 ± 0.02; $[ATc > A]$ p = 0.002, z-score = 3.059, Cohen's d = 1.136; $[ATc > ATi]$ p = 0.015, z-score = 2.432, Cohen's d = 0.991). Importantly, the encoding of the music envelope was also greater for audio-tactile congruent trials compared to the additive model (A+T ρ 0.234 ± 0.05; $[ATc > A + T]$ p = 0.003, z-score = 2.981,

Cohen's d = 1.024). The reconstruction of incongruent trials did not outperform the additive model, which indicates the relevance of temporal coherence in obtaining multisensory enhancement as already shown in the case of audio-visual speech [Crosse et al., 2015, Crosse et al., 2016a] (Fig. 3.3A).

Subsequently, the neural tracking of the melodies was assessed in the cocktail-party scenario. Again, the correlation coefficient of the audio-tactile condition was greater than that of the unisensory one and the additive model (A cp ρ 0.187 ± 0.07; AT cp ρ 0.222 ± 0.05; A cp + T ρ 0.216 ± 0.06; $[ATcp > Acp]$ p = 0.015, z-score = 2.432, Cohen's d = 0.991; $[ATcp > Acp + T]$ p = 0.022, z-score = 2.275, Cohen's d = 0.76). Crucially, the multisensory melody was also decoded with a greater accuracy with respect to the concurrent auditory one ($[ATcp\, boosting]$ p = 0.009, z-score = 2.589, Cohen's d = 0.991) suggesting that the tactile stimulus enhanced the perception of the temporally congruent melody (Fig. 3.3B).

Finally, to assess the time windows over which integration occurred [O'Sullivan et al., 2015, Crosse et al., 2016b], we trained three more models at different time-lags from the stimulus onset: early (0-150ms), middle (150-300ms) and late (300-450ms) responses ((Fig. 3.3C)). We tested again MSI (Fig. 3.3D) as the difference between the reconstruction of the congruent AT trial and the additive model. Early and middle windows of integration showed that the neural tracking of ATc trials was greater than the A+T ones ($p < 0.05$, Holm-Bonferroni corrected) whereas later lags did not show any MSI effect (p = 0.33) suggesting that, coherently with previous literature, integration of audio and tactile signals happened at early time windows [Crosse et al., 2016b, Riecke et al., 2019].

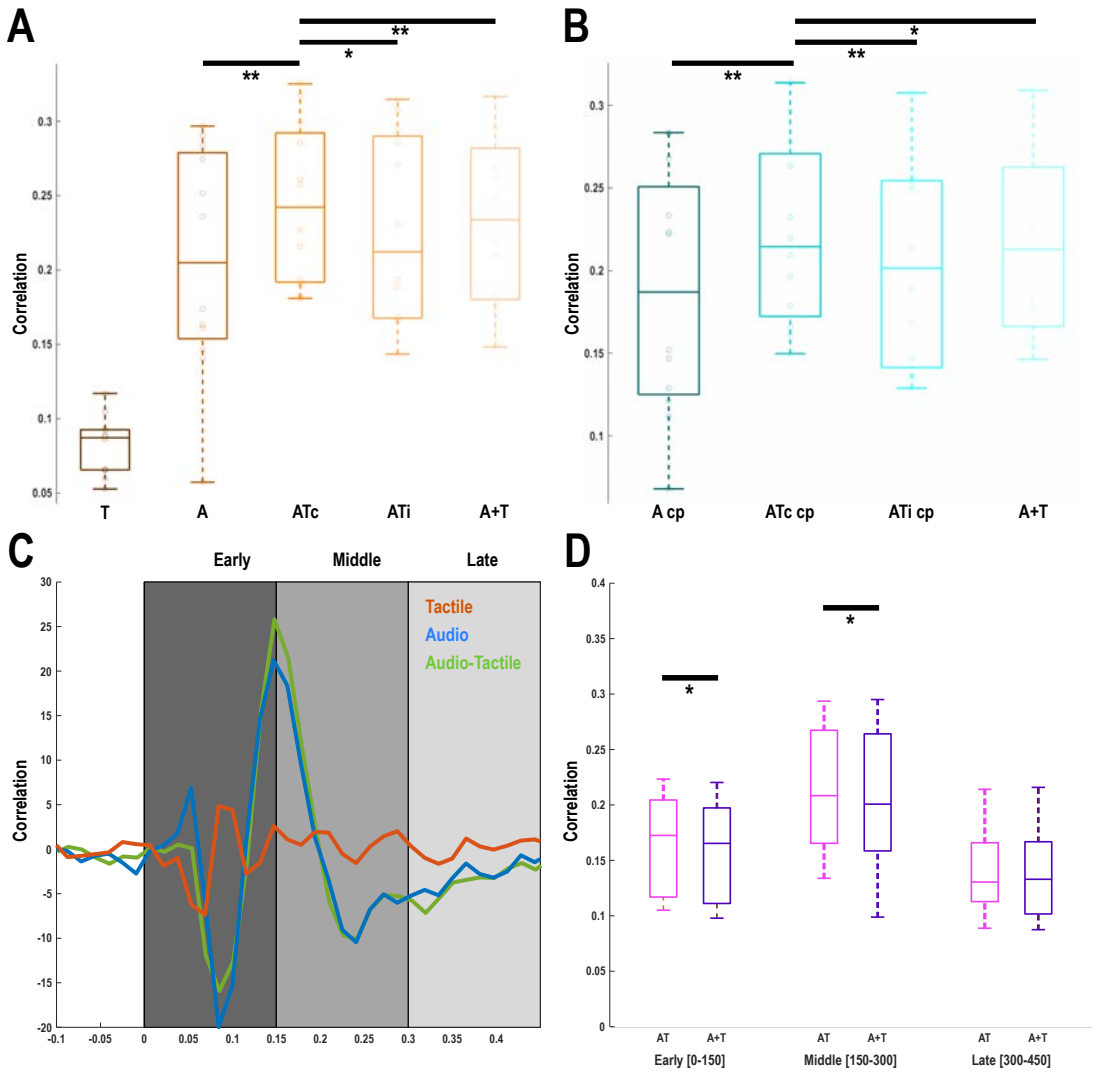

Figure 3.3: Multisensory integration (MSI) in decoding envelope of music. (A) Decoding accuracy obtained for each of the monophonic conditions. (B) Similarly to the monophonic condition, the correlation between the reconstructed envelope and the original one was computed for the cocktail-party trials. (C) Example of kernel learnt during the training phase for each EEG channel (Fz in figure). Time/Lag-plots represent the correlation as a function of lag with respect to the onsets of the stimuli for audio, tactile and AT conditions. The windows of integration reflect early, middle and late brain responses. (D) MSI effect in the separate time windows reflected by the significant difference in reconstruction between ATc and the A+T model. Bar represents statistical differences in decoding accuracy (*: p<0.05 **: p<0.01). T=Tactile; A=Audio; ATc=Audio tactile congruent; ATi= Audio tactile incongruent; A cp=Audio cocktail-party; ATc cp=Audio tactile cocktail-party congruent stream; ATi cp=Audio tactile cocktail-party incongruent stream

3.4.2 FMRI bold - Block contrasts

Superadditivity Multisensory integration was tested in accordance to the superadditivity criterion $AT > A + T$ across monophonic and cocktail-party conditions (AND conjunction). The results show a superadditive effect in the temporal areas including early auditory areas, planum temporale and STG (see Table 3.4.2 and Fig. 3.4 for a summary). Superadditivity was mainly driven by a large temporal deactivation during the perception of unisensory tactile stimuli. This effect is consistent with previous research showing that attention to non-acoustic modalities deactivates temporal regions, while multimodal signals eliminate this effect [Laurienti et al., 2002, Beauchamp et al., 2004, Petkov et al., 2004]. Interestingly, superadditivity was found also in insular and opercular regions often associated with processing of sounds as well as part of multisensory networks [Bamiou et al., 2003, Herdener et al., 2009, Remedios et al., 2009, Zhang et al., 2019, Sepulcre, 2014, Ro et al., 2013, Pérez-Bellido et al., 2018].

Enhancement As discussed in the previous section the superadditivity effect was influenced by the deactivation of tempo-parietal regions during tactile stimulation. To better understand which of these areas conveyed boosted multisensory responses to music perception, absolute multisensory enhancement ($\mathbf{AT} > \mathbf{max}(\mathbf{A}, \mathbf{T})$) was also tested for both monophonic and cocktail-party conditions. While no effect was found for monophonic multisensory enhancement, the AT cocktail-party condition showed greater activation in the bilateral parietal operculum (pOP). Interestingly, the second somatosensory cortex (SII) has not only been associated with tactile and vibrotactile perception [Burton et al., 1993, Burton et al., 2008b, Francis et al., 2000, Reed et al., 2004, Eickhoff et al., 2006], but is also a highly relevant brain area for tactile attention [Nelson et al., 2004, Hämäläinen et al., 2002, Fujiwara et al., 2002, Burton et al., 2008a, Johansen-Berg et al., 2000].

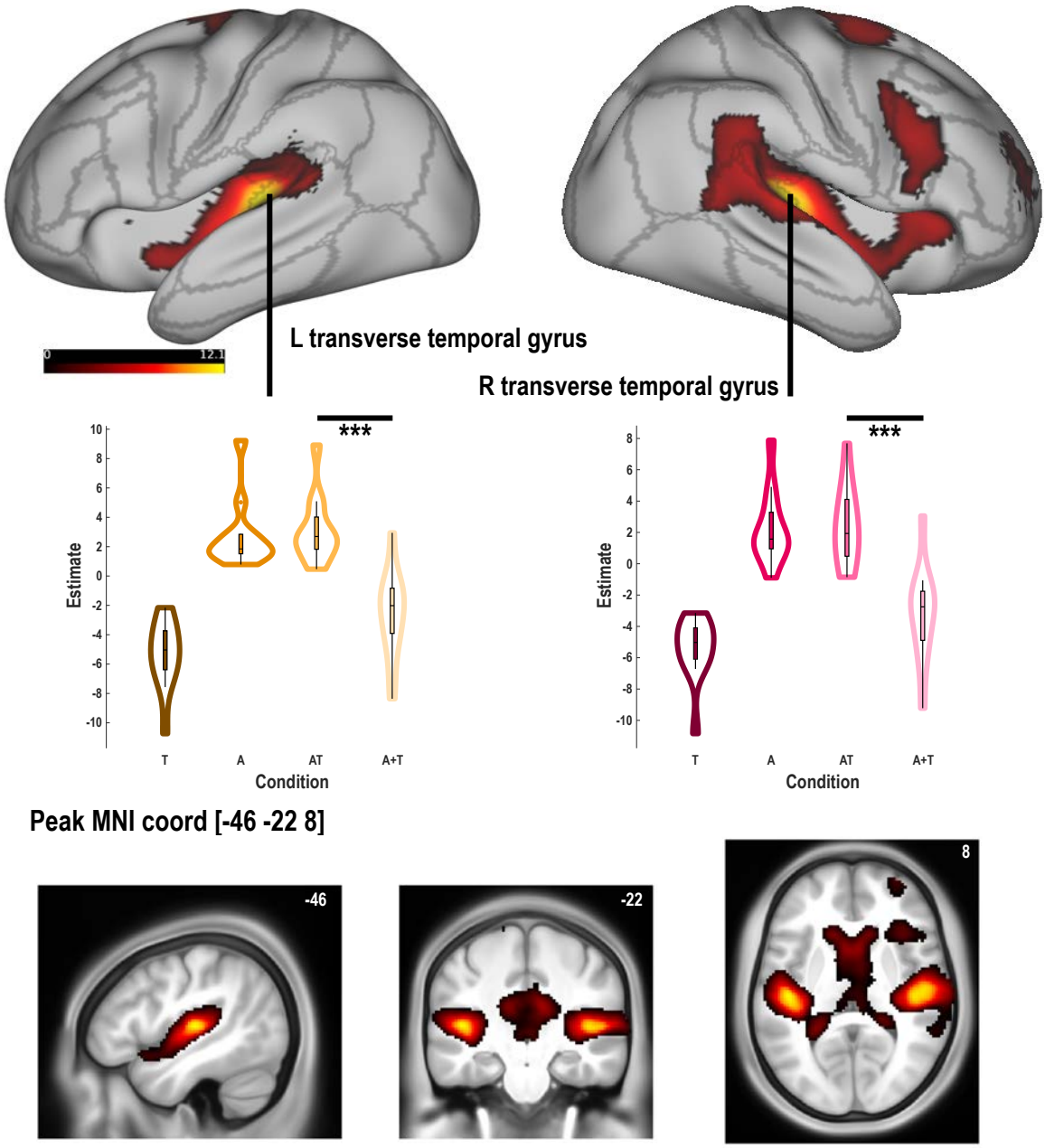

Figure 3.4: Multisensory integration (MSI) shown as superadditivity across bilateral auditory areas. Reported activation of MSI in the conjunction between monophonic and cocktail-party conditions on the inflated brain [Glasser et al., 2013] ($p<0.05$ FWE-cluster level corrected; light grey regions represent Brodmann areas [Brodmann, 1909]). The violin plots represent the mean value of the cluster across subjects for each condition: unisensory tactile (T), unisensory auditory (A) and multisensory audio-tactile (AT)

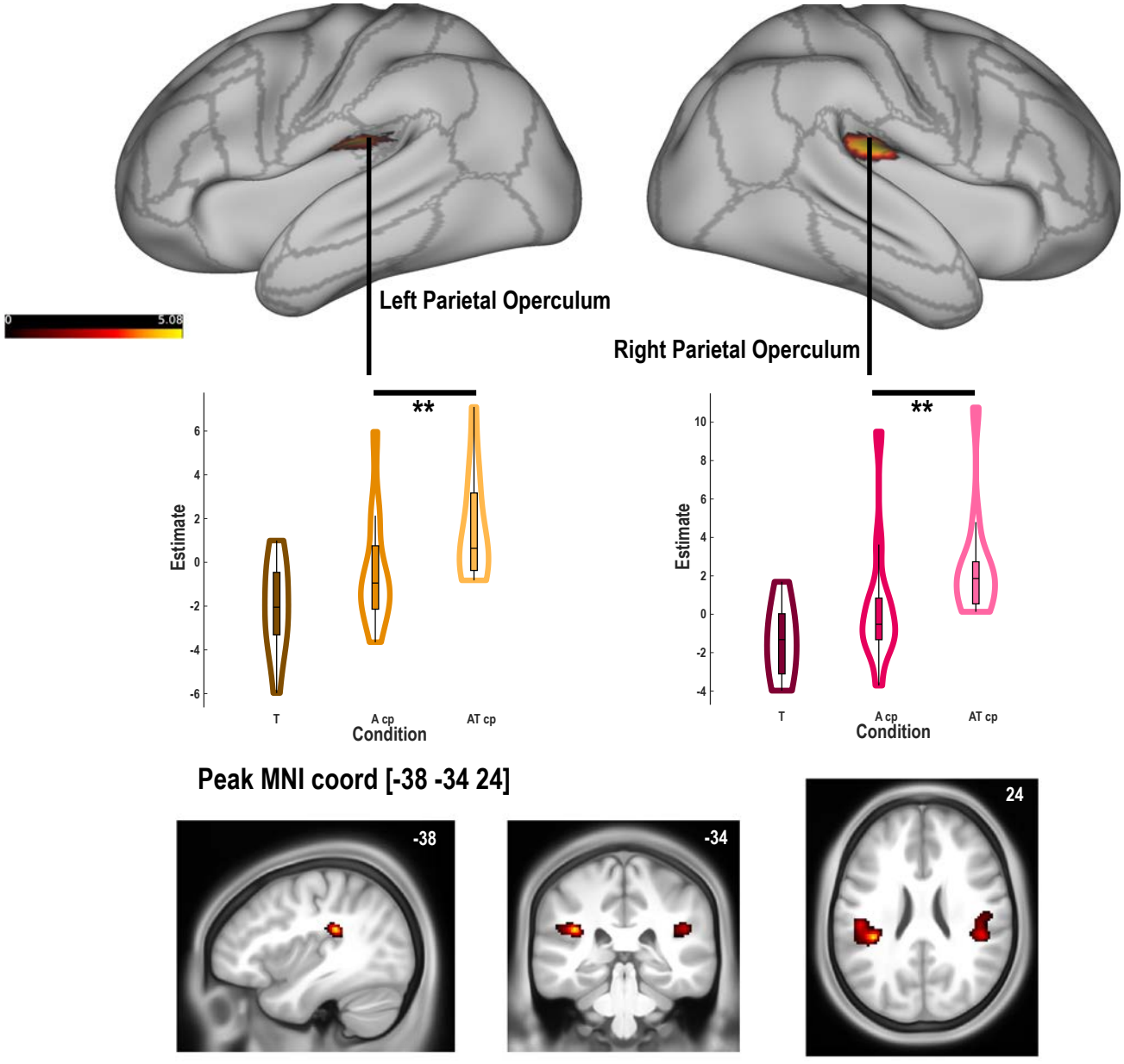

Figure 3.5: Multisensory enhancement during the cocktail party shown as **ATcp** $>$ **max**(**Acp**, **T**). Significantly different activation are reported on the inflated brain ($p<0.05$ FWE-cluster level corrected). The violin plots represent mean value of the cluster across subjects for each condition:unisensory tactile (T), unisensory auditory cocktail party (Acp) and multisensory audio tactile cocktail party (ATcp)

.

Contrast: ATc $>$ A $+$ T $\cap$ ATcp $>$ Acp $+$ T Brain region	p_{FDR} (Cluster)	Cluster size	Z score	x {mm}	y {mm}	z {mm}
	.000	17196				
Left transverse temporal gyrus			$>$8	-46	-22	8
Right transverse temporal gyrus			$>$8	52	-16	6
Right lateral ventricle			6.58	12	-20	24
Left lateral ventricle			6.41	-16	-30	16
Right Thalamus			5.67	2	-34	-8
Planum temporale			5.46	-58	-38	14
Right supramarginal gyrus			5.36	54	-38	24
	.000	1191				
Anterior cingulate cortex			4.59	6	18	40
Pre supplementary motor area			3.83	4	26	58
	.000	889				
Left Cerebellum			5.29	-4	-74	-24
	.002	486				
Right superior frontal gyrus			4.80	18	-2	70
	.005	435				
Left superior frontal gyrus			4.80	-12	-8	74
	.006	415				
Middle frontal gyrus			4.38	32	60	10

Table 3.1: pValues reported in the table were FWE-corrected at a cluster level for multiple comparisons. Uncorrected auxiliary peak was thresholded at p<0.001. A=Audio; T=Tactile; Acp=Audio cocktail-party; ATc=Audio Tactile congruent; ATcp=Audio Tactile cocktail-party

Contrast: ATcp $>$ **max**$(\mathbf{Acp}, \mathbf{T})$ Brain region	p_{FDR} (Cluster)	Cluster size	Z score	x {mm}	y {mm}	z {mm}
Left Operculum	0.002	518	5.69	-38	-34	24
			5.13	-50	-18	18
			4.70	-48	-24	24
Right Operculum	0.006	407	5.04	48	-20	20
			4.67	48	-32	24

Table 3.2: pValues reported in the table were FWE-corrected at a cluster level for multiple comparisons. Uncorrected auxiliary peak was thresholded at p<0.001. T=Tactile; Acp=Audio cocktail-party; ATcp=Audio Tactile cocktail-party

3.4.3 FMRI bold - Envelope contrasts

To further investigate the role played by temporal congruency in AT integration, the encoding of envelope information in the brain was tested again for super/sub additive effects with a specific focus on the role played by incoherent tactile signals during multisensory stimulation. While no effect of superadditivity was found for ATc conditions, different brain areas -namely parietal operculum, supplementary motor cortex and bilateral post and precentral gyri- showed deactivation during incongruent trials (see Table 3.4.3 and Fig3.6 for a comparison with unisensory modalities). Crucially, as showed in the previous section, these same areas were also involved in the multisensory enhancement during the AT cocktail-party condition. Moreover, subadditivity effects during incongruent trials occurred in motor regions, which have recently been shown to be involved in auditory rhythm perception and execution [Chen et al., 2008, Penhune and Zatorre, 2019, Rimmele et al., 2018]. Interestingly, these effects were mainly found in the right hemisphere, which has been associated with spectro-temporal processing of music [Peretz and Zatorre, 2005, Albouy et al., 2020].

As mentioned in previous sections, to further investigate the role of temporal coherence in the encoding of music envelope during multisensory stimulation, the contrast $ATcong > ATincong$ was evaluated. Brain regions similar to those that were discussed before (e.g. parietal operculum) exhibited different activation patterns when the tactile stimulus was coherent (or not) with the auditory one (see Table 3.4.3). This further confirmed the relevant role played by the parietal somatosensory areas in AT multisensory integration over time.

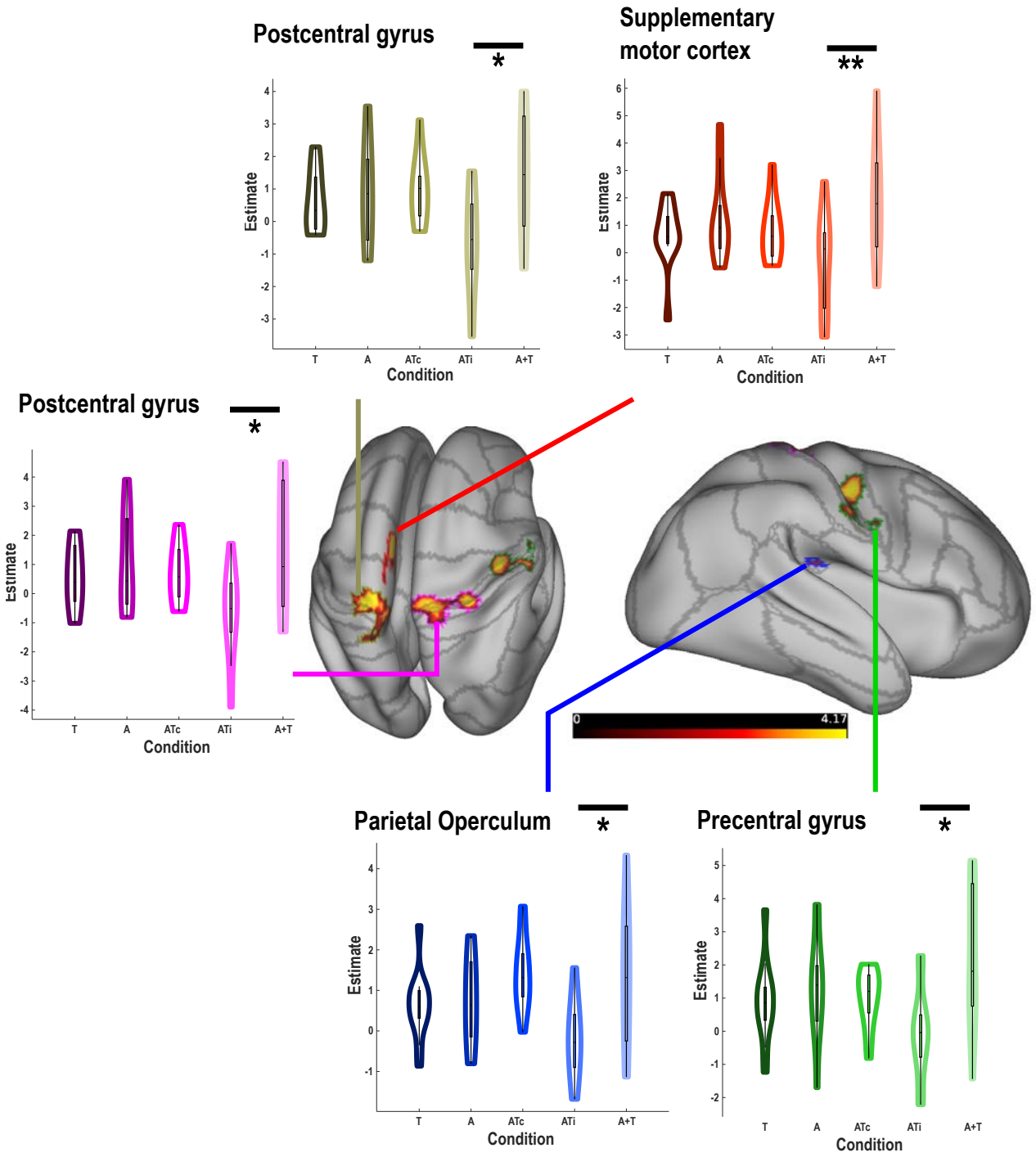

Figure 3.6: Multisensory subadditivity measured for the envelope in the incongruent condition. Activations are reported on the inflated brain ($p<0.05$ FWE-cluster level corrected). The violin plots represent mean value of the cluster across subjects for each condition: unisensory tactile (T), unisensory auditory (A) and multisensory audio tactile congruent (ATc) and incongruent (ATi).

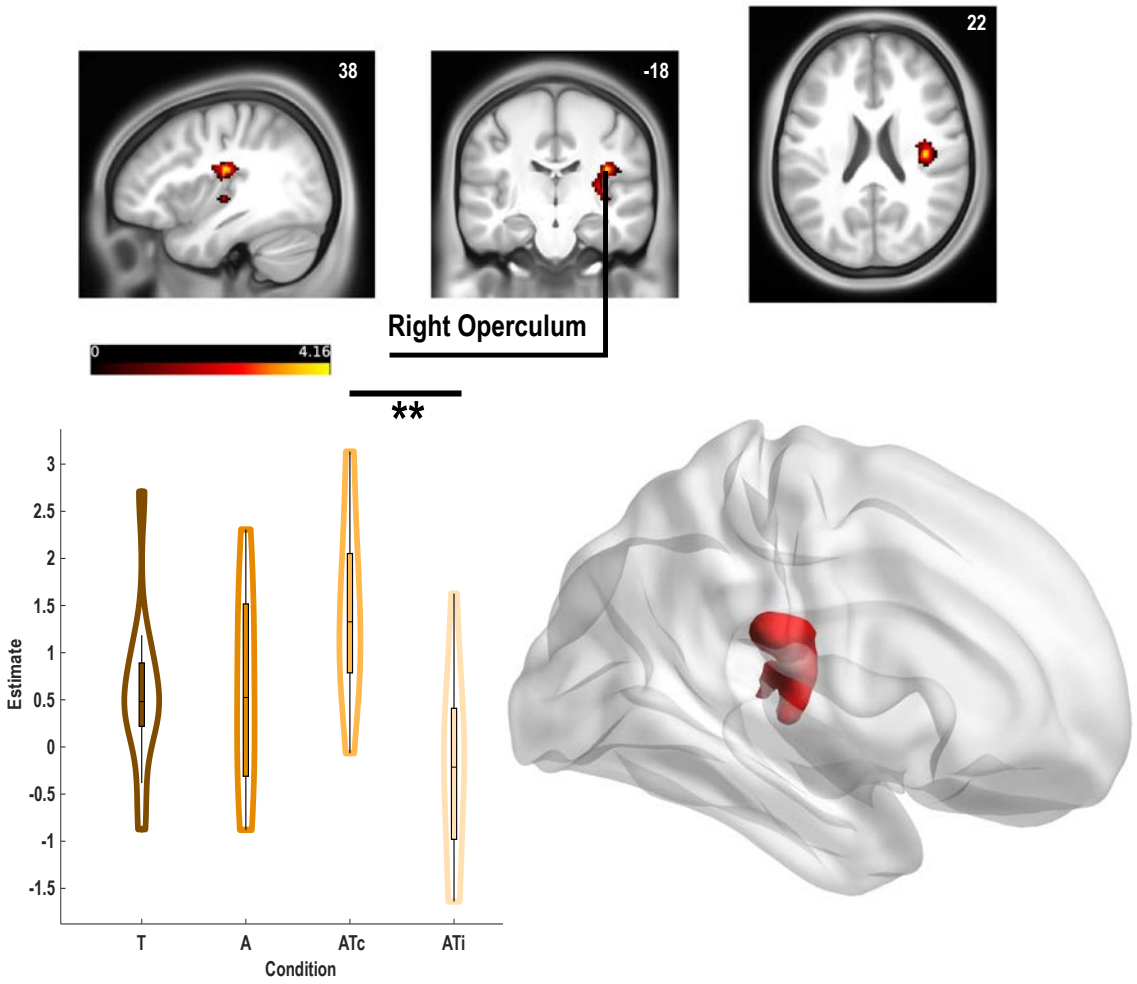

Figure 3.7: Differences between congruent and incongruent trials for the envelope regressor. Activations are reported on the inflated brain ($p<0.05$ FWE-cluster level corrected). The violin plots represent mean value of the cluster across subjects for each condition: unisensory tactile (T), A unisensory auditory (A) and multisensory audio tactile congruent (ATc) and incongruent (ATi)

Contrast : ATi $<$ A $+$ T(Envelope) Brain region	p_{FDR} (Cluster)	Cluster size	Z score	x {mm}	y {mm}	z {mm}
	0.04	260				
Right posterior insula			4.16	32	-16	10
Right operculum			3.81	40	-16	20
Right thalamus			3.81	20	-26	10
	0.03	277				
Right precentral gyrus			4.12	50	-6	48
Right postcentral gyrus			3.43	48	-14	36
	0.008	387				
Left supplementary motor cortex			4.10	-2	6	60
Right supplementary motor cortex			3.43	6	-6	52
	0.022	306				
Right precentral gyrus			4.10	40	-28	66
Right postcentral gyrus			3.87	22	-34	72
Right superior parietal lobule			3.38	18	-48	66
	0.012	355				
Left postcentral gyrus			3.88	-18	-28	70
Left precentral gyrus			3.60	-12	-28	72
Left medial PoG			3.48	-4	-36	66
Right superior parietal lobule			3.41	-14	-44	68

Table 3.3: pValues reported in the table were FWE-corrected at a cluster level for multiple comparisons. Uncorrected auxiliary peak was thresholded at p<0.001. A=Audio; T=Tactile; ATi=Audio Tactile incongruent;

Contrast : ATc $>$ ATi Brain region	p_{FDR} (Cluster)	Cluster size	Z score	x {mm}	y {mm}	z {mm}
	0.0057	425				
Right operculum			4.08	38	-18	22
Right insula			3.97	32	-16	8
Right thalamus			3.61	22	-26	8
			3.60	36	-18	0
			3.57	22	-30	6

Table 3.4: pValues reported in the table were FWE-corrected at a cluster level for multiple comparisons. Uncorrected auxiliary peak was thresholded at p<0.001. ATc=Audio Tactile congruent; ATi=Audio Tactile incongruent;

3.5 Discussion

In this study, EEG and fMRI techniques were used to assess the temporal and spatial dynamics of neural activations involved in audio-tactile binding during music perception. In particular, the decoding of the music envelope from coherent EEG neural tracking [Mesgarani et al., 2009, Crosse et al., 2016a] was used to determine the extent of MSI effect in time as well as stimulus-driven attentional effects present during auditory scene analysis [Crosse et al., 2016b, Zion Golumbic et al., 2013]. In addition, a GLM was computed from fMRI data to address two purposes: (1) to quantify superadditivity in the cortex with a particular focus on early auditory areas [Noppeney, 2012, James and Stevenson, 2012, Bizley et al., 2016]; (2) to compare the variations in envelope encoding between congruent and incongruent AT conditions.

As suggested by the results on the temporal dynamics of MSI, the formation of audio-tactile objects appears to start at an early time scale, which is in agreement with previous studies on audio-speech integration [Luo et al., 2010, Crosse et al., 2015],) In fact, audio-tactile congruent trials showed greater decoding accuracy when compared to the additive model at early and middle latencies , while no MSI effect was observed at later time windows. These findings indicate that, like lipreading, redundant tactile information can enhance the perception of an auditory stimulus at the early stages of stimulus representation. This is furthermore confirmed by the differences in neural tracking of the envelope found between congruent and incongruent tactile streams. Additionally, temporal coherence between neural activations and multisensory stimuli has a relevant role for the segregation of auditory objects [Shamma et al., 2011, Ding and Simon, 2012], which is in line with the evidence obtained from the decoding of cocktail-party conditions, where concurrent vibratory events boosted neural tracking of congruent melodies

over unisensory ones.

As expected, fMRI results showed superadditivity in auditory regions during both monophonic and cocktail-party trials. As previously shown in audio-visual experiments [Beauchamp et al., 2004, Laurienti et al., 2002], while presentation of unisensory tactile stimulation produced inhibition in early auditory cortex, likely due to attention driven away from hearing [Fritz et al., 2007], cross-modal conditions led to sustained activation in sensory areas [Kayser et al., 2005]. Interestingly, an enhancement of the processing AT information during cocktail-party events was found in the opercula (OP). These parietal regions, which are connected to auditory temporal belts [Ro et al., 2013], are also included in the tactile attentional networks that involve both the primary and secondary somatosensory cortex [Reed et al., 2004, Burton et al., 2008b, Johansen-Berg et al., 2000]. Crucially, the OP and insula showed also greater encoding of the music envelope when auditory and tactile information was congruent rather than incongruent. These results might thus suggest not only that SII is recruited for sustaining the binding of audio-tactile objects with the goal of solving complex auditory scenarios, but also that it is a region modulated by the temporal coherence of the multisensory stimuli. It is also worth noting that envelope encoding of incongruent AT trials did suppress the first somatosensory cortex as well as premotor and supplementary motor areas. Therefore, it could be speculated that, since the latter regions are part of a well-known network employed for rhythm and beat perception [Merchant et al., 2015], their inhibition during auditory perception might reflect the interference introduced by the incongruent tactile stimuli. This idea is supported by recent results in AV speech studies, showing how the motor cortex does, in fact, encode cross-modal synergistic information [Park et al., 2018].

One major drawback of the present study, related to the assessment of the functional relevance of AT object formation, is the absence of behavioural results due to

the passivity of the experimental design. Recent evidence suggests that, if perceptual outcomes are not taken into account, classical activation tends to reflect more stimulus-driven neural activity rather than actual behaviourally relevant representations. Keitel at al. illustrated that, when AV speech formations are restricted to actual behaviour outcomes, the overlapping regions shared across modalities are relatively few [Keitel et al., 2020] and do not include areas of the auditory cortex that previously showed encoding of lipreading information [Sams et al., 1991, Calvert et al., 1997, Pekkola et al., 2005]. Hence, the importance for similar future studies to incorporate behavioural measures of multisensory integration to neuroimaging outcomes. Finally, it would be also interesting to perform a better controlled manipulation of stimulus tracking. In this experiment, the low correlation between neuronal and tactile stimuli, suggest that somatosensory processing of melodic information did not, in a broad sense, entrain neural populations as the auditory stimuli. Hence, following the principle of inverse effectiveness [Stein and Meredith, 1993], a stronger evidence of audio-tactile binding might be obtained by maximizing behavioural benefits when stimulus reliability is similar across unimodal signals, boosting object formation and superadditivity effects [Noppeney, 2012, Werner and Noppeney, 2010].

Chapter 4

Audio-tactile integration in music during different levels of awareness

Giulio Degano, David Rollings, Uta Noppeney

Computational Cognitive Neuroimaging lab, Centre of Human Brain Health, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, Birmingham, UK

Citation:

Degano, G., Rolling, D., Noppeney, N. (in preparation). Audio-tactile integration in music during different levels of awareness.

Authors contributions:

Experiment conceptualisation and design: Giulio Degano, Uta Noppeney.

Data collection: Giulio Degano.

Data analysis: Giulio Degano (supervised by Uta Noppeney).

Sleep scoring: David Rollings.

Writing: Giulio Degano (supervised by Uta Noppeney).

4.1 Abstract

Sleep is a vital process during which our brain undergoes recovery and memory consolidation. Indeed, the sensory barriers that isolate the cortex from external stimuli have been suggested to be a necessary mean for the recalibration of neural synapses. Nevertheless, the brain needs to balance between these restorative demands and the need to wake up in the occurrence of important events happening in the surrounding environment. Indeed, recent evidence highlighted the existence of sensory signals processing even in deep stages of unresponsiveness. Yet, the integration of concurrent information incoming from different senses is still unclear. In this study, audio-tactile stimulation was used to assess the extent of multisensory binding across different levels of awareness. Using EEG recordings, the current experiment tried to address two questions: (1) is a multisensory signal able to elicit an enhancement of cortical activity during sleep? (2) can a tactile signal drive bottom-up attentional processes even during unresponsiveness? Multivariate decoding of acoustic features was used to quantify the temporal coherence between sensory information and neural populations within a cocktail party scenario.

The results show that the cortical signature of multisensory binding that was evoked during wakefulness persisted during the first stages of sleep, but gradually decreased with the depth of unawareness. Importantly, it is the increasing unbalance between the processing of acoustic and tactile information during sleep that seems to limit the integration of cross-modal sensory information.

4.2 Introduction

Sleep is essential for human well-being as it allows our body to rest, recover and does even promote cognitive functions such as memory consolidation [Rasch and Born, 2013]. Particularly relevant and deeply researched are the stages of non rapid eye-movement (NREM), i.e. profoundly unresponsive states that have been associated with synaptic down-scaling, learning, energy saving, hormone regulations [Léger et al., 2018] or even cognitive dysfunctions due to sleep loss [Chee and Chuah, 2008]. Neurally, NREM windows are hallmarked by slow oscillatory rhythms called slow waves (SW), which are so stable that they are clearly visible even by naked eye during EEG recordings. Although their functional scopes are multiples, in an opinion paper from 1994 McCormic and Bal framed their importance in what has been know as the ”Thalamic Gating Hypothesis” [Mccormick, 1994]. In this landmark perspective it is suggested that the thalamus, in its active role of modulator of SW oscillations, can be seen as a biological gatekeeper of information to the cortex, creating a strong interference in the perception of sensory messages. It can be then deduced that this subcortical region impedes the processing of audio, tactile or visual signals during sleep in order to avoid unnecessary waste of energy. However, this represents just half of the truth. Although extremely important, the Thalamic Gating Hypothesis fails to address what happens in those time windows during which sleep oscillations do not occur and why unresponsiveness is not only related to SWs but can also happen before their appearance on EEG recordings [Ogilvie, 2001]. Importantly, recent research integrating different imaging techniques has unequivocally shown that sensory processing occurs during sleep and that it does not only involve low-level sensory features [Atienza et al., 2001, Portas et al., 2000, Bastuji and García-Larrea, 1999, Sanders et al., 2013] but also more complex cognitive tasks, such as detection of saliency information (e.g. name of the subjects) [Oswald et al., 1960], odd

sounds [Ruby et al., 2008] and semantic violations [Ibáñez et al., 2006, Bastuji et al., 2002].

That said, the brain struggles to maintain information in time during deep sleep. Classic markers of predictive coding are disrupted in sleep when compared to wakefulness [Strauss et al., 2015] and more than one study has failed to find windows of sensory information processing during NREM that would surpass few seconds [Ruby et al., 2008, Sharon et al., 2017]. However, it is worth noting that these results are influenced by the fact that the sensory inputs during NREM stages cannot be easily attended to since the involvement of high-level prefrontal regions, which activate during task-related events, is limited -if not absent- while subjects are unresponsive [Muzur et al., 2002]. Therefore, if tasks that require higher cognitive abilities are automatized before sleep, it might be possible for the brain to bypass these prefrontal regions and perform more complex tasks, even if unresponsive [Kouider and Dehaene, 2007, Andrillon and Kouider, 2020]. Along this line, in a recent study, researchers investigated the ability to solve a cocktail-party problem during different stages of human sleep [Legendre et al., 2019] . The experiment was run in two different phases: firstly, participants practiced listening to naturalistic stories presented dichotically with Jabberwocky speech; secondly, during sleep, the ability of participants to track the original stories was analyzed by testing which of the two streams was better decoded. Interestingly, subjects demonstrated an ability to segregate meaningful signals from irrelevant ones even during unresponsive states such as NREM2.

The presence of semantic differences in the stimuli is not the only tool that humans use to segregate relevant information in challenging conditions. In fact, a large corpus of studies demonstrated that when sources of information come from different sensory modalities, bottom-up attentional mechanisms are activated, allowing the brain to attend to a specific stimulus more efficiently [Sumby and Pollack, 1954, Ross et al.,

2007, Callan et al., 2003]. In particular, audio-visual integration during speech comprehension has been shown to increase neural tracking of speech features via the introduction of redundant information that reduces ambiguity in auditory scene analysis [Crosse et al., 2016a, Luo et al., 2010, Zion Golumbic et al., 2013, Atilgan et al., 2018]. Moreover, because of the interplay between feedforward and feedback processes involved in sensory integration [Foxe and Schroeder, 2005], recent evidence has shown that the building of multisensory objects can influence selective attention even across different features of the stimuli, creating a strong cross-modal binding effect [Maddox et al., 2015, Bizley et al., 2016, Atilgan et al., 2018].

The current study was designed to assess the ability of subjects to track relevant information in a complex auditory scenario, addressing the following questions: 1) is a multisensory object able to boost the neural tracking of the congruent stream of information while subjects are in an unresponsive state?; 2) Does the multisensory stream decoding differ between light and deep sleep?

Specifically, we aimed at evoking cross-modal binding between audio-tactile stimuli in participants who listened to polyphonic pieces of music created by a composer. The reason behind the selection of audio-tactile stimuli was two-fold: first, it provides redundant information that does not deteriorate with closed eyes; second, it creates a temporally coherent multisensory signal that boost the binding between the two sensory modalities [Bizley et al., 2016, Atilgan et al., 2018, Degano et al., Prep]. Moreover, the choice of stimulating the somatosensory system poses its roots on a strong background of studies suggesting the occurrence of feedforward integration and tactile modulations in early auditory areas [Schroeder et al., 2001, Schroeder and Foxe, 2005, Kayser et al., 2005, Lakatos et al., 2007, Soto-Faraco and Deco, 2009] as well as the existence of direct connections between somatosensory and auditory cortex [Cappe and Barone, 2005, Hackett et al., 2007]. Importantly, recent studies also showed that audio-tactile

interactions have functional validity in naturalistic scenarios like music [Brochard et al., 2008, Ammirante et al., 2016, Tranchant et al., 2017, Degano et al., Prep] and speech [Riecke et al., 2019].

4.3 Materials and Methods

4.3.1 Participants

After giving written consent, 12 volunteers from a previous study (Chapter 3) participated in the sleep session (9 females, age mean = 27.75 SD = 4.08). None of the subjects reported any history of neurological or psychiatric conditions. All volunteers were right handed based on the Edinburgh Handedness Inventory [Oldfield, 1971] (mean Laterality Quotient ($L.Q.$) = 85, with $L.Q. \in [-100, 100]$). Subjects were deprived of stimulants before and on the day of the sleep session to increase the probability of falling asleep when exposed to an auditory stimulus. The recording session was scheduled after 10pm to maximise the amount of NREM sleep. None of the participants was a trained musician (Index of Music Instrument Playing (IMIP) resulted $< .4$ [Chin and Rickard, 2012]) and satisfied the congruency exclusion criteria of the experiment (Chapter 3).
Participants slept on the bed mounted inside the laboratory between 10.00pm until the moment when they were woken up (6.00 am). One participant could not fall asleep after 2.30 am and asked to stop the experiment. Subjects rested on average 22 minutes in N1 (±16 min), 46 minutes in N2 (±22 min), 160 minutes in N3 (±70 min).

The volunteers were reimbursed for taking part in the experiment based on the amount of hours spent in the laboratory (7£ per hour). The study was approved by the University of Birmingham Ethics Committee.

4.3.2 Stimulation

Twenty-four different counterpoint music compositions were presented diotically from the previous study (Chapter 3). Audio and tactile signals were synthesized at 44100 Hz from a MIDI score using Linux MultiMedia Studio 1.1.3 (LMMS 2004, `github.com/LMMS/lmms`) with a Yamaha YDP piano sound font for the auditory signal and a Triple Oscillator for the tactile one.

The envelope extraction, normalization and permutation were computed with a combination of custom MATLAB code (MathWorks, 2019a) and Audacity (The Audacity Team, 2019).

4.3.3 Experimental design and procedure

Participants listened to and perceived 3 different stimuli: 1) a monophonic Tactile (T); 2) an Auditory cocktail party (Acp), with two concurrent monophonic pieces; 3) an Audio-Tactile cocktail party (ATcp), with a tactile monophonic piece that matched one of the two monophonic tracks presented in the auditory modality (Fig.3.2).

The stimuli had a duration of 28 s and were presented with a fixed inter-trial interval of 2s. In order to avoid habituation or prediction effects, the stimuli order were counterbalanced between runs for each participant.

During wakefulness, each condition was presented 24 times, across 6 different runs. To control for vigilance, participants were asked to respond to the appearance of full-screen flashes via pedal-press within $[100ms - 2000ms]$ after flash onset (group average accuracy score: 0.893 ± 0.002, see Chapter 3 for details). During the sleep session, the stimuli were presented continuously overnight until the participant woke up. The sleep session followed the wakefulness testing phase (Fig. 4.2).

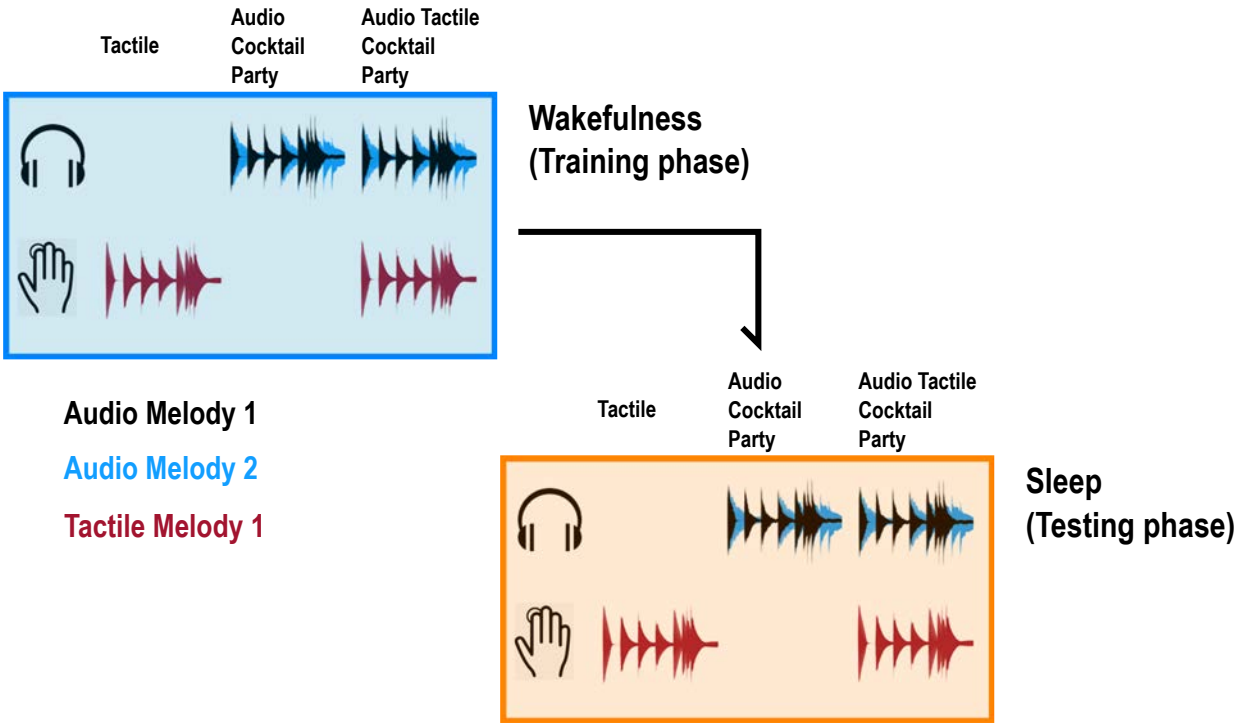

Figure 4.1: Three conditions were presented: one consisted of unisensory tactile trials (T); two conditions consisted of cocktail-party trials. The latter were either unisensory Audio (Acp) or multisensory Audio-Tactile, where the tactile signal matched one of the two auditory streams (ATcp). Participants perceived the stimulation of the trials during both wakefulness and sleep.

4.3.4 Experimental setup

Stimuli were presented using Psychtoolbox version 3.0.15 [Brainard, 1997] (`http://psychtoolbox.org/`) under MATLAB R2018b (MathWorks) on a laptop running Linux Ubuntu.

EEG setup During the training phase (wakefulness), visual flashes were presented via a 30in LCD monitor with a resolution of 2560 x 1600 pixels, at a frame rate of 60 Hz. Auditory stimuli were presented diotically at a sampling rate of 44.1 kHz via EEG compatible earplugs (EARTONE, Insert Earphone 3A) and an ASUS Xonar DSX sound

card. The same sound card was used to drive the tactile vibrations (piezoelectric system (PTS-C2, Dancer Design, UK)). The stimulation was applied to both hands at the fingertip of each index finger. Participants rested their head on a chin rest at a distance of 600 mm from the monitor and at a height that matched participants' ears with the horizontal midline of the monitor. Participants responded by pressing a pedal with their right foot (SODIAL, Shenzhen IMC Digital Technology Co.). During the test phase (sleep), participants were put in a bed and asked to relax and try to fall asleep. Since the goal of the experiment was to asses passive boosting of the congruent AT melody during sleep, no additional task was added.

Background white noise was additionally played through external speakers (65 dB sound pressure level) to mask eventual sounds coming from the tactile vibrations.

4.3.5 Feature extraction

The envelope was computed by first band-pass filtering each music piece with a filter bank of 8 logarithmically-spaced filters between 100-8000 Hz. The Hilbert transform of each signal obtained from the filter bank was then calculated. The final envelope was obtained by averaging the 8 analytic signals together [Yang et al., 1992].

4.3.6 EEG acquisition

Continuous EEG signals were recorded from 64 channels using AgAgCl active electrodes arranged in a 10–20 layout (ActiCapSlim, Brain Products GmbH, Gilching, Germany). Signals were digitised at 5000 Hz with an anti-aliasing filter at 1000 Hz and down-sampled to 1000 Hz. Subsequently, the data was high-pass filtered at 0.1 Hz and low-pass filtered at 500 Hz. Electrode impedances were kept below 20 kOhm. Triggers from the stimulus-control computer were sent via a LabJack to the EEG acquisition

computer.

4.3.7 EEG sleep scoring

Standard sleep scoring was performed using American Academy of Sleep medicine (AASM) criteria in the FASST open-source software `http://www.montefiore.ulg.ac.be/~phillips/FASST.html` [Phillips et al., 2011] and custom code in MATLAB. Data were segmented into chunks of 30 s and referenced to linked-mastoids. Sleep stages were assessed by an experienced neurophysiologist (D.R.). If, after artifact rejection, less than 20 trials were retained per condition in both the NREM1 and NREM2 stage, the participant was fully excluded (subject 11 for NREM1 analysis).

4.3.8 EEG preprocessing

Preprocessing was performed with the FieldTrip toolbox [Oostenveld et al., 2011] (`http://www.fieldtriptoolbox.org/`). Raw data was high-pass filtered at 0.3 Hz, low-pass filtered at 150 Hz and band-stop filtered around the line noise and its harmonics (49-51 Hz, 99-101 Hz, and 149-151 Hz), and epoched for each trial. The epoch length was from -1 s to 28 s. Subsequently, trials were visually inspected and independent component analysis was used to remove artifacts due to eye movement. The first trials of each condition in each sleep stage were then selected up to a maximum amount of 30 to allow for comparison between stages (N1 mean: 47 trial ±30, N2: 93 trial average ±45). The EEG recording was finally rereferenced using the two mastoids (TP9 and TP10).

4.3.9 EEG analysis

Wake Neural tracking of the auditory envelope was assessed by computing the accuracy of the reconstruction of the stimulus $s(t)$ from the EEG activity $r(t)$ during wakeful-

ness. The envelope was predicted with the following linear model:

$$\hat{S}(t) = \sum_{ch=1}^{64} \sum_{\tau=-150ms}^{500ms} r(t+\tau, ch) g(\tau, ch)$$

Where ch represents the EEG channel ranging from 1 to 64, τ is the time lag considered in the model and $g(\tau, ch)$ the modelled impulse response function of the brain.
The condition-specific kernel $g(\tau, ch)$ was estimated using ridge regression with an 8-fold (2 runs per fold) nested cross validation [Crosse et al., 2016a]. A least square estimation was computed as follow:

$$\hat{g} = (D^T D + \lambda I)^{-1} D^T S \tag{4.1}$$

Where S is the 1 x T stimuli matrix, D is the N x T neural data, g are the linear coefficient and λ is the regularization coefficient or Lagrange multiplier. The correlation between $S(t)$ and $\hat{S}(t)$ was used to assess the accuracy with which the features were predicted. The procedure described above was used to predict the envelope of the two melodies presented during T, Acp and ATcp conditions. The best lambda parameter of each melody and condition was selected during the nested cross validation and subsequently used to estimate the decoder across the complete wakefulness session. Importantly, the final training gave four different estimated impulse responses $g(\tau, ch)$: one unisensory tactile, two unisensory auditory (one for each melody) and two multisensory audio-tactile. The Matlab functions use to compute the temporal responses were taken from the mTRF toolbox `https://github.com/mickcrosse/mTRF-Toolbox.git`.

Sleep The four kernels estimated during wakefulness were tested on the trials during NREM1 and NREM2 stages. The correlation between S(t) and $\hat{\mathrm{S}}$(t) estimated for NREM1 and NREM2 was used to assess the *reconstruction accuracy* of the envelope

as:

$$\rho = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(\hat{s}_i - \bar{\hat{s}}_i)}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2}\sqrt{\sum_{i=1}^{N}(\hat{s}_i - \bar{\hat{s}})^2}}$$

As in the previous Chapter 3, the neural tracking of the melodies was compared between multimodal and unisensory conditions.

Attention Decoding: To compare stimulus-driven attentional effects during the AT cocktail-party condition, the trials were first segmented into chunks of 7 seconds and then used to reconstruct the multisensory signal and the unisensory one that were presented concurrently. Each trial was considered to be decoded if the correlation between the estimated envelope and the original one was greater for the AT melody than for the auditory one ($\rho_{\text{Audio-tactile}} > \rho_{\text{Audio}}$) [O'Sullivan et al., 2015, Legendre et al., 2019]. Finally the *decoding-accuracy* was assessed as:

$$\text{Dec-Acc} = \frac{\sum \text{Trial decoded}}{\text{Total number of trial}}$$

The decoding-accuracy results were then combined across subjects and compared to chance level in a second-level statistical analysis [Legendre et al., 2019].

4.4 Results

Melody reconstruction The neural tracking of cross-modal music objects was assessed during different level of awareness using multivariate regression methods. First, linear kernels were learnt and tested during wakefulness, then decoders were used to predict the envelope of the melodies (tactile or auditory) from the EEG data during different conditions and stages of sleep. The decoding accuracy during wakefulness trials was greater for AT conditions compared to unisensory ones. As in Chapter 3, this result suggests an enhancement of melody tracking due to cross-modal binding effects (T ρ 0.08 ± 0.01 ;A cp ρ 0.187 ± 0.07; AT cp ρ 0.222 ± 0.05; A cp + T ρ 0.216 ± 0.06; [ATcp >Acp]

p = 0.015, z-score = 2.432, Cohen's d = 0.991;). Interestingly, this result subsisted during light sleep (NREM1), but faded away during stages of deeper unresponsiveness (NREM2) (*N1*: T $\rho\ 0.008 \pm 0.02$; A cp $\rho\ 0.037 \pm 0.048$; AT cp $\rho\ 0.048 \pm 0.04$; [ATcp >Acp] *p* = 0.041, z-score = 1.733, Cohen's d = 0.589; *N2*: T $\rho\ 0.003 \pm 0.01$; A cp $\rho\ 0.031 \pm 0.03$; AT cp $\rho\ 0.02 \pm 0.03$; [ATcp >Acp] *p* = 0.872, z-score = -1.137, Cohen's d = -0.475).

Attention Decoding The decoding accuracy of the two melodies presented during cocktail-party conditions was tested to determine whether stimulus-induced attentional effects, driven by cross-modal saliency, occurred during sleep.

As previously described, the successfully decoded trials ($\rho_{\text{Audio-tactile}} > \rho_{\text{Audio}}$) were used to estimate the overall performance in a group-level analysis. During wakefulness, cross-modal binding drove the tracking of auditory information to the temporally congruent audio-tactile melodic object (Dec-Avg: 0.54, t-test [ATmelody > Amelody] *p* = 0.032, tval = 2.44). Coherently with this result, the decoding-accuracy obtained during wakefulness persisted during NREM1, but weakened in NREM2, indicating that cross-modal saliency effects were modulated by sleep stages (*N1* Dec-Avg: 0.55, t-test [ATmelody > Amelody] *p* = 0.035, tval = 2.43; *N2* Dec-Avg: 0.467, t-test [ATmelody > Amelody] *p* = 0.064, tval = -2.04).

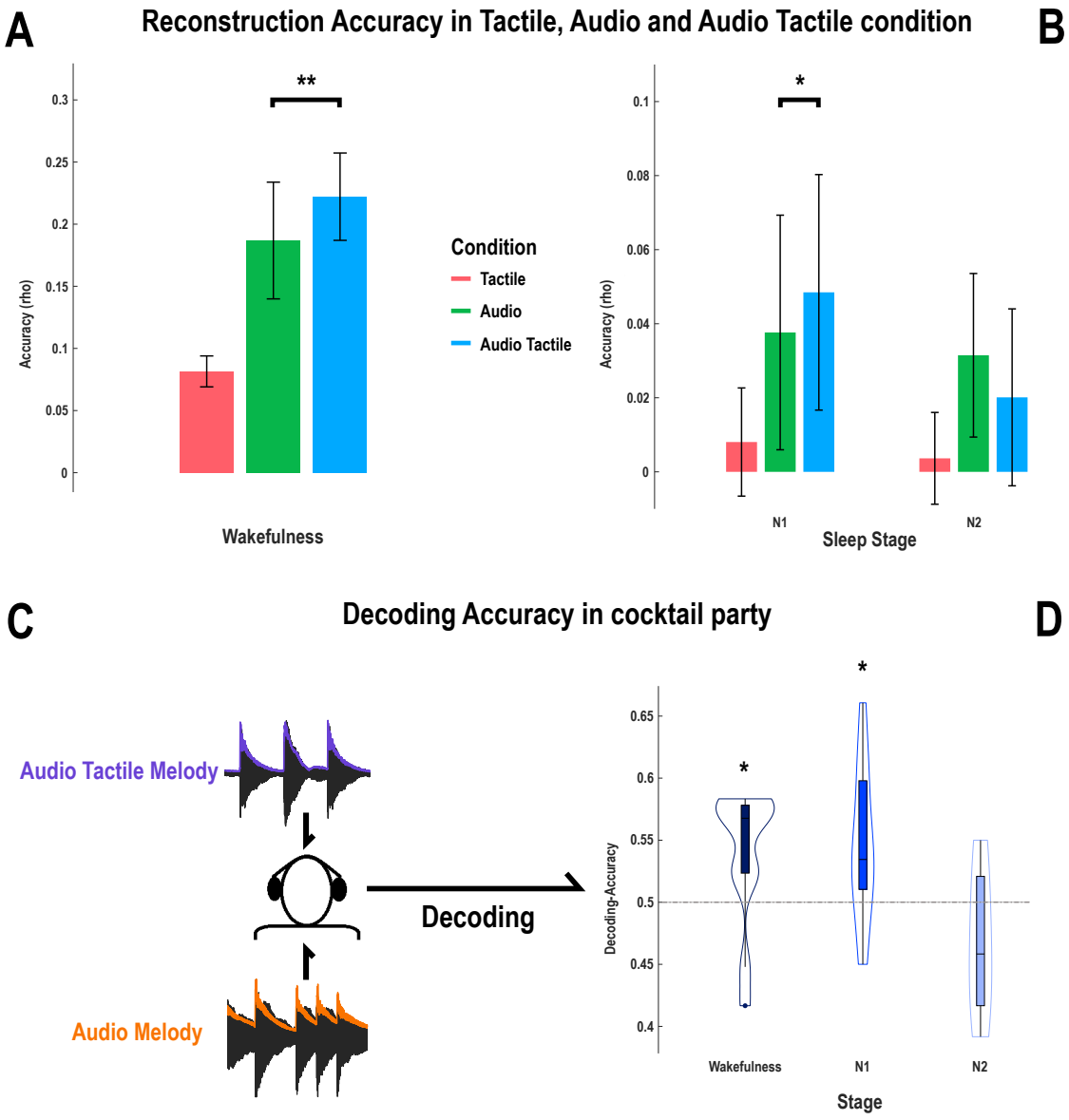

Figure 4.2: *EEG reconstruction and decoding.* (A) Reconstruction accuracy during wakefulness shows enhancement of neural tracking during multisensory conditions. (B) During light sleep the envelope of AT trials was reconstructed with higher accuracy although the effect was not replicated during NREM2. (C) Scheme of the decoding procedure: if the accuracy of multivariate analysis was higher in AT melodies during cocktail party conditions, the trial was labelled as decoded. (D) Decoding results in different stages of awareness (second-level t-test, decoding vs chance). Significance levels are *: $p<0.05$ **: $p<0.01$

4.5 Discussion

In this study, electroencephalography was used to determine how different levels of awareness affect the neural tracking of cross-modal objects. More specifically, polyphonic piano melodies were composed to generate rich auditory scenarios and, through audio-tactile congruency, elicit stimuli-driven attentional outcomes. Multivariate reconstruction approaches were used to assess the neural dynamics during unresponsiveness states as well as their coherence with the auditory and tactile stimuli provided [Mesgarani et al., 2009, Crosse et al., 2016a]. Importantly, the influence of sleep stages on cross-modal saliency was investigated in the context of cocktail-party conditions by estimating the segregation of piano melodies [Shamma et al., 2011, O'Sullivan et al., 2015, Legendre et al., 2019].

The results obtained from these two different but complementary analyses reveal that deeper sleep stages tend to cancel AT effects. While light sleep stages like NREM1 show a consistent (although reduced) multisensory boosting of neural activation, more profound states of unawareness are likely to diminish these cross-modal interactions. It is important to notice that this does not mean that sensory information is not processed: on the contrary it seems that, even during NREM2, auditory information does reach the cortex as previously shown [Portas et al., 2000], but there is a strong unbalance between somatosensory and hearing perceptions. In agreement with previous results on speech perception during sleep, our data suggest that, during light sleep, the brain is able to enhance the neural response to one stream over the other. Indeed, while Legendre et al. reported significant segregation during cocktail party conditions in NREM stages driven by extensive training during wake, the current experiment shows that similar results can be achieved via multisensory saliency. However, this work fails to tack such behavior at deeper stages of sleep, as demonstrated in the case of speech [Legendre

et al., 2019]. In Chapter 3 it has been discussed that when audio-tactile stimulation is employed, the strength of the neural responses strongly depend on the presented modality. In other words, the neural tracking of audio and tactile signals is dissimilar. This suggests that the one used is not the most favourable scenario for obtaining robust multisensory effects (see inverse effectiveness [Stein and Meredith, 1993].). Moreover, although we can confidently state that the perception of the tactile stimulus is weaker than the auditory one, our results indicate that touch is still able to evoke a measurable neural response during wakefulness. However, the latter is presumably damped by thalamic inhibitions during sleep [Mccormick, 1994]. It is therefore possible that the sensation of touch is processed when the brain is aware of the coherent musical message carried by vibrations, but is gradually ignored while falling asleep. This neglection is likely due to habituation effects [Simpson, 1977] or to the fact that touch signals are not treated as meaningful events [Andrillon and Kouider, 2020].

In summary, this naturalistic study suggests that audio-tactile enhancement of cortical tracking seems to subsist during early stages of sleep or drowsiness, although the same multisensory benefits do not persist during longer unresponsiveness states. It is suggested that cross-modal binding likely fades away due to strong unbalances in the processing of auditory and tactile information, which occur gradually with increased level of unawareness.

Chapter 5

Discussions

The goal of this work was to assess the neural correlates of audio-tactile binding during ecologically-valid stimulation. More specifically, the two studies presented in this thesis aimed at addressing different issues regarding AT object formation in the brain. First, the temporal and spatial nature of AT binding was examined during free-behaving music perception. Second, the enhancement of the representation of acoustic streams via congruent AT stimulation was investigated in a naturalistic cocktail-party scenario. Third, the modulation of the AT binding by different levels of awareness was addressed during sleep.

In the following sections a summary of the main findings will be given. Finally, the limitations of the proposed studies will be discussed together with possible future work on AT binding and its potential applications.

5.1 Findings

5.1.1 Early cross modal formation and effects of temporal congruency

In Chapter 3, audio-tactile binding was investigated in the context of naturalistic music perception. Neuroimaging data was used to compare neural activations in multisensory conditions against unisensory ones. Moreover, congruent conditions were compared to the additive model (A+T) to quantify non-linear responses commonly associated to pure multisensory effects [Murray and Wallace, 2012]. To do so, superadditivity was assessed spatially with fMRI [Noppeney, 2012] and temporally with EEG [Crosse et al., 2015].

Since the aim of these two experiments was to extend the definition of audio-visual object formation to AT pairings, the spatial and temporal criterion of multisensory binding

needed to be satisfied at a neural level [Bizley et al., 2016]. Indeed, the EEG analysis showed that audio-tactile interactions happened at early time windows of integration [0-150ms], satisfying the temporal criterion of cross-modal object formation. Secondly, findings of multisensory effects in auditory areas confirmed that low-level regions must be implicated in audio-tactile binding, thus verifying also the spatial hypothesis.

In the present thesis, neural activations in the cortex have been shown to be modulated by temporal congruency between audio and tactile stimuli. The reconstruction of musical melodies from EEG data was in fact greater if the tactile signal was coherent with the auditory one, demonstrating that all the features of a cross-modal object must consistently unwrap over time. These results did in fact replicate findings from previous studies that employed audio-visual speech parings [Zion Golumbic et al., 2013, Crosse et al., 2015, Crosse et al., 2016b], further confirming the relevance of temporal coherence between sensory signals for multisensory binding. Spatially, the right parietal operculum (PO) showed significantly different encoding of envelope information between congruent and incongruent trials. Crucially, recent evidence showed that the PO is highly interconnected with auditory regions [Cappe and Barone, 2005, Ro et al., 2013], represents temporal frequency content of tones [Pérez-Bellido et al., 2018] and is functionally relevant for multisensory integration [Beauchamp and Ro, 2008]. These findings suggest that the operculum might play a role in facilitating AT binding via representation of coherent AT information.

5.1.2 Tactile stimuli drive selective attention

It is well established that, during cocktail-party scenarios, different neural populations selectively respond to each sound source [Maddox and Shinn-Cunningham, 2012, Middlebrooks and Bremen, 2013]. Concurrent information coming from other senses can

bind to an auditory stream, resulting in an enhancement of its neural representation [Bizley et al., 2016]. Emerging evidence in audio-visual formations have shown that coherent stimuli can help the segregation of a cocktail-party scenario even before the actual activation of selective attentional mechanisms [Atilgan et al., 2018]. Thus, the assessment of brain responses during stimuli-competition becomes an essential aspect for the investigation of cross-modal objects [Bizley et al., 2016]. To this end, tactile enhancement of neural activations during auditory scene analysis was also assessed in Chapter 3.

The neural tracking of the acoustic features of music was quantified using decoding of EEG data. This method reflects the amount of neurons that respond coherently to one of the two music streams that were concurrently presented [O'Sullivan et al., 2015]. The results obtained from this analysis confirmed the expectation that tactile information selectively enhanced the cortical responses to the acoustic stimulus to which it was bounded.

The evaluation of somatosensory enhancement of neural activations in fMRI during AT cocktail-party conditions revealed the parietal operculum as a possible network involved in maintaining stream segregation over time. As discussed in the previous section, incongruent trials were also found modulating the representation of envelope information in this same region. Interestingly, these consistent findings furthermore corroborate the role of PO in the representation of temporally coherent AT signals.

As mentioned in the Introduction, the temporal and spatial effects of multisensory integration *and* the neural enhancement due to cross-modal interactions that allowed for a selective segregation of the cocktail party scenario constitute the neural basis of the formation of an audio-tactile object. Indeed, these complementary results highlight

the fact that the brain is able to infer at early time windows (from 0 to 150 ms after the presentation of sounds) the concurrent information shared between the somatosensory and auditory perception. This inference across cross-modal features is subsequently used to compute and maintain over time the segregation of acoustic streams, thus forming a cross-modal bind between these two sensory modalities.

5.1.3 Modulation of awareness

In the second part of the present thesis, AT cross-modal formation was studied across different levels of awareness. The motivation behind this research question was to assess whether the automatic enhancement of neural activations due to cross-modal signals was sustained during stimuli-competition even during sleep. To address this question, stimulus features were decoded from EEG data during cocktail-party scenarios with subject specific linear decoders learnt during wakefulness (Chapter 3): these were tested in both NREM1 and NREM2 stages [Legendre et al., 2019].

The reconstruction accuracy obtained during sleep highlighted that multisensory binding degraded as the sleep stages deepened. In fact, while the neural tracking of musical melodies increased for multisensory conditions presented during light sleep, NREM2 sleep disrupted completely the neural gain carried by cross-modal objects. The decoding of the AT stream during cocktail-party conditions revealed similar results for both the wakefulness and the NREM1 stage, but did not exceeded chance level at deeper stages.

Taken together, these results suggest that brain responses involved in AT object formation during wakefulness do not generalize to profound unresponsive states. A possible reason for the neural tracking degradation of the multisensory object is that tactile information processing was almost absent during sleep. Indeed, it is likely that the vibratory touch presented to participants during sleep did not evoke a response salient

enough to be processed in the cortex, differently to what occurred during wakefulness [Andrillon and Kouider, 2020].

It can be debated that sensory signals during NREM can evoke cortical representations that are different from those observed during awareness states and varies between modalities [Laurino et al., 2014]. Thus, it is possible that the responses to the multisensory stimulus do not generalize well from wakefulness as the saliency weights between modalities have been changed. In other words, training and testing of cross-modal cortical models should be assessed within the same sleep stage to uncover NREM specific multisensory responses. This hypothesis would still be valid even in the perspective of sensory processing during unresponsiveness put forward by Andrillon et al. Indeed, the feedforward characteristics of cross-modal objects that bypass higher-level brain regions might recalibrate during sleep and subsequently favour auditory scene segregation with different cortical weights [Andrillon and Kouider, 2020]. Moreover, the pattern of results observed during the NREM2 stage between the auditory and multisensory condition, comes in reinforcement of such interpretation. The higher reconstruction results of the auditory envelope obtained in the NREM2 stage for the unisensory audio condition rather than the audio-tactile one (see Fig.4.2), suggests that the weights of the brain responses were changing between sensory modalities across wake and deep sleep. This phenomenon would indeed reduce the reconstruction ability of the linear model as the brain responses would no longer reflect the learnt cross-modal interaction, thus performing worse than the pure unisensory condition.

Taken together, the present thesis provided new empirical evidence on the neural correlates of naturalistic cross-modal integration and its modulation by awareness. The aim to extend the definition of multisensory binding from previous studies on audio-visual formations to the audio-tactile interaction was achieved at a neural level in Chapter 3.

Indeed, audio-tactile information has been shown to verify temporal and spatial criteria of cross-modal formation as well as elicit stimulus-driven attentional processes for auditory scene analysis. Moreover, the findings on sleep data revealed that unawareness might interfere with AT binding by employing different inhibitory strengths on audio and tactile signals. In fact, results of Chapter 4 demonstrated that, while the cortical dynamics of multisensory integration were sustained during light sleep, deeper stages of unresponsiveness cancelled out the cortical gain of cross-modal object formation assessed during wakefulness.

5.2 Limitations

The main limitation of the two studies is the unbalance between the neural tracking of acoustic and the tactile envelopes. As reported in the first experiment as well as during sleep, the reconstruction of auditory information was always greater than that of the tactile one, even during cocktail-party scenarios. This is not the most favourable multisensory scenario as more similar levels of cortical tracking between acoustic and tactile features could results in higher cross-modal gains, as asserted by the principle of inverse effectiveness [Stein and Meredith, 1993].

As previously stated in the introduction, the assessment of behavioural outcomes for the study of AT integration was not among the aims of the present thesis. Indeed, the research questions that were addressed in Chapter 3 and 4 purposely avoided to include top-down attentional effects that could have possibly masked neural signatures involved in the automatic AT object formation. The only task used to select participants (see ”Screening” in Chapter 3) had purposely high accuracy and very low behavioural variability (participants responded almost at ceiling; mean accuracy 0.966 ± 0.011) in order to avoid subjects that could not clearly match the tactile stimuli with the audi-

tory one. Moreover, the impossibility to measure task related questions during sleep drove the decision to create an entire study that consistently avoided perceptual measures across wakefulness and different level of awareness. That said, it is important to acknowledge that the present thesis, although verifying the neural criteria for which audio-tactile binding is satisfied, comes short in assessing the functional relevance of the AT object formation. In fact, perceptual benefits of multisensory integration can, for example, improve word comprehension in challenging environments [Sumby and Pollack, 1954]. Moreover, it is important to take into account that behavioural-dependent representations of naturalistic stimuli might be different across modalities. Recent evidence in audio-visual speech suggests that brain areas involved in the representation of auditory and visual stimulus features are spatially distinct from the regions that are correlated with perceptual outcomes (word-comprehension) [Keitel et al., 2020]. These results, while not invalidating the stimulus-driven nature of cross-modal binding investigated in the present studies, uncover a relevant facet of multisensory object representations across the hierarchy of brain processes that justifies the inclusion of behavioural outcomes to future study designs.

5.3 Future work

In Chapter 3, the envelope and pitch features extracted from the musical pieces were used to create a temporally coherent tactile signal. This choice was taken in order to maximise the amount of musical information shared across modalities, thus eliciting greater redundancy for the integration of audio and tactile stimuli and favouring stream segregation during cocktail-party scenarios. However, the formation of cross-modal objects should also allow the enhancement of the representation of acoustic features that are orthogonal to the one shared by two sensory modalities [Bizley et al., 2016, Atilgan

et al., 2018, Maddox et al., 2015]. On this line, a possible next step in the investigation of audio-tactile binding should be the assessment of perceptual benefits on those acoustic dimensions that are not shared between the audio stimulus and the tactile one [Maddox et al., 2015]. More specifically, a natural continuation of this work is represented by the inclusion of a target detection task built to assess multisensory benefits across pitch and envelope features of melodies. Indeed, positive perceptual benefit of congruent somatosensory information on these features would complement and corroborate the current neural findings on the AT object formation. Moreover, the correlation of behavioural outcomes with neuroimaging data would also identify those brain regions that contain perceptually-relevant information for the representation of cross-modal objects. An interesting application of the proposed AT paradigm is for the investigation of multisensory integration in ageing. It is renowned that cognitive functions gradually degrade over the course of our lifespan together with the perception of different sensory modalities, including hearing [Kunelskaya et al., 2005, Peelle and Wingfield, 2016] and touch [Wickremaratchi and Llewelyn, 2006]. Cross-modal information has been shown to improve perceptual abilities of older adults -especially in term of response time [Laurienti et al., 2006, Diederich et al., 2008]- although it is not yet clear if these benefits are due to the inverse effusiveness principle, top-down control architectures or other mechanisms [Mozolic et al., 2012]. Moreover, the current literature is falling short in framing the modulation of cross-modal object formation across the lifespan as only few studies have investigated ageing effects on the integration of temporally coherent multisensory stimuli [Brooks et al., 2018]. The present paradigm -together with recent evidence of AT enhancement in speech perception [Riecke et al., 2019]- can be employed to address the former research questions not only to characterize changes of network dynamics in older adults but also to define the perceptual benefits of multisensory integration.

References

[Albouy et al., 2020] Albouy, P., Benjamin, L., Morillon, B., and Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, 1047(February):1043–1047.

[Alday, 2018] Alday, P. M. (2018). M/EEG analysis of naturalistic stories: a review from speech to language processing. *Language, Cognition and Neuroscience*.

[Aller and Noppeney, 2019] Aller, M. and Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian Causal Inference.

[Alluri et al., 2012] Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, 59(4):3677–3689.

[Ammirante et al., 2016] Ammirante, P., Patel, A. D., and Russo, F. A. (2016). Synchronizing to auditory and tactile metronomes: a test of the auditory-motor enhancement hypothesis. *Psychonomic Bulletin and Review*, 23(6):1882–1890.

[Andrillon and Kouider, 2020] Andrillon, T. and Kouider, S. (2020). The vigilant sleeper: neural mechanisms of sensory (de)coupling during sleep. *Current Opinion in Physiology*, 15:47–59.

[Asaridou and McQueen, 2013] Asaridou, S. S. and McQueen, J. M. (2013). Speech and music shape the listening brain: Evidence for shared domain-general mechanisms. *Frontiers in Psychology*, 4(JUN):1–14.

[Ashburner and Friston, 2005] Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3):839–851.

[Atienza et al., 2001] Atienza, M., Cantero, J. L., and Escera, C. (2001). Auditory information processing during human sleep as revealed by event-related brain potentials. *Clinical Neurophysiology*, 112(11):2031–2045.

[Atilgan et al., 2018] Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K., and Bizley, J. K. (2018). Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. *Neuron*, 97(3):640–655.e4.

[Bamiou et al., 2003] Bamiou, D. E., Musiek, F. E., and Luxon, L. M. (2003). The insula (Island of Reil) and its role in auditory processing: Literature review. *Brain Research Reviews*, 42(2):143–154.

[Bastuji and García-Larrea, 1999] Bastuji, H. and García-Larrea, L. (1999). Evoked potentials as a tool for the investigation of human sleep.

[Bastuji et al., 2002] Bastuji, H., Perrin, F., and Garcia-Larrea, L. (2002). Semantic analysis of auditory input during sleep: Studies with event related potentials. *International Journal of Psychophysiology*, 46(3):243–255.

[Beauchamp et al., 2004] Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41(5):809–823.

[Beauchamp and Ro, 2008] Beauchamp, M. S. and Ro, T. (2008). Neural substrates of sound-touch synesthesia after a thalamic lesion. *Journal of Neuroscience*, 28(50):13696–13702.

[Bernstein et al., 2004] Bernstein, L. E., Auer, E. T., and Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, 44(1-4 SPEC. ISS.):5–18.

[Besle et al., 2004] Besle, J., Fort, A., and Giard, M.-H. (2004). Interest and validity of the additive model in electrophysiological studies of multisensory interactions. *Cognitive Processing*, 5(3):189–192.

[Bidet-Caulet et al., 2007] Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P. E., Giard, M. H., and Bertrand, O. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *Journal of Neuroscience*, 27(35):9252–9261.

[Bizley et al., 2016] Bizley, J. K., Maddox, R. K., and Lee, A. K. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. *Trends in Neurosciences*, 39(2):74–85.

[Bolognini et al., 2013] Bolognini, N., Convento, S., Rossetti, A., and Merabet, L. B. (2013). Multisensory processing after a brain damage: Clues on post-injury crossmodal plasticity from neuropsychology. *Neuroscience and Biobehavioral Reviews*, 37(3):269–278.

[Brainard, 1997] Brainard, D. H. a. (1997). The Psychophysics Toolbox. *Spatial Vision*.

[Bregman and McAdams, 1990] Bregman, A. S. and McAdams, S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. *The Journal of the Acoustical Society of America*, 95(2):1177–1178.

[Bregman and McAdams, 1994] Bregman, A. S. and McAdams, S. (1994). Auditory Scene Analysis: The Perceptual Organization of Sound. *The Journal of the Acoustical Society of America*, 95(2):1177–1178.

[Brochard et al., 2008] Brochard, R., Touzalin, P., Després, O., and Dufour, A. (2008). Evidence of beat perception via purely tactile stimulation. *Brain Research*, 1223:59–64.

[Broderick et al., 2018] Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*.

[Brodmann, 1909] Brodmann, K. (1909). Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues. *Barth*, 44(0).

[Brooks et al., 2018] Brooks, C. J., Chan, Y. M., Anderson, A. J., and McKendrick, A. M. (2018). Audiovisual Temporal Perception in Aging: The Role of Multisensory Integration and Age-Related Sensory Loss. *Frontiers in Human Neuroscience*, 12(May):1–9.

[Bullmore and Sporns, 2012] Bullmore, E. and Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349.

[Burton et al., 2008a] Burton, H., Sinclair, R. J., and McLaren, D. G. (2008a). Cortical network for vibrotactile attention: A fMRI study. *Human Brain Mapping*, 29(2):207–221.

[Burton et al., 2008b] Burton, H., Sinclair, R. J., Wingert, J. R., and Dierker, D. L. (2008b). Multiple parietal operculum subdivisions in humans: Tactile activation maps. *Somatosensory and Motor Research*, 25(3):149–162.

[Burton et al., 1993] Burton, H., Videen, T. O., and Raichle, M. E. (1993). Tactile-vibration-activated foci in insular and parietal-opercular cortex studied with positron emission tomography: Mapping the second somatosensory area in humans. *Somatosensory & Motor Research*, 10(3):297–308.

[Burunat et al., 2015] Burunat, I., Ristaniemi, T., Brattico, E., Alluri, V., Sams, M., Toiviainen, P., and Bogert, B. (2015). The reliability of continuous brain responses during naturalistic listening to music. *NeuroImage*, 124:224–231.

[Busse et al., 2005] Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., and 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(51):18751–18756.

[Butler et al., 2012] Butler, J. S., Foxe, J. J., Fiebelkorn, I. C., Mercier, M. R., and Molholm, S. (2012). Multisensory Representation of Frequency across Audition and Touch: High Density Electrical Mapping Reveals Early Sensory-Perceptual Coupling. *Journal of Neuroscience*, 32(44):15338–15344.

[Buzśaki and Wang, 2012] Buzśaki, G. and Wang, X. J. (2012). Mechanisms of gamma oscillations. *Annual Review of Neuroscience*, 35:203–225.

[Caetano and Jousmäki, 2006] Caetano, G. and Jousmäki, V. (2006). Evidence of vibrotactile input to human auditory cortex. *NeuroImage*, 29(1):15–28.

[Callan et al., 2003] Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Callan, A. M., and Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport*, 14(17):746–748.

[Calvert, 2001] Calvert, G. A. (2001). Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cerebral Cortex*, 11(12):1110–1123.

[Calvert et al., 1999] Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., and David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *NeuroReport*, 10(12):2619–2623.

[Calvert et al., 1997] Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, 276(5312):593–596.

[Calvert et al., 2004] Calvert, G. A., Spence, C., and Stein, B. E., editors (2004). *The handbook of multisensory processes.* MIT Press, Cambridge, MA, US.

[Campbell, 2008] Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1001–1010.

[Cappe and Barone, 2005] Cappe, C. and Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience*, 22(11):2886–2902.

[Cappe et al., 2009] Cappe, C., Morel, A., Barone, P., and Rouiller, E. M. (2009). The thalamocortical projection systems in primate: An anatomical support for multisensory and sensorimotor interplay. *Cerebral Cortex*, 19(9):2025–2037.

[Cash et al., 2009] Cash, S. S., Halgren, E., Dehghani, N., Rossetti, A. O., Thesen, T., Bromfield, E., and Ero, L. (2009). The Human K-Complex Represents an Isolated Cortical Down-State. (May):1084–1088.

[Chandrasekaran et al., 2009] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, 5(7).

[Chee and Chuah, 2008] Chee, M. W. and Chuah, L. Y. (2008). Functional neuroimaging insights into how sleep and sleep deprivation affect memory and cognition. *Current Opinion in Neurology*, 21(4):417–423.

[Chen et al., 2008] Chen, J. L., Penhune, V. B., and Zatorre, R. J. (2008). Listening to musical rhythms recruits motor regions of the brain. *Cerebral Cortex*, 18(12):2844–2854.

[Cherry, 1953] Cherry, E. C. (1953). Some experiments on the recognition of speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25(5):975–979.

[Chin and Rickard, 2012] Chin, T. and Rickard, N. S. (2012). The Music USE (MUSE) Questionnaire: An Instrument to Measure Engagement in Music. *Music Perception: An Interdisciplinary Journal*, 29(4):429–446.

[Chuen and Schutz, 2016] Chuen, L. and Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. *Attention, Perception, and Psychophysics*, 78(5):1512–1528.

[Cirelli and Tononi, 2008] Cirelli, C. and Tononi, G. (2008). Is sleep essential? *PLoS Biology*, 6(8):1605–1611.

[Cohen, 2011] Cohen, M. (2011). It's about time. *Frontiers in Human Neuroscience*, 5:2.

[Crosse et al., 2015] Crosse, M. J., Butler, J. S., and Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience*, 35(42):14195–14204.

[Crosse et al., 2016a] Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016a). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*.

[Crosse et al., 2016b] Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (2016b). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. *Journal of Neuroscience*.

[Darwin, 2008] Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1011–1021.

[Davis and Johnsrude, 2003] Davis, M. H. and Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, 23(8):3423–3431.

[De Gennaro and Ferrara, 2003] De Gennaro, L. and Ferrara, M. (2003). Sleep spindles: an overview. *Sleep Medicine Reviews*, 7(5):423–440.

[Degano et al., Prep] Degano, G., Ferrari, A., and Noppeney, U. (Prep). Audio-tactile integration in music.

[Di Liberto et al., 2020] Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., de Cheveigné, A., Shamma, S., and Mesgarani, N. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife*, 9:1–26.

[Diederich et al., 2008] Diederich, A., Colonius, H., and Schomburg, A. (2008). Assessing age-related multisensory enhancement with the time-window-of-integration model. *Neuropsychologia*, 46(10):2556–2562.

[Ding et al., 2017] Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., and Poeppel, D. (2017). Temporal Modulations in Speech and Music. *Neuroscience & Biobehavioral Reviews*.

[Ding and Simon, 2012] Ding, N. and Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11854–9.

[Ding and Simon, 2014] Ding, N. and Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience*.

[Disbergen et al., 2018] Disbergen, N. R., Valente, G., Formisano, E., and Zatorre, R. J. (2018). Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music. *Frontiers in Neuroscience*.

[Doelling and Poeppel, 2015] Doelling, K. B. and Poeppel, D. (2015). Cortical entrainment to music and its modulation by expertise. *Proceedings of the National Academy of Sciences*.

[Driver and Noesselt, 2008] Driver, J. and Noesselt, T. (2008). Multisensory Interplay Reveals Crossmodal Influences on 'Sensory-Specific' Brain Regions, Neural Responses, and Judgments. *Neuron*, 57(1):11–23.

[Eickhoff et al., 2006] Eickhoff, S. B., Amunts, K., Mohlberg, H., and Zilles, K. (2006). The human parietal operculum. II. Stereotaxic maps and correlation with functional imaging results. *Cerebral Cortex*, 16(2):268–279.

[Elton et al., 1997] Elton, M., Winter, O., Heslenfeld, D., Loewy, D., Campbell, K., and Kok, A. (1997). Event-related potentials to tones in the absence and presence of sleep spindles. *Journal of Sleep Research*, 6(2):78–83.

[Ernst and Bülthoff, 2004] Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept.

[Fan et al., 2016] Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., and Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex*, 26(8):3508–3526.

[Fiedler et al., 2017] Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., and Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of Neural Engineering*.

[Fishman and Michael, 1973] Fishman, M. C. and Michael, C. R. (1973). Integration of auditory information in the cat's visual cortex. *Vision Research*, 13(8):1415–1419.

[Foxe et al., 2000] Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., and Schroeder, C. E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Cognitive Brain Research*.

[Foxe and Schroeder, 2005] Foxe, J. J. and Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *NeuroReport*, 16(5):419–423.

[Foxe et al., 2002] Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., Ritter, W., and Murray, M. M. (2002). Auditory-somatosensory multisensory processing in auditory association cortex: An fMRI study. *Journal of Neurophysiology*, 88(1):540–543.

[Francis et al., 2000] Francis, S. T., Kelly, E. F., Bowtell, R., Dunseath, W. J., Folger, S. E., and McGlone, F. (2000). fMRI of the responses to vibratory stimulation of digit tips. *NeuroImage*, 11(3):188–202.

[Frassinetti et al., 2002] Frassinetti, F., Bolognini, N., and Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, 147(3):332–343.

[Friston et al., 2007] Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping*. Academic Press, London.

[Friston et al., 1994a] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994a). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210.

[Friston et al., 1994b] Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., and Evans, A. C. (1994b). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1(3):210–220.

[Fritz et al., 2007] Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007). Auditory attention - focusing the searchlight on sound. *Current Opinion in Neurobiology*, 17(4):437–455.

[Fu et al., 2003] Fu, K. M. G., Johnston, T. A., Shah, A. S., Arnold, L., Smiley, J., Hackett, T. A., Garraghty, P. E., and Schroeder, C. E. (2003). Auditory cortical neurons respond to somatosensory stimulation. *Journal of Neuroscience*, 23(20):7510–7515.

[Fujiwara et al., 2002] Fujiwara, N., Imai, M., Nagamine, T., Mima, T., Oga, T., Takeshita, K., Toma, K., and Shibasaki, H. (2002). Second somatosensory area (SII) plays a significant role in selective somatosensory attention. *Cognitive Brain Research*, 14(3):389–397.

[Gau et al., 2020] Gau, R., Bazin, P. L., Trampel, R., Turner, R., and Noppeney, U. (2020). Resolving multisensory and attentional influences across cortical depth in sensory cortices. *eLife*, 9:1–26.

[Gescheider, 1997] Gescheider, G. A. (1997). *Psychophysics: The fundamentals, 3rd ed.* Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

[Ghazanfar and Schroeder, 2006] Ghazanfar, A. A. and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, 10(6):278–285.

[Gillmeister and Eimer, 2007] Gillmeister, H. and Eimer, M. (2007). Tactile enhancement of auditory detection and perceived loudness. *Brain Research*, 1160(1):58–68.

[Giraud and Poeppel, 2012] Giraud, A. L. and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations.

[Glasser et al., 2013] Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., and Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.

[Gobbelé et al., 2003] Gobbelé, R., Schürmann, M., Forss, N., Juottonen, K., Buchner, H., and Hari, R. (2003). Activation of the human posterior parietal and temporoparietal cortices during audiotactile interaction. *NeuroImage*.

[Gordon, 1987] Gordon, J. W. (1987). The perceptual attack time of musical tones. *Journal of the Acoustical Society of America*, 82(1):88–105.

[Grant, 2001] Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America*, 109(5):2272–2275.

[Grimault et al., 2002] Grimault, N., Bacon, S. P., and Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *The Journal of the Acoustical Society of America*, 111(3):1340–1348.

[Guest et al., 2002] Guest, S., Catmur, C., Lloyd, D., and Spence, C. (2002). Audiotactile interactions in roughness perception. *Experimental Brain Research*.

[Hackett et al., 2007] Hackett, T. A., Smiley, J. F., Ulbert, I., Karmos, G., Lakatos, P., De La Mothe, L. A., and Schroeder, C. E. (2007). Sources of somatosensory input to the caudal belt areas of auditory cortex. *Perception*, 36(10):1419–1430.

[Hämäläinen et al., 2002] Hämäläinen, H., Hiltunen, J., and Titievskaja, I. (2002). Activation of somatosensory cortical areas varies with attentional state: An fMRI study. *Behavioural Brain Research*, 135(1-2):159–165.

[Harding et al., 2019] Harding, E. E., Sammler, D., Henry, M. J., Large, E. W., and Kotz, S. A. (2019). Cortical tracking of rhythm in music and speech. *NeuroImage*, 185:96–101.

[Haufe et al., 2014] Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110.

[Hausfeld et al., 2018] Hausfeld, L., Riecke, L., and Formisano, E. (2018). Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex. *NeuroImage*, 173(February):472–483.

[Henry and Obleser, 2012] Henry, M. J. and Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 109(49):20095–20100.

[Herdener et al., 2009] Herdener, M., Lehmann, C., Esposito, F., Di Salle, F., Federspiel, A., Bach, D. R., Scheffler, K., and Seifritz, E. (2009). Brain responses to auditory and visual stimulus offset: Shared representations of temporal edges. *Human Brain Mapping*, 30(3):725–733.

[Hickok and Poeppel, 2007] Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature PERSPECTIVES*, 8(May):393–402.

[Higgins et al., 2008] Higgins, N. C., Escabí, M. A., Rosen, G. D., Galaburda, A. M., and Read, H. L. (2008). Spectral processing deficits in belt auditory cortex following early postnatal lesions of somatosensory cortex. *Neuroscience*, 153(2):535–549.

[Hoefer et al., 2013] Hoefer, M., Tyll, S., Kanowski, M., Brosch, M., Schoenfeld, M. A., Heinze, H. J., and Noesselt, T. (2013). Tactile stimulation and hemispheric asymmetries modulate auditory perception and neural responses in primary auditory cortex. *NeuroImage*.

[Hoefle et al., 2018] Hoefle, S., Engel, A., Basilio, R., Alluri, V., Toiviainen, P., Cagy, M., and Moll, J. (2018). Identifying musical pieces from fMRI data using encoding and decoding models. *Scientific Reports*, 8(1):1–13.

[Ibáñez et al., 2006] Ibáñez, A., López, V., and Cornejo, C. (2006). ERPs and contextual semantic discrimination: Degrees of congruence in wakefulness and sleep. *Brain and Language*, 98(3):264–275.

[James and Stevenson, 2012] James, W. J. and Stevenson, A. R. (2012). *The Use of fMRI to Assess Multisensory Integration*. CRC Press Taylor and Francis.

[Johansen-Berg et al., 2000] Johansen-Berg, H., Christensen, V., Woolrich, M., and Matthews, P. M. (2000). Attention to touch modulates activity in both primary and secondary somatosensory areas. *NeuroImage*, 11(5 PART II):1237–1241.

[Joon Kim et al., 2007] Joon Kim, Y., Grabowecky, M., Paller, K. A., Muthu, K., and Suzuki, S. (2007). Attention induces synchronization-based response gain in steady-state visual evoked potentials. *Nature Neuroscience*, 10(1):117–125.

[Jousmäki and Hari, 1998] Jousmäki, V. and Hari, R. (1998). Parchment-skin illusion: sound-biased touch. *Current Biology*, 8(6):R190–R191.

[Kassuba et al., 2013] Kassuba, T., Menz, M. M., Röder, B., and Siebner, H. R. (2013). Multisensory interactions between auditory and haptic object recognition. *Cerebral Cortex*, 23(5):1097–1107.

[Kayser et al., 2005] Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron*, 48(2):373–384.

[Kayser et al., 2007] Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *Journal of Neuroscience*, 27(8):1824–1835.

[Keitel et al., 2020] Keitel, A., Gross, J., and Kayser, C. (2020). Shared and modality-specific brain regions that mediate auditory and visual word comprehension. *eLife*, 6.

[Kouider and Dehaene, 2007] Kouider, S. and Dehaene, S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):857–875.

[Kunelskaya et al., 2005] Kunelskaya, N. L., Levina, Y. V., Garov, E. V., Dzuina, A. V., Ogorodnikov, D. S., Nosulya, E. V., and Luchsheva, Y. V. (2005). Presbyacusis. *Lancet*, 366:1111–20.

[Lakatos et al., 2007] Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., and Schroeder, C. E. (2007). Neuronal Oscillations and Multisensory Interaction in Primary Auditory Cortex. *Neuron*, 53(2):279–292.

[Large and Palmer, 2002] Large, E. W. and Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science*, 26(1):1–37.

[Laurienti et al., 2006] Laurienti, P. J., Burdette, J. H., Maldjian, J. A., and Wallace, M. T. (2006). Enhanced multisensory integration in older adults. *Neurobiology of Aging*, 27(8):1155–1163.

[Laurienti et al., 2002] Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y. F., Field, A. S., and Stein, B. E. (2002). Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of Cognitive Neuroscience*, 14(3):420–429.

[Laurienti et al., 2005] Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., and Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, 166(3-4):289–297.

[Laurino et al., 2014] Laurino, M., Menicucci, D., Piarulli, A., Mastorci, F., Bedini, R., Allegrini, P., and Gemignani, A. (2014). Disentangling different functional roles of evoked K-complex components: Mapping the sleeping brain while quenching sensory processing. *NeuroImage*, 86:433–445.

[Legendre et al., 2019] Legendre, G., Andrillon, T., Koroma, M., and Kouider, S. (2019). Sleepers track informative speech in a multitalker environment. *Nature Human Behaviour*.

[Léger et al., 2018] Léger, D., Debellemaniere, E., Rabat, A., Bayon, V., Benchenane, K., and Chennaoui, M. (2018). Slow-wave sleep: From the cell to the clinic. *Sleep Medicine Reviews*, 41:113–132.

[Leonardelli et al., 2015] Leonardelli, E., Braun, C., Weisz, N., Lithari, C., Occelli, V., and Zampini, M. (2015). Prestimulus oscillatory alpha power and connectivity patterns predispose perceptual integration of an audio and a tactile stimulus. *Human Brain Mapping*, 36(9):3486–3498.

[Loomis et al., 1935] Loomis, A. L., Harvey, E. N., and Hobart, G. (1935). Potential rhythms of the cerebral cortex during sleep. *Science*, 81(2111):597–598.

[Luo et al., 2010] Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biology*, 8(8):25–26.

[Macaluso, 2006] Macaluso, E. (2006). Multisensory processing in sensory-specific cortical areas. *Neuroscientist*, 12(4):327–338.

[Macaluso et al., 2000] Macaluso, E., Frith, C. D., and Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science*, 289(5482):1206–1208.

[Macaluso et al., 2016] Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., and Adam, R. (2016). The Curious Incident of Attention in Multisensory Integration: Bottom-up vs. Top-down. *Multisensory Research*, 29(6):557–583.

[Maddox et al., 2015] Maddox, R. K., Atilgan, H., Bizley, J. K., and Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *eLife*, 4:1–11.

[Maddox and Shinn-Cunningham, 2012] Maddox, R. K. and Shinn-Cunningham, B. G. (2012). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *JARO - Journal of the Association for Research in Otolaryngology*, 13(1):119–129.

[Maguire, 2012] Maguire, E. A. (2012). Studying the freely-behaving brain with fMRI. *NeuroImage*, 62(2):1170–1176.

[Martuzzi et al., 2007] Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J. P., Maeder, P. P., Clarke, S., and Meuli, R. A. (2007). Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cerebral Cortex*, 17(7):1672–1679.

[Matusz et al., 2019] Matusz, P. J., Dikker, S., Huth, A. G., and Perrodin, C. (2019). Are We Ready for Real-world Neuroscience? *Journal of Cognitive Neuroscience*, 31(3):327–338.

[Mccormick, 1994] Mccormick, D. A. (1994). Sensory gating mechanisms of the thalamus. *Current Opinion in Neurobiology*, pages 550–556.

[Merchant et al., 2015] Merchant, H., Grahn, J., Trainor, L., Rohrmeier, M., and Fitch, W. T. (2015). Finding the beat: A neural perspective across humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664).

[Meredith, 2002] Meredith, M. A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research*, 14(1):31–40.

[Meredith and Stein, 1983] Meredith, M. A. and Stein, B. E. (1983). Interactions Among Converging Sensory Inputs in the Superior Colliculus. *Science*, 369.

[Meredith and Stein, 1984] Meredith, M. A. and Stein, B. E. (1984). Descending Efferents from the Superior Colliculus Relay Integrated Multisensory Information. *Science*, 30.

[Mesgarani and Chang, 2012] Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception.

[Mesgarani et al., 2009] Mesgarani, N., David, S., Fritz, J. B., and Shamma, S. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of neurophysiology*, 102(6):3329.

[Middlebrooks and Bremen, 2013] Middlebrooks, J. C. and Bremen, P. (2013). Spatial stream segregation by auditory cortical neurons. *Journal of Neuroscience*, 33(27):10986–11001.

[Middlebrooks et al., 2017] Middlebrooks, J. C., Simon, J. Z., Popper, A. N., and Fay, R. R., editors (2017). *The Auditory System at the Cocktail Party*, volume 60 of *Springer Handbook of Auditory Research*. Springer International Publishing, Cham.

[Molholm et al., 2002] Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research*, 14(1):115–128.

[Moray, 1959] Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology*, 11(1):56–60.

[Mozolic et al., 2012] Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., and Laurienti, P. J. (2012). Multisensory Integration and Aging. *The Neural Bases of Multisensory Processes*, (2008):1–12.

[Mudrik et al., 2014] Mudrik, L., Faivre, N., and Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences*, 18(9):488–496.

[Murray and Wallace, 2012] Murray, M. and Wallace, M., editors (2012). *The Neural Bases of Multisensory Processes*. Boca Raton CRC Pres.

[Murray et al., 2005] Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2005). Grabbing your ear: Rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cerebral Cortex*.

[Musacchia and Schroeder, 2009] Musacchia, G. and Schroeder, C. E. (2009). Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. *Hearing Research*.

[Muzur et al., 2002] Muzur, A., Pace-Schott, E. F., and Hobson, J. A. (2002). The prefrontal cortex in sleep. *Trends in Cognitive Sciences*, 6(11):475–481.

[Naselaris et al., 2011] Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI.

[Nelson et al., 2004] Nelson, A. J., Staines, W. R., Graham, S. J., and McIlroy, W. E. (2004). Activation in SI and SII; The influence of vibrotactile amplitude during passive and task-relevant stimulation. *Cognitive Brain Research*, 19(2):174–184.

[Noppeney, 2012] Noppeney, U. (2012). Characterization of Multisensory Integration with fMRI. *The Neural Bases of Multisensory Processes*, pages 1–21.

[Noppeney and Lee, 2018] Noppeney, U. and Lee, H. L. (2018). Causal inference and temporal predictions in audiovisual perception of speech and music. *Annals of the New York Academy of Sciences*, 1423(1):102–116.

[Nozaradan et al., 2011] Nozaradan, S., Peretz, I., Missal, M., and Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(28):10234–10240.

[Odgaard et al., 2004] Odgaard, E. C., Arieh, Y., and Marks, L. E. (2004). Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cognitive, Affective and Behavioral Neuroscience*, 4(2):127–132.

[Ogilvie, 2001] Ogilvie, R. D. (2001). The process of falling asleep. *Sleep Medicine Reviews*, 5(3):247–270.

[Oldfield, 1971] Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9(1):97–113.

[Oostenveld et al., 2011] Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*.

[O'Sullivan et al., 2019] O'Sullivan, A. E., Lim, C. Y., and Lalor, E. C. (2019). Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations. *European Journal of Neuroscience*, 50(8):3282–3295.

[O'Sullivan et al., 2015] O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, 25(7):1697–1706.

[Oswald et al., 1960] Oswald, I., Taylor, A. M., and Treisman, M. (1960). Discriminative responses to stimulation during human sleep. *Brain*, 83(3):440–453.

[Oyanedel et al., 2020] Oyanedel, C. N., Durán, E., Niethard, N., Inostroza, M., and Born, J. (2020). Temporal associations between sleep slow oscillations, spindles and ripples. *European Journal of Neuroscience*, 52(12):4762–4778.

[Park et al., 2018] Park, H., Ince, R. A., Schyns, P. G., Thut, G., and Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS Biology*, 16(8):1–26.

[Park et al., 2016] Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*.

[Peelen and Kastner, 2014] Peelen, M. V. and Kastner, S. (2014). Attention in the real world: Toward understanding its neural basis. *Trends in Cognitive Sciences*, 18(5):242–250.

[Peelle and Wingfield, 2016] Peelle, J. E. and Wingfield, A. (2016). The Neural Consequences of Age-Related Hearing Loss. *Trends in Neurosciences*, 39(7):486–497.

[Pekkola et al., 2005] Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., and Sams, M. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *NeuroReport*, 16(2):125–128.

[Penhune and Zatorre, 2019] Penhune, V. B. and Zatorre, R. J. (2019). Rhythm and time in the premotor cortex. *PLoS Biology*, 17(6):1–6.

[Peretz and Zatorre, 2005] Peretz, I. and Zatorre, R. J. (2005). Brain Organization for Music Processing. *Annual Review of Psychology*, 56(1):89–114.

[Pérez-Bellido et al., 2018] Pérez-Bellido, A., Anne Barnes, K., Crommett, L. E., and Yau, J. M. (2018). Auditory Frequency Representations in Human Somatosensory Cortex. *Cerebral Cortex*, 28(11):3908–3921.

[Petkov et al., 2004] Petkov, C. I., Kang, X., Alho, K., Bertrand, O., Yund, E. W., and Woods, D. L. (2004). Attentional modulation of human auditory cortex. *Nature Neuroscience*, 7(6):658–663.

[Phillips et al., 2011] Phillips, C., Leclercq, Y., Schrouff, J., Noirhomme, Q., and Maquet, P. (2011). FMRI artefact rejection and sleep scoring toolbox. *Computational Intelligence and Neuroscience*, 2011.

[Poldrack et al., 2020] Poldrack, R. A., Huckins, G., and Varoquaux, G. (2020). Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*, 77(5):534–540.

[Portas et al., 2000] Portas, C. M., Krakow, K., Allen, P., Josephs, O., Armony, J. L., and Frith, C. D. (2000). Auditory Processing across the Sleep-Wake Cycle. *Neuron*, 28(3):991–999.

[Pourtois et al., 2005] Pourtois, G., Gelder, B. D., Bol, A., and Crommelinck, M. (2005). Perception of Facial Expressions and Voices and of Their. *Cortex*, 41(1):49–59.

[Rasch and Born, 2013] Rasch, B. and Born, J. (2013). About sleep's role in memory. *Physiological Reviews*, 93(2):681–766.

[Reed et al., 2004] Reed, C. L., Shoham, S., and Halgren, E. (2004). Neural Substrates of Tactile Object Recognition: An fMRI Study. *Human Brain Mapping*, 21(4):236–246.

[Reisberg et al., 1987] Reisberg, D., McLean, J., and Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. *Hearing by eye: The psychology of lip-reading*, pages 97–113.

[Remedios et al., 2009] Remedios, R., Logothetis, N. K., and Kayser, C. (2009). An auditory region in the primate insular cortex responding preferentially to vocal communication sounds. *Journal of Neuroscience*, 29(4):1034–1045.

[Riecke et al., 1995] Riecke, F., Bodnar, D., and Bialek, W. (1995). Naturalistic stim uli increase the rate and efficiency of inform ation transm ission by prim ary auditory afferents. *Proceedings Royal Society B*, 262:259–265.

[Riecke et al., 2019] Riecke, L., Snipes, S., van Bree, S., Kaas, A., and Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *NeuroImage*, 202(April):116134.

[Rimmele et al., 2018] Rimmele, J. M., Morillon, B., Poeppel, D., and Arnal, L. H. (2018). Proactive Sensing of Periodic and Aperiodic Auditory Patterns. *Trends in Cognitive Sciences*, 22(10):870–882.

[Ro et al., 2013] Ro, T., Ellmore, T. M., and Beauchamp, M. S. (2013). A neural link between feeling and hearing. *Cerebral Cortex*, 23(7):1724–1730.

[Ro et al., 2009] Ro, T., Hsu, J., Yasar, N. E., Caitlin Elmore, L., and Beauchamp, M. S. (2009). Sound enhances touch perception. *Experimental Brain Research*, 195(1):135–143.

[Ross et al., 2007] Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17(5):1147–1153.

[Ruby et al., 2008] Ruby, P., Caclin, A., Boulet, S., Delpuech, C., and Morlet, D. (2008). Odd sound processing in the sleeping brain. *Journal of Cognitive Neuroscience*, 20(2):296–311.

[Sams et al., 1991] Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., and Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, 127(1):141–145.

[Sanders et al., 2013] Sanders, R. D., Tononi, G., Laureys, S., and Sleigh, J. (2013). Unconsciousness, not equal to unresponsiveness. *Anesthesiology*, 116(4):946–959.

[Santoro et al., 2014] Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*.

[Scheirer, 1998] Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103(1):588–601.

[Schroeder and Foxe, 2005] Schroeder, C. E. and Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology*, 15(4):454–458.

[Schroeder and Foxe, 2002] Schroeder, C. E. and Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research*, 14(1):187–198.

[Schroeder et al., 2008] Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, 12(3):106–113.

[Schroeder et al., 2001] Schroeder, C. E., Lindsley, R. W., Specht, C., Marcovici, A., Smiley, J. F., and Javitt, D. C. (2001). Somatosensory input to auditory association cortex in the macaque monkey. *Journal of Neurophysiology*, 85(3):1322–1327.

[Schroeder et al., 2003] Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O'Connell, M. N., and Hackett, T. A. (2003). Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology*, 50(1-2):5–17.

[Schürmann et al., 2006] Schürmann, M., Caetano, G., Hlushchuk, Y., Jousmäki, V., and Hari, R. (2006). Touch activates human auditory cortex. *NeuroImage*, 30(4):1325–1331.

[Sepulcre, 2014] Sepulcre, J. (2014). Functional streams and cortical integration in the human brain. *Neuroscientist*, 20(5):499–508.

[Shamma et al., 2011] Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3):114–123.

[Shannon et al., 1995] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304.

[Sharon et al., 2017] Sharon, O., Ding, N., and Ben-shachar, M. (2017). Sleep disrupts high-level speech parsing despite significant basic auditory processing. *Journal of Neuroscience*.

[Shinn-Cunningham, 2008] Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5):182–186.

[Simpson, 1977] Simpson, J. A. (1977). Handbook of Sensory Physiology Volume 3 Auditory System, Clinical and Special Topics.

[Singh, 1987] Singh, P. G. (1987). Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre. *Journal of the Acoustical Society of America*, 82(3):886–899.

[Smiley et al., 2007] Smiley, J. F., Hackett, T. A., Ulbert, I., Karmas, G., Lakatos, P., Javitt, D. C., and Schroeder, C. E. (2007). Multisensory convergence in auditory cortex, I. Cortical connections of the caudal superior temporal plane in macaque monkeys. *The Journal of Comparative Neurology*, 502(6):894–923.

[Soto-Faraco and Deco, 2009] Soto-Faraco, S. and Deco, G. (2009). Multisensory contributions to the perception of vibrotactile events. *Behavioural Brain Research*, 196(2):145–154.

[Soto-Faraco et al., 2019] Soto-Faraco, S., Kvasova, D., Biau, E., Ikumi, N., Ruzzoli, M., Morís-Fernández, L., and Torralba, M. (2019). *Multisensory Interactions in the Real World*. Elements in Perception. Cambridge University Press.

[Stein, 2012] Stein, B. E., editor (2012). *The New Handbook of Multisensory Processing*. MIT Press.

[Stein and Meredith, 1993] Stein, B. E. and Meredith, M. A. (1993). *The merging of the senses.* Cognitive neuroscience. The MIT Press, Cambridge, MA, US.

[Stein and Stanford, 2008] Stein, B. E. and Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4):255–266.

[Steinmetz et al., 2000] Steinmetz, P. N., Roy, A., Fitzgerald, P. J., Hsiao, S. S., Johnson, K. O., and Niebur, E. (2000). Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, 404(6774):187–190.

[Steriade, 2003] Steriade, M. (2003). The corticothalamic system in sleep. *Frontiers in bioscience : a journal and virtual library*, 8:1093–9946.

[Strauss et al., 2015] Strauss, M., Sitt, J. D., King, J.-R., Elbaz, M., Azizi, L., Buiatti, M., Naccache, L., van Wassenhove, V., and Dehaene, S. (2015). Disruption of hierarchical predictive coding during sleep. *Proceedings of the National Academy of Sciences*.

[Sumby and Pollack, 1954] Sumby, W. H. and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, 26(2):212–215.

[Summerfield, 1992] Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 335(1273):71–78.

[Todd, 1912] Todd, J. W. (1912). *Reaction to multiple stimuli.* Archives of psychology. The Science Press.

[Tranchant et al., 2017] Tranchant, P., Shiell, M. M., Giordano, M., Nadeau, A., Peretz, I., and Zatorre, R. J. (2017). Feeling the beat: Bouncing synchronization to vibrotactile music in hearing and early deaf people. *Frontiers in Neuroscience*, 11(SEP):1–8.

[Treisman and Gelade, 1980] Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

[Van Essen, 1997] Van Essen, D. C. (1997). A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature*, 385(6614):313–318.

[van Noorden, 1975] van Noorden, L. P. A. S. (1975). Temporal Coherence in the Perception of Tone Sequences. *Institute for Perception Research*, Ph. D.(1975).

[Vroomen and de Gelder, 2000] Vroomen, J. and de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision.

[Vyazovskiy and Harris, 2013] Vyazovskiy, V. V. and Harris, K. D. (2013). Sleep and the single neuron: The role of global slow oscillations in individual cell rest. *Nature Reviews Neuroscience*, 14(6):443–451.

[Werner and Noppeney, 2010] Werner, S. and Noppeney, U. (2010). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex*, 20(8):1829–1842.

[Wickens, 2001] Wickens, T. D. (2001). *Elementary Signal Detection Theory*. Oxford University Press,.

[Wickremaratchi and Llewelyn, 2006] Wickremaratchi, M. M. and Llewelyn, J. G. (2006). Effects of ageing on touch. *Postgraduate Medical Journal*, 82(967):301–304.

[Wilsch et al., 2018] Wilsch, A., Neuling, T., Obleser, J., and Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. *NeuroImage*, 172:766–774.

[Wilson et al., 2010] Wilson, E. C., Braida, L. D., and Reed, C. M. (2010). Perceptual interactions in the loudness of combined auditory and vibrotactile stimuli. *The Journal of the Acoustical Society of America*, 127(5):3038–3043.

[Yang et al., 1992] Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory Representations of Acoustic Signals. *IEEE Transactions on Information Theory*, 38(2):824–839.

[Yau et al., 2009] Yau, J. M., Olenczak, J. B., Dammann, J. F., and Bensmaia, S. J. (2009). Temporal Frequency Channels Are Linked across Audition and Touch. *Current Biology*, 19(7):561–566.

[Yau et al., 2010] Yau, J. M., Weber, A. I., and Bensmaia, S. J. (2010). Separate mechanisms for audio-tactile pitch and loudness interactions. *Frontiers in Psychology*.

[Zampini et al., 2007] Zampini, M., Torresan, D., Spence, C., and Murray, M. M. (2007). Auditory-somatosensory multisensory interactions in front and rear space. *Neuropsychologia*, 45(8):1869–1877.

[Zhang et al., 2019] Zhang, Y., Zhou, W., Wang, S., Zhou, Q., Wang, H., Zhang, B., Huang, J., Hong, B., and Wang, X. (2019). The Roles of Subdivisions of Human Insula in Emotion Perception and Auditory Processing. *Cerebral Cortex*, 29(2):517–528.

[Zion Golumbic et al., 2013] Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a ”cocktail party”. *The Journal of neuroscience*, 33(4):1417–1426.

[Zumer et al., 2020] Zumer, J., White, T. P., and Noppeney, U. (2020). The neural mechanisms of audiotactile binding depend on asynchrony. *European Journal of Neuroscience2*, pages 0–3.