The cognition of auditory smiles: a computational approach

Pablo Arias Sarah

DOCTORAL THESIS FROM SORBONNE UNIVERSITÉ

The cognition of auditory smiles: a computational approach

Author:
Pablo ARIAS SARAH

Supervisors:
Jean-Julien AUCOUTURIER
Pascal BELIN
Patrick SUSINI

Reviewers:
Rachael JACK
Tecumseh FITCH

Examiners:
Julie GRÈZES
Catherine PELACHAUD
Martine GAVARET

CNRS - IRCAM - Sorbonne Université - ED3C

December 18, 2018

Abstract

Pablo ARIAS SARAH

The cognition of auditory smiles: a computational approach

Emotions are the fuel of human survival and social development. Not only do we undergo primitive reflexes mediated by ancient brain structures, but we also consciously and unconsciously regulate our emotions in social contexts, affiliating with friends and distancing from foes. One of our main tools for emotion regulation is facial expression and, in particular, smiles. Smiles are deeply grounded in human behavior: they develop early, and are used across cultures to communicate affective states. The mechanisms that underlie their cognitive processing include interactions not only with visual, but also emotional and motor systems. Smiles, trigger facial imitation in their observers, reactions thought to be a key component of the human capacity for empathy.

Smiles, however, are not only experienced visually, but also have audible consequences. Although visual smiles have been widely studied, almost nothing is known about the cognitive processing of their auditory counterpart. This is the aim of this dissertation. In this work, we characterise and model the smile acousitc fingerprint, and use it to probe how auditory smiles are processed cognitively. We give here evidence that (1) auditory smiles can trigger unconscious facial imitation, that (2) they are cognitively integrated with their visual counterparts during perception, and that (3) the development of these processes does not depend on pre-learned visual associations. We conclude that the embodied mechanisms associated to the visual processing of facial expressions of emotions are in fact equally found in the auditory modality, and that their cognitive development is at least partially independent from visual experience.

Résumé

Pablo ARIAS SARAH

The cognition of auditory smiles: a computational approach

Les émotions sont essentielles à notre survie et à notre développement social. Non seulement elles nous servent au travers de réflexes primitifs, mais aussi comme moyen de régulation de nos interactions sociales. Les outils fondamentaux de cette communication sociale sont les expressions faciales émotionnelles, dont une des plus importantes est le sourire. Le sourire se développe tôt pendant l'enfance, et est utilisé comme un instrument de communication affectif à travers les cultures. Les mécanismes cognitifs responsables de sa perception impliquent des interactions avec des systèmes visuels, émotionnels et moteurs. En particulier, l'observation d'un sourire entraîne typiquement une imitation faciale spontanée, une réaction qui est considérée comme essentielle à notre capacité d'empathie avec l'autre.

Malgré de grandes avancées scientifiques sur la cognition des sourires visuels, très peu de travaux se sont intéressés à leur perception auditive. C'est le but de cette thèse. Ici, nous caractérisons et modélisons les conséquences acoustiques des sourires, et nous utilisons ces modèles pour étudier comment ils sont traités cognitivement. Nos résultats montrent (1) que les sourires auditifs induisent de l'imitation faciale chez leurs auditeurs, (2) qu'ils sont intégrés de façon multimodale aux indices visuels, et (3) que ces processus ne dépendent pas d'associations visuelles préalablement établies. Nous concluons que les mécanismes de cognition incarnés typiquement associés au traitement visuel des expressions émotionnelles se retrouvent dans la modalité auditive, et que leur développement est au moins partiellement indépendant de l'expérience visuelle.

Acknowledgements

I would like to deeply acknowledge my doctoral advisors. First, Jean-Julien Aucouturier for his constant support, effervescent energy and deep intellectual stimulation. Jean-Julien made of this thesis a highly pleasurable, didactic, and enjoyable experience —besides the usual PhD instability that one has to go through while doing a PhD, which, in retrospective, may be most of the fun. Second, Pascal Belin, for his enthusiastic and emotional comments throughout the thesis, and for teaching me some of the social codes in the cognitive sciences, as well as smart and strategic scientific moves. Third, Patrick Susini, not only for making this thesis possible, but also for his dedication to the research team, and his wise experimental and meta-theoretical advise.

I am also very grateful for the thesis committee, not only for accepting to review this work, but also for being, through their own work, an important source of inspiration.

None of this work would have been possible without the IRCAM institution and all the people being part of it. Thank you for accompanying me physically and intellectually from master student to doctor, naturally combining work, pleasure, music, art, and innovation. To all (my) wonderful, full-of-knowledge, post-doc colleagues: Emmanuel, Louise, Guillaume, Michael and Marco; thank you for teaching me so much, and always being there to listen or give advice, be it in musical, experimental, statistical or human terms.

I'm also extremely grateful with the constantly-evolving Perception and Sound Design team (Olivier, Nicolas, Mondher, Maxime, and all the interns) for providing a perfectly well-balanced cocktail of pleasurable, warm and smooth interactions during winter, and calm and concentration in the summertime. A very special acknowledgment to all the people who worked in the CREAM (Cracking the Emotional Code of Music) project during these four years: Vasso, Lou, Emmanuele, Laura, Daniel, Louise Vasa, Mailys, Mael...More generally, warm greetings to all my collaborators and colleagues from different institutions: Caren Bellman, Ousemma Bouafif, Juan-Jose

Burred, Greg Bryant, Ed Large, Benjamin Morillon, Ani Patel, Axel Roebel, Ana Saitovitch, Daniele Schön, Renaud Seguier, Catherine Soladié, Indiana Wollman, Monica Zilbovicus, Sarra Zaeid.

Finally, I can not forget to thank all the team from INSEAD (Huong, Jean-Yves, Germain, Nico, Seb, Liselott) as well as all the people that made this facility available to the PhD students from Sorbonne Université. It was an extraordinary place, not only to spend some time outside the laboratory, but also to collect data very effectively. Without this research facility my thesis would have at least one chapter less.

More personally, this work would not have been possible without the help and support of my family in the distant Colombia. Particularly, with the help of my mother Monique, for whom I am incredibly grateful. Not only for all the skype-based virtual engagement, management and encouragement, but also for the unconditional listening, interest and love. To my brothers Santiago and Camilo for being who they are; to my father for teaching me how to live; and to Kika for her deep listening.

Finally, cheers to all my dear friends in France, Colombia, and through the world, who made of this journey a profoundly gratifying experience: Adrien, Fefo, Olivier, Léo, Yahia, Pam, Abreo, David, Solenne, Vivi, and many many more. And then there is Audrey, whom I can't thank enough. Thank you for your support, patience, profound humanity, tears and laughs...

Contents

1 Introduction

Foreword

This dissertation introduces and uses the concept of auditory smile, the acoustic consequences of smiling while speaking, in order to probe the mechanisms that underly the cognitive processing of emotional facial expressions. As we will see, the study of auditory smiles allows us to take a step back from the unimodal approach usually taken to study facial expressions, to a broader multimodal approach which considers both their visual and auditory consequences.

Auditory smiles lie at the intersection of the four scientific fields of study: emotion, facial mimicry, speech perception and audiovisual integration. Thus, this introductory chapter will review theory and evidence from these four fields, and show that studying auditory smiles has potential not only to draw important theoretical parallels between these fields, but also to complement current models on e.g. the cognitive processing of emotional facial expressions.

First, because auditory smiles are a signal, in the auditory modality, of an emotional facial expression, I will briefly review key concepts of emotions: their definition, their function, but also how they are communicated, and what their underlying neural processes are. Particular care will be given to *facial* expressions of emotions, as they shed light on the origin and function of emotions, which will be of importance in the rest of the manuscript.

Second, because an important aspect of the cognitive processing of visual smiles is that they are often imitated by observers during social interactions, and because a large portion of the work in this dissertation will explore facial imitative responses to auditory smiles, I will review the field of mimicry and its theoretical link to empathy. Recent models suggest that these imitatory/embodied mechanisms are used in social contexts as tools for emotion communication and regulation and that thus support our basic capacity to

appraise the inner (e.g. affective) state of an individual. I'll support these views with the data from the emotional mimicry literature, which measures the microscopic facial reactions produced by people when observing others' emotional facial expressions. Emotional mimicry will serve as an experimental paradigm to probe the emotional processing of auditory smiles in chapter 4 and 6.

Third, because perceiving smiles from the voice is akin to decoding a motor gesture (the contraction of the zygomatic major muscle) from its acoustic consequence in speech sounds, I will review how vocal gestures are produced and perceived. To this aim, I will introduce key concepts about the vocal apparatus, its main components and acoustic properties. Then, I will present the motor theory of speech perception, and the more recent dual-stream model of speech perception, which separates neural processing pathways into ventral and dorsal streams, the latter being responsible for acoustico-motor coupling and articulation processing.

Finally, because in real-world interactions auditory smiles are often perceived simultaneously with the visual signal of the smile, and because part of this thesis will use eye-tracking to examine how auditory and visual smiles interact, I will review the literature on audiovisual integration, with a specfic emphasis on the findings in the fields of eye tracking and face-voice integration. As such, I will present data showing that audiovisual integration takes place at the very low-levels of sensory processing, but also at higher levels of emotional processing, which suggest that visual and auditory information are jointly integrated during perception, creating illusions, misperceptions, and shared multimodal percepts.

To conclude this introduction, I will describe the case study of interest in this dissertation, auditory smiles. I will briefly present the smiling gesture as one of the most important emotional behaviors in humans, recognised across cultures, developed early during infancy, and serving a crucial affiliation function. I will then present how the acoustic consequences of smiles in speech relate to emotion perception, and the different views on the acoustic features involved in smiled speech. For theoretical purposes, and to distanciate this dissertation from the study of the more broad smiled or amused speech, I will define auditory smiles as "the acoustic consequence of the smile facial expression in the voice" —which perception involves the auditory processing of the smile motor gesture. This simple formulation will allow us to ensure

causality from gesture to sound, perception and action, across the dissertation, and to raise some straightforward questions, which will be addressed in the form of separate chapters (2-6):

- Chapter 2. Do auditory smiles have a specific acoustic fingerprint? How are they internally represented?
- Chapter 3. Can the acoustic consequences of smile articulation be modelled computationally?
- Chapter 4. Does the perception of auditory and visual smiles share emotional processes? Specifically, do auditory smiles trigger the imitative behaviors usually associated with their visual counterparts?
- Chapter 5. How is smile-specific information from visual and auditory inputs integrated during perception?
- Chapter 6. Are the motor reactions observed during the perception of auditory smiles independent from visual experience?

1.1 Emotion

1.1.1 What is an emotion?

Definition

Emotions are one of the most intriguing human behaviors. Scientists across disciplines (biology, psychology, philosophy and neuroscience) have tried to tackle the difficult task of defining something as close to the human nature as emotions are, without yet agreeing on a broadly accepted definition (Izard, 2009). From emotions being dissmissed as a *"category of fictional causes of behavior"* (Skinner 1953), to more modern advances on emotional schemas, neural networks, neurochemestry, and the most recent frameworks integrating consciousness to the emotional experience, science has made a great progress through the debate. In this process, several scientific fields have contributed to dissociate three key characteristics of emotions.

First, emotions manifest themselves by rapid physiological changes, which are in some cases termed physiological arousal. Emotions can trigger pupil dilation, heart beat, skin conductance and somatic variations, as well as changes in body language and facial expressions (Hoehl et al., 2017; Gross, 1998; Darwin, 1872).

That emotions are associated by bodily changes has been known for a long time. In fact, in one of the first attempts to define emotions, 134 years ago, James (1884) equated them with the sole cognitive evaluation of these physiological reactions. The classic example given to explain James' theory is the following. Imagine you are walking down a forest and suddenly you see an enormous bear. As soon as you see the bear, your body will react in several ways: you will begin to sweat, your heart rate will increase, and you will start to run. James provocative view (at least what was kept of it in subsequent debates; Ellsworth, 1994) suggested that these physiological responses are what we call emotions. In other words, *"we feel sorry because we cry, angry because we strike, afraid because we tremble, and [it is] not that we cry, strike, or tremble, because we are sorry, angry, or fearful"* (James, 1884; p190). Although it is now known that this sequentiality of events (physiology preceding appraisal) is not causally univocal (as the same physiological states seem to trigger different emotional responses; Barrett, 2017), physiological reactions are still considered a key component of emotions.

Second, it is now accepted that emotions have (at least partially) dedicated neural systems. Thanks to extended animal research, we know that we share core emotional operating systems in ancient subcortical regions with some mamals (Panksepp, 2007). The very inventor of the term *affective neuroscience*, the psychobiologist and neuroscientist Jaak Panksepp, describes in his work seven systems, each one depending on specific neural substrates, present both in humans and non-human animals : Seeking, rage, fear, lust, care, panic and play. Consistently, causal manipulations of such systems (chemical or electrical) in both humans and animals modify emotional behaviors and emotional feelings (Panksepp, 2004). Concomitant evidence corrobored these views in humans, as fMRI of people experiencing sadness, fear, anger, and joy evoke similar brain patterns than those in animals (Damasio et al., 2000).

Third, emotions are thought to be feeling processes that motivate, organize and interact with cognition and action. As Caroll Izard describes it in Izard (2009), we possess emotion schemas, which are defined as dynamic emotion-cognition interactions. For instance, seeing a snake in real life or in a zoo will not have the same emotional consequences in people. These emotion schemas depend on each persons' subjective experience and personality, and usually emerge during development (Izard, 2007; Izard, 2009). More recently, higher-order emotion theories to account for the emergence of conscious feelings, or "emotional states of consciousness" were suggested (LeDoux and Brown, 2017). In this framework, the sub-cortical circuits presented above interact with the (most frontal) brain system responsible for general consciousness, to create an overall conscious emotional experience.

In sum, the three main aspects of emotions are specific physiological reactions, ancient neural systems, shared with other mammals, and their interaction with higher-order cognitive systems such as consciousness, which are influenced by our environment and our personal experiences.

Functions of emotion

But why exactly do we feel emotions? What are the functions of these profound and rapid reactions? The classic perspective on the evolutionary function of emotions is that emotional responses increase the probability of an individuals' survival (e.g. in a danger situation) or reproductive success. Emotions are functional tools as they help individuals to address and overcome

TABLE 1.1: Emotions and their social function. Adapted from Fischer and Manstead, 2008

Affiliation function	**Distancing function**
Happiness	Anger
Love	Hate
Gratitude	Contempt
Admiration	Disgust
Sadness	Social fear
Guilt	Schadenfreude
Shame	Pride about self
Regret	Disappointment in others

problems. As we saw earlier with the example of the bear, it is straightforward to think that people who run from bears have more chances of survival than those passively awaiting to be devoured. In this specific example, emotion is adaptive as the individuals who have the capacity to experience fear when seeing the bear, and run away, are more likely to escape the threat of predators (Cosmides and Tooby, 2000).

A more recent functional view is that emotions also serve a social function. Indeed, humans are a particularly social species. We live in groups, interact with others to gather food, and seek love and group bonding. Research on ostracism reports that social isolation can lead to poor health, well-being, and decreased emotional and cognitive skills (Williams, 2002). As such, emotions are thought to be one of the key tools to maintain these social links. On the one hand, emotions can have an affiliation function as they can help to maintain positive and social relationships or avoid conflict. On the other hand, emotion can have a distancing function as they can serve to establish/maintain a social position in a group (e.g., identity, power), even at the expense of other group members (Fischer and Manstead, 2008). Table 1 presents a selection of emotions and their corresponding social function.

In this line, regulating one's emotional overt reactions in social contexts is crucial. Individuals usually control their emotional displays to e.g. establish or maintain closeness and cooperation with some, or separation and distance with others. Individuals also help others control their own emotions by adaptively controling their emotional reactions depending on the situation, expressing empathy when one's friend is in profound sadness, or portraying

happiness in a positive context, a process also called emotion co-regulation (Bruder et al., 2012).

In sum, emotions seem to serve two basic functions. First, an adaptive function aiming at improving the survival and reproduction of individuals, and second, a social function to regulate social group dynamics.

1.1.2 Facial expression of emotions

Definition

Emotions are the basis of ancient survival mechanisms and are key when regulating group dynamics. But how do we manage to do all these with emotions? The answer is straightforward: our emotion communication tools are remarkable. It is because of the panoply of such tools that we manage to serve the social functions associated with emotions, and accurately communicate signals in critical contexts, such as in the utmost danger situations.

One of the most remarkable tools for emotion communication are facial expressions. Facial expressions of emotions are facial muscle configurations (muscle patterns) which allow us to accurately communicate internal emotional/affective states.

Facial expressions of emotions have been studied for centuries. To put the progress of science in this field into perspective, here is a (beautiful) description of the fright expression by the French painter Charles LeBrun in 1667, which shows how metaphysical concepts were by then usual to describe the strangeness of these biological phenomena. *"The eyebrow, which is lowered on one side and raised on the other, gives the impression that the raised part wants to attach itself to the brain, to protect it from the evil the soul perceives; and the side that is lowered, and which looks swollen, seems to be placed in that position by the phantoms that pour fourth from the brain, as though to shield the soul and protect it from the evil it fears; the wide-open mouth manifests the shock to the heart caused by the blood flowing back to it, which forces it to work harder to draw a breath, which is why the mouth gapes widely open and why, when the breath passes through the larynx and speech organs, it emits an inarticulate sound; for if the muscles and the veins seem swollen, it is only through the action of phantoms the brain sends forth."*

More than 200 years after this text, Charles Darwin published one of biology's most influential books *"Expression Of Emotions in Man and Animals"*,

which exposes the general principles of emotional expressions. In this book, Darwin reports on an inventory of emotional facial expressions, observes that they are used by humans and animals in similar ways, and suggests these expressions have evolved following his previously introduced concept of natural selection (Darwin, 1872).

Darwin's travels, observations and correspondences led him to suggest that humans across cultures have distinct facial expressions for specific emotions, and that these expressions are produced involuntarily as a result of these emotions. Darwin reduced the number of emotional expressions to six "core" expressions: anger, fear, surprise, disgust, happiness and sadness, for which he detailed the motor configurations. Figure 1.1 presents some of the images Darwin based his theory on, which he notably collected from French physiologist Duchenne de Boulogne —from whom we will hear more later in this introduction.

FACS (Facial Action Coding System): measuring facial expressions

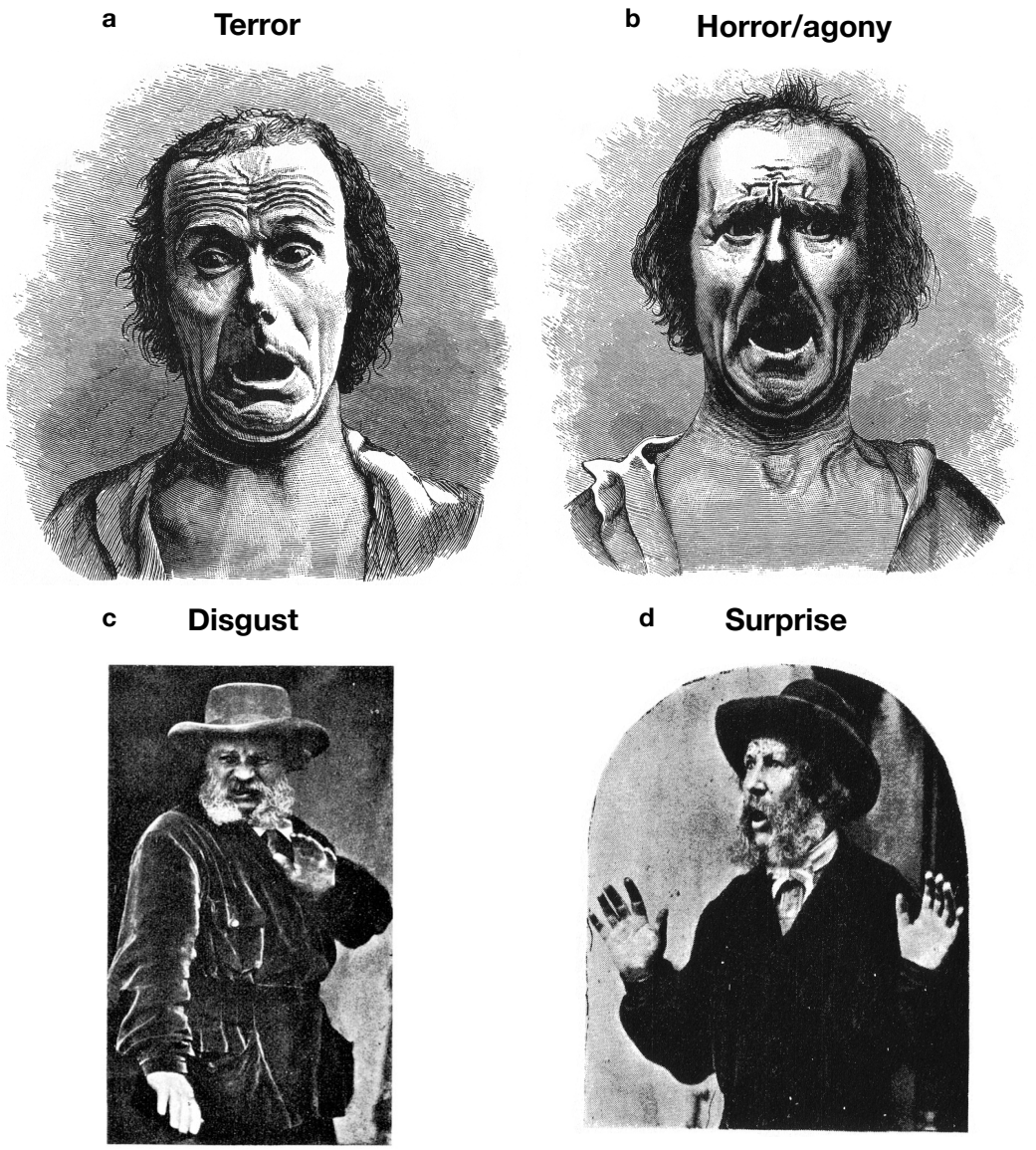
To study facial expressions, researchers categorised the specific facial movements possible with the human facial muscles and developped what is known as Action Units (AUs, Ekman and Friesen, 1978). Using AUs, Ekman and colleagues created the Facial Action Coding System (FACS), a system which describes most of the visible facial AU movements possible with the human face. With this system, the six basic facial expressions of emotion can be described. For instance, the expression of happiness includes the Cheek Raiser (AU6) and Lip Corner Puller (AU12), whereas 'sad' is composed of the Inner Brow Raiser (AU1), Brow Lowerer (AU4) and Lip Corner Depressor (AU15).

The hypothesis of universality

Darwin's idea led researchers to hypothesize the existence of *"affect programs"*, innate, hard-wired, and genetically transmitted mechanisms that link certain evokers to distinguishable and universal displays for each of the primary affects: interest, joy, surprise, fear, anger, distress, disgust-contempt, and shame (Tomkins, 1962a).

The idea of universality was experimentally tested by psychologist Paul Ekman. Ekman went to literate and preliterate cultures such as New Guinea,

FIGURE 1.1: Examples of prototypical facial expressions of emotions of (a) terror (b) horror/agony (c) disgust and (d) surprise. adapted from Darwin 1872.



Borneo, United States, Brazil and Japan, some of which had been in very limited contact with western cultures, to investigate their use of emotional facial expressions (Ekman, Sorenson, and Friesen, 1969). Using forced-choice tasks, Ekman found that happiness, disgust, anger, surprise and sadness were recognised above chance in these cultures, and concluded that these core facial expressions where shared across cultures because of their evolutionary origins, findings which were supported by subsequent studies (Izard, 1994).

Parallel evidence in favor of this hypothesis came from the study of emotion expressions in congenitally blind persons. Although these individuals have never seen facial expressions, there is significant evidence reporting that blind individuals spontaneously produce the same patterns of facial expressions as their sighted conspecifics (for a review see Valente, Theurel, and Gentaz, 2017, for a specific case of a blind and deaf children see Eibl-Eibesfeldt, 1973).

Most of the data supporting claims of universality, however, is collected using similar experimental protocols. These studies usually test whether the average recognition rate of a specific expression is above chance level, using fixed emotion labels, which can bias the measures or hide some specificities of the phenomenon. To solve this issue, researchers have recently developped a data-driven technique to disentangle what specific action unit patterns are shared across cultures (Jack et al., 2016). To do this, researchers first identified a set of highly familiar and typical emotion words both in Chinese and English, and then measured the perceived semantic similarity between these words, deriving a semantic network of emotion words for each culture. Using clustering techniques they identified groups of words that were conceptually related (like happy, glad, pride, cheerful). Afterwards, researchers used a psychophysical technique, reverse correlation, combined with a face generator, to model the dynamic facial expressions associated with these emotion words in each culture. Finally, using Non-negative Matrix Factorization (NMF), a technique allowing investigators to reduce the dimensionality of the data set, researchers showed that there were four basic Action Unit patterns shared between these cultures: Happy, anxious, surprise and disgust, all presented in figure 1.2. Interestingly, the analysis of the specific contribution of each AU in these patterns revealed that the most important movements seem to be the Lip corner puller for the happy expression (AU12; the one used to smile); the Lip Stretcher for the expression of

anxiety; the Jaw Drop for the expression of surprise; and the Lip funneler for the expression of disgust. These action units are the ones that explain the most variance across the data set, suggesting that these are the specific facial configurations which are shared across cultures, and likely have biological origins.

Facial expression of emotions in non-human primates

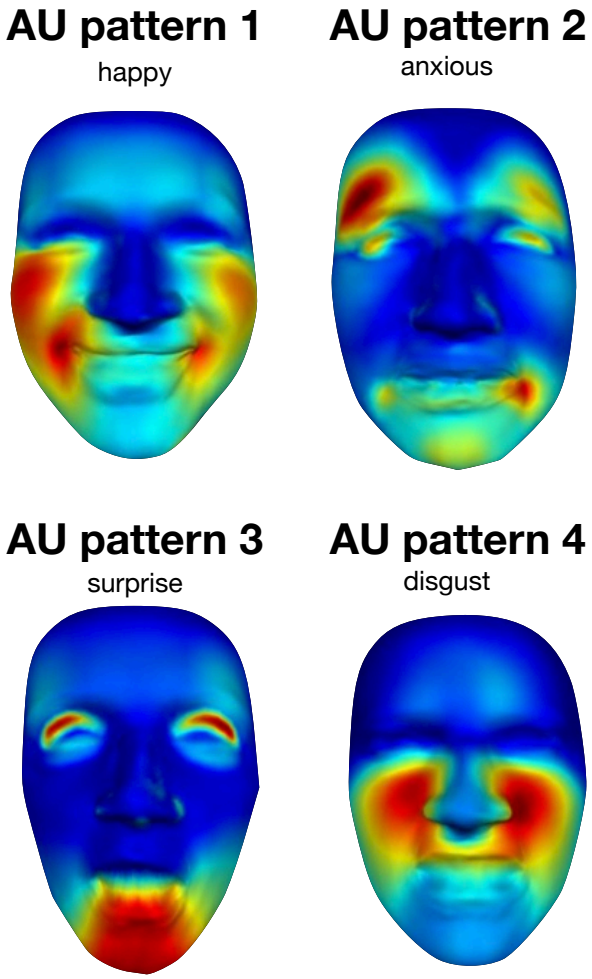
Interestingly, as already noted by Darwin, non-human primates also present distinctive facial displays. Although it is not yet known if these express emotions, researchers have identified four main facial displays, or facial behaviors in macaques: Threat, Silent Bared Teeth, Lipsmack, Relaxed open mouth (also known as the play face; Bliss-Moreau and Moadab, 2017). These displays occur in a large variety of contexts and are used for several social functions. Moreover, there is evidence that primates use these displays differentially depending on the social context, and during interactions with conspecifics (Scheider et al., 2016), e.g. while playing (Chevalier-Skolnikoff, 1974). Moreover, macaques perform above chance when identifying facial expressions regardless of individual identity (Micheletta et al., 2015). Finally, research suggests some aspects of facial perception are evident in primates and a other social mammals (Leopold and Rhodes, 2010), which reinforce the view that facial displays have evolved to suit the needs of complex social communication.

1.1.3 Vocal expression of emotions

The voice : an auditory face

The expression of emotions is not only limited to our facial displays. The voice is an equally relevant tool to communicate affective states. In fact, there is a specific neural structure for the processing of voice in the auditory cortex (Belin et al., 2000), which develops early during infancy (Blasi et al., 2011). This voice structure, called the Temporal Voice Area, is shared between humans and macaques, and responds specifically to conspecifics vocalisations (Petkov et al., 2008). TMS (Transcranial Magnetic Stimulation) in this temporal structure was found to disrupt voice detection, showing that the region has a causal role in voice processing (Bestelmeyer, Belin, and Grosbras,

FIGURE 1.2: Culturally Common Action Unit Patterns in Facial Expressions of Emotion. Color-coded face maps show the four common AU patterns. Red color-coding indicates stronger AU presence (i.e., factor weight) and blue indicates weaker AU presence. AU pattern 1 (the expression of happiness) comprises Lip Corner Puller (AU12) and Cheek Raiser (AU6); AU pattern 2 (the expression of anxiety) included Brow Lowerer (AU4) and Eyes Closed (AU43); AU pattern 3 (the expression of suprise) included Upper Lid Raiser (AU5) and Jaw Drop (AU26); AU pattern 4 (the expression of disgust) involved Nose Wrinkler (AU9) and Upper Lip Raiser (AU10). Adapted from Jack 2016



2011). Importantly, the activity in this voice areas is also sensitive to affective prosody. For instance, a classifier trained with the Temporal Voice Area activity can discern different vocal emotions (Ethofer et al., 2009), even in infants (Blasi et al., 2011). In sum, we possess ancient and highly specialised structures for extracting affective, social and semantic information from conspecifics vocalisations.

Voice's Action Units

But how does the voice communicate emotions? What are its "action units"? The main non-verbal emotion dimensions used by the voice to communicate emotions are pitch contour (intonation), speech rate and timbre.

Among these features, intonation (characterised by the temporal dynamics of the fundamental frequency) is one of the most important. It is used to convey emotions (Banse and Scherer, 1996; Bachorowski and Owren, 1995), but also social attitudes (Ponsot et al., 2018). In fact, algorithmically manipulating an individual's voice in real-time by incrementing their mean pitch and pitch variation (with dynamic inflections) can change the emotional state of the speaker itself (Aucouturier et al., 2016).

Speech rate is also linked to mood, to the extent that there are significant negative correlations between depression tests and speech-rate (Cannizzaro et al., 2004), with slower speech-rate associated with higher levels of depression.

Finally, timbre, studied for instance by the presence of rough cues, whose production by the vocal apparatus is usually due to the saturation of the vocal folds, is an acoustic cue often used as an expression of arousal in humans, but also across species, extending to mammals (Fitch, Neubauer, and Herzel, 2002; Arnal et al., 2015).

Cross-cultural expression of vocal emotions

Just like facial expressions of emotions, vocal communication of affect seems to be mediated both by species-specific and cultural patterns.

For instance, prototypical vocalizations communicating Ekman's basic emotions (anger, disgust, fear, joy, sadness, and surprise) are bidirectionally recognized between individuals from a culturally isolated Namibian villages, in the southwest of Africa, and westerner individuals (Sauter et al., 2010).

We are so adept at extracting information from affective vocalisations that we are able to detect affiliation with a very brief exposure to individuals' co-laughter (Bryant et al., 2016). In Bryant et al. (2018), listeners from 21 societies across six world regions were able to differentiate whether laughter produced by English speakers was fake or real.

Moreover, increasing pitch variation (by algorithmically manipulating its standard deviation with a pitch-shift) was also found to be happier than non transformed voices in English, French and Japenese, in a continuous parametric fashion (the higher the pitch shift/pitch variation, the higher the emotional intensity; Rachman et al., 2017).

1.1.4 Neural processing of emotions

What are the neural bases of emotions? Although this question is broader than the specific topic of this work, it is important to highlight the large networks and areas involved in processing emotion. The conscious experience of emotions, their vocal/facial perception or their regulation all have specific neural pathways.

There are key brain structures for processing emotions in the brain. For instance, the historically-called limbic system (whose main structures are the hypothalamus, the amygdala, the thalamus and the hippocampus), has brain structures that are activated during several emotional processes. The amygdala is known to be higly activated when processing fearful stimuli, be it facial (Pessoa et al., 2005) or vocal (Fecteau et al., 2007). As such, the roughness of a sound, characterised by its fast temporal modulations, and often present in screams, is enough to trigger amygdala activity (Arnal et al., 2015). Moreover, patients with bilateral damage to the amygdala have impaired recognition of fearful facial expressions and musical emotions (Adolphs et al., 1994; Gosselin et al., 2007). and have non-prototypical mental representations of fearful facial expressions (Adolphs et al., 2005). Similarly, there is evidence that humans possess a specific neural substrate for disgust, as the neural response to facial expressions of disgust in others closely relates to

the appraisal of distasteful stimuli (Phillips et al., 1997). I will introduce in more details the neural basis of joy processing, specifically, the neural processing of smiles, and their relation to reward circuits in the last part of this introduction.

It is interesting to highlight that some emotion processes seem to be independent from the stimuli causing it. Emotional musical reactions (such as "chills") correlate with activity in brain regions involved in pleasure and reward, such as food, sex, and drugs of abuse (Blood and Zatorre, 2001). In the same vein, unpleasant (dissonant) music contrasted with pleasant (consonant) music show activations of similar brain areas (amygdala, hippocampus, parahippocampal gyrus; Koelsch et al., 2006). For affective sound processing in general, a recent review suggests there is a broad core neural network for the processing of affect across sound sources. In this network, the amygdala and the auditory cortex play a key role in the decoding of emotional meaning from several sound sources, whereas other regions (inferior frontal cortex, insula, cerebellum) support evaluation and adaptive responses (Frühholz, Trost, and Kotz, 2016).

1.1.5 Section summary

Emotion

I have presented in this section the main characteristics of emotions (physiological reactions, neural systems, and emotion schemas), as well as their adaptive and social functions (affiliation and distanciation). I also presented our two main emotion communication tools: the face and the voice.

By studying the history of the research in "facial expressions of emotions", specifically with Darwin's and James' theories, we saw what facial expressions of emotions are, and how they can be characterized by specific action unit patterns, which are used across cultures when feeling or communicating affective states.

From a neural perspective, I presented the main neural substrate involved in emotion processing, highlighting key brain areas, like the amygdala, which are activated during the perception and the production of affect, and during the visual and auditory perception of emotional stimuli. In the auditory domain, these regions are supported by voice-specific brain structures in the auditory cortex, which activity is enough to decode emotions.

In the next section, we will dive into a more social view of emotion communication by reviewing the mechanisms used when perceiving others' emotions, which can be dissociated from conscious awareness, but still support their core social functions.

1.2 Empathy and facial mimicry

One of the main experimental contributions of this thesis is to report that auditory smiles can trigger the low-level facial reactions usually associated with their visual counterparts. These facial reactions, also called facial mimicry (sometimes emotional mimicry), have important theoretical foundations in the field of empathy. As an introduction to the studies in Chapters 4 and 6, I will briefly review in this section the current models framing the study of empathy, and their relationship to the empirical contributions from the study of mimicry.

1.2.1 Empathy

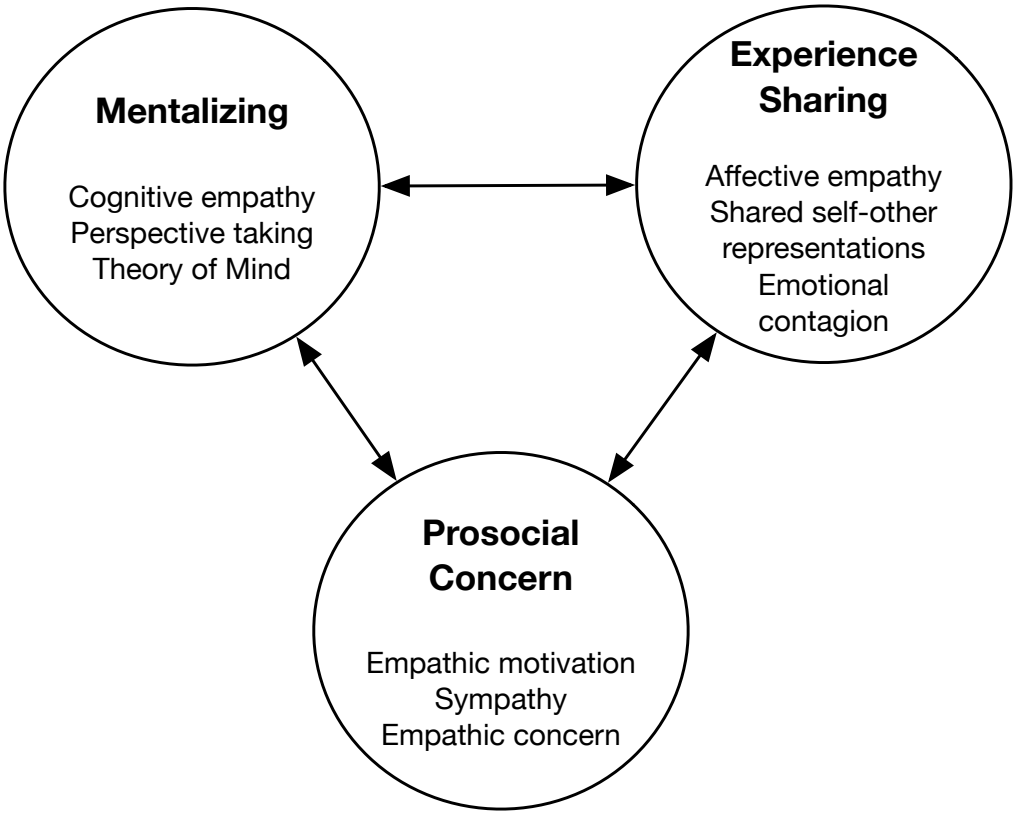
The transition from recognising specific facial expressions to regulating interpersonal affects in social groups is not trivial. It is not because we express affective states with similar facial displays that we can assure the social functions of emotion seen above.

One key component in the regulation of emotions is empathy. Empathy is our capacity to understand the inner (e.g. affective) state of an individual, and to appraise that state. The rapid advance of social neuroscience has suggested interesting frameworks on how to study empathy. In such frameworks, empathy is composed of three main components: Experience sharing, Mentalizing and Prosocial behavior (Figure 1.3; Zaki and Ochsner, 2012).

Experience sharing, often associated to 'neural resonance', is the tendency to engage similar mental systems both when perceiving and experiencing a mental state. Behaviourally, individuals mimic others' bodily postures, facial movements or moods when observing them (Chartrand and Bargh, 1999; Hess and Fischer, 2013; Neumann and Strack, 2000).

Neuroimaging evidence suggests that similar neural structures are activated both when seing and experiencing an action. For instance, when perceiving a disgusting smell and when seeing someone smell something disgusting, people activate overlapping patterns of brain activity (Wicker et al., 2003). In the same manner, experiencing a mild electric shock to induce pain and seeing someone experience that same pain stimuli triggers activity in overlapping brain structures (Singer et al., 2004), namely, the Anterior Insula (AI)

FIGURE 1.3: Three components of empathy. (1) *Experience sharing*, the tendency to take on, resonate with, or 'share' the emotions of others (2) *Mentalizing* the ability to explicitly reason and draw inferences about others' mental states (3) *Prosocial motivation* helps others as a result of using one or both of the other facets to share and/or cognitively understand the emotions they are experiencing. Legend and image adapted from Zaki 2012



and the rostral cingulate cortex (rCC). Such studies suggest a neural substrate for *experience sharing*, highlighting the key role played by AI and rCC (Keysers and Gazzola, 2018). Interestingly, individuals that report experiencing more empathy, show stronger activation in these areas when witnessing others' disgust and pain (Jabbi, Swart, and Keysers, 2007; Singer et al., 2004). Studies report similar effects for positive affects as e.g. both experiencing and observing others experience pleasant tastes triggers activations in the AI (Jabbi, Swart, and Keysers, 2007).

Contrary to a hard-wired, feed-forward wiew of this mechanism, these effects are strongly modulated by social features (Keysers and Gazzola, 2018). Experimentaly manipulating the fairness, in/out-group, responsibility, voluntary empathy, or prior experiences can change the empathic reactions

measured by the activity of these brain structures (Keysers and Gazzola, 2018). For instance, De Vignemont and Singer (2006b) had participants play a prisoner's dilemma game with two research confederates. One of them was fair; the other unfair. Thereafter, the researchers measured AI and rCC activity while participants witnessed both confederates receiving electric shocks. Male participants showed lower activations when witnessing the unfair player receiving shocks, suggesting that empathic neural processes depend on the social link we have with individuals. *Experience sharing* therefore seems to be a self-motivated process, which can be deployed differently depending on the target person and the individual's past experiences.

The second key component of empathy is mentalizing. Mentalizing is our ability to do inferences about others' mental states, including affective states, but also cognitive perspectives. For instance, mentalizing is involved when inferring, from a facial expression how another person is feeling, but also what the aim of a PhD is when reviewing a series of broad fields.

The last component of empathy is *prosocial behavior* (also termed prosocial motivation), through which individuals who share and understand targets' mental states are compeled to help those targets or react in specific ways (Zaki and Ochsner, 2016).

In sum, our capacity for empathy helps to co-regulate emotions, and to achieve their social functions, via its three main components *Experience sharing*, *Mentalizing* and *Prosocial behavior*. In the following we focus mainly on the first two components (*Experience sharing*, *Mentalizing*).

1.2.2 Imitation as social glue

Mimicry

Of particular theoretical importance for our study of auditory smiles is the component of *shared experience* in empathy, which I will illustrate here with the empirical contributions of the literature on mimicry. As seen in the previous section, one key component of empathy is our tendency to engage similar mental systems both when perceiving and experiencing states, in other words, *"it's the tendency of perceivers to take on the sensorimotor, visceral, and affetive states of targets"* as put by Zaki and Ochsner (2016). Substantial evidence demonstrates that people imitate several behaviors when observing others. For instance, few days old infants cry in response to another infant

cry (Simner, 1971). Similarly, we have all experienced uncontrollable laughing or yawning after hearing such behaviors in conspecifics (Provine, 2004; Warren et al., 2006; Yoon and Tennie, 2010). This often spontaneous process of imitation is called mimicry and is seen across a wide range of human behaviors.

I would like to share two personal experiences to illustrate these behaviors. I am originally from Colombia, in South America, and thus my mother tongue is Spanish. During my PhD I had the chance to go to Japan to present some of my results (Chapter 4). When interacting with people from Japan, they all asked me the same question: -"Are you French?" -"No, why?" I responded. I realised I had a French accent when speaking in English. Being exposed for 10 years to the English accent of French speakers had the consequence of my English phonemes, dynamics, and intonations sounding as those of my French colleagues. Similarly (perhaps more impressively), I once visited the main square of Santa Elena, Tolima, a village in Colombia in the middle of the Andean Mountains isolated from civilisation by the civil war, where my father and his family grew up, but where no member of my family had been in the last two decades. While walking through the main square I heard two women in a balcony say to each other: "Look at that guy there, he walks like an Arias". I was in shock, how could they know who my parents were only by seeing me walk? These women recognised my way of walking although my family had not been in the region for years.

These are not mere anecdotes. Empirical data from the seventies and eighties confirm that people imitate each others' dialects, accents, speech rate, vocal intensity and syntax (Giles, 1973; Street Jr, 1984; Natale, 1975; Levelt and Kelter, 1982). In a similar way, the more two people have lived together, the more they are judged to behave similarly (Zajonc et al., 1987) and during interactions, we also imitate the body gestures of our partners, such as their foot tapping, face touching and mannerisms (Cheng and Chartrand, 2003).

Closer to the interest of this dissertation, there is also evidence of mimicry of psychological states. For instance, participants adopt the moods of the voices they hear (Neumann and Strack, 2000). In a recent study Aucouturier et al. (2016), showed that covertly manipulating participants' voice with a specific emotion shifts participants' emotional states towards that same emotion.

Interestingly, these mimicry reactions do not happen automatically, and are mediated by an important social component. For instance, real mother-baby interactions are judged as more synchronous by external observers (Bernieri,

Reznick, and Rosenthal, 1988), and foot tapping imitation can be used as an affiliatory strategy (Cheng and Chartrand, 2003). Chartrand and Bargh (1999) manipulated mimicry in the context of dyadic interactions. Participants were either mimicked or not mimicked by a research confederate during a joint task. Participants that were mimicked more strongly liked the research confederate and perceived their interactions as smoother than unmimicked participants. More recently, pupil mimicry was also found to promote trust through similar *theory of mind* mechnisms (Prochazkova et al., 2018).

In sum, there is large empirical data to support the view that imitative behaviors serve a social function. Experimental psychologists think of mimicry as social glue, as it is experimentally linked to rapport, empathy, affiliation and prosocial behavior. Theories suggest that our capacity to mimic is an adaptive behavior which serves an important evolutionary function (Chartrand and Dalton, 2009). We mimic some, and not others, depending on the social context, our ambitions, and the targets, often with affiliative goals in mind.

Imitation in infancy

This capacity to spontaneously imitate others' states is thought to be essential to our development. Some developmental psychologists even describe humans as Homo Imitans (Meltzoff, 1988), suggesting that without imitation, we would not be able to develop critical capacities such as language, social skills and even empathy.

Andrew Meltzoff arrived to this conclusion after studying newborn imitations. One of his most spectacular studies tested newborns while they were still at the maternity ward (the youngest participant was 42 minutes old). The experiment consisted in looking at newborn babies and either produce a tongue protrusion or an open-mouth gesture. The results showed that newborn infants can imitate both facial displays (Meltzoff and Moore, 1983). Subsequent studies highlighted that imitation helps babies to learn how to use objects (Meltzoff, 1985), to produce speech sounds (Kuhl and Meltzoff, 1996; Kuhl and Meltzoff, 1982) and, generally to learn from others, accelerating drastically an infant's capacity to remember motor skills (Meltzoff, 1985).

As a side note, there has been a recent spirited debate on whether or not infants have the innate capacity to imitate. On the one hand Oostenbroek et al. (2016) claim that infants do not have the capacity to imitate from birth, whereas Meltzoff et al. (2018) support their claims of innate imitation (see

also Oostenbroek et al., 2018). In the scope of this dissertation, taking both sides into consideration, we can confidently assume that there is enough evidence to support that infants imitate at least when being a few weeks old (Oostenbroek et al., 2016).

Mirror neurons

I can not talk about imitation, empathy and speech perception, without at least mentioning another controversial discovery: Mirror Neurons. Mirror neurons are a type of neuron discovered by an italian research group in Parma, lead by neurophysiologist Giacomo Rizzolatti in 1992, which have strongly influenced the views on how the brain processes the actions of others (Di Pellegrino et al., 1992).

Mirror neurons were first discovered using single unit recordings (recording individual cells) in the ventral premotor cortex of the macaque monkey (the equivalent of Broca's area in the human brain) and were described as neurons that fire both when the monkey sees and performs an action (Gallese et al., 1996). More recently, they have also been shown to be activated when the monkey hears the sound of the action (Keysers et al., 2003). Such "cell behavior" is remarkable, and fitted perfectly well in several theories in which motor activity was thought to be essential to understand action, such as, for instance, the motor theory of speech perception (more on this later).

In the 2000s, several fMRI studies in humans mapped all brain regions that had voxels that were activated both while executing and witnessing the actions of others (Keysers and Gazzola, 2018). For instance, Gazzola and Keysers (2008), asked participants both to execute and to observe goal-directed hand actions while measuring their brain, and concluded that the observation and execution of actions share motor and somatosensory voxels.

However, there is no way to prove with fMRI that humans have mirror neurons. The fact that there is overlapping voxel activity both during action observation and action execution is indeed compatible with the hypothesis that there are voxels containing mirror neurons, but is not a direct proof of their presence. Indeed, a voxel contains millions of neurons and can contain both neurons that respond to the observation and to the execution of an action.

It is thus difficult to prove the existence of mirror neurons in humans. Ethical limitations do not allow researchers to implant electrodes in people brains as

they do with macaques. For this reason, it's only until 2010 that mirror neuron activity was recorded in human brains. In Keysers and Gazzola (2010), researchers recorded 21 patients who were being treated for epilepsy, and who had already intracranial depth electrodes implanted for clinical purposes; In such patients, the researchers found a small number of neurons that showed mirror characteristics.

The problem with mirror neurons, as explained by Gregory Hickock in his book *"The Myth of Mirror Neurons"*, is that after their discovery they generated a lot of attention because of their perfect fit with Meltzoff's imitation theories, but also speech theories. People began to speculate that these mirror neurons were responsible for empathy, or even imitation, while at the time there was no proof of their existence in humans. Mirror neurons were suggested to be the basis of imitation, even though monkeys (the ones in whom these were first discovered) did not have an imitation system as developed as humans. Similar arguments are developed in the field of speech perception. In other words, the existence of mirror neurons is not enough to explain neither language development, nor imitation, and if they are indeed involved in such processes, there has to be a panoply of adjacent mechanisms supporting them (Hickok, 2014).

In sum, fMRI studies have shown that several motor structures show mirror-like responses in humans, and that humans do possess mirror neurons. But, to my knowledge, it is not known whether these "mirror-like" responses are due to actual mirror neurons. In the same vein, although phenomena as imitation, mimicry, or empathy have all some sort of "mirror-like" component, there is no evidence yet that these mechanisms are mediated by mirror neurons in the human brain (Hickok, 2014).

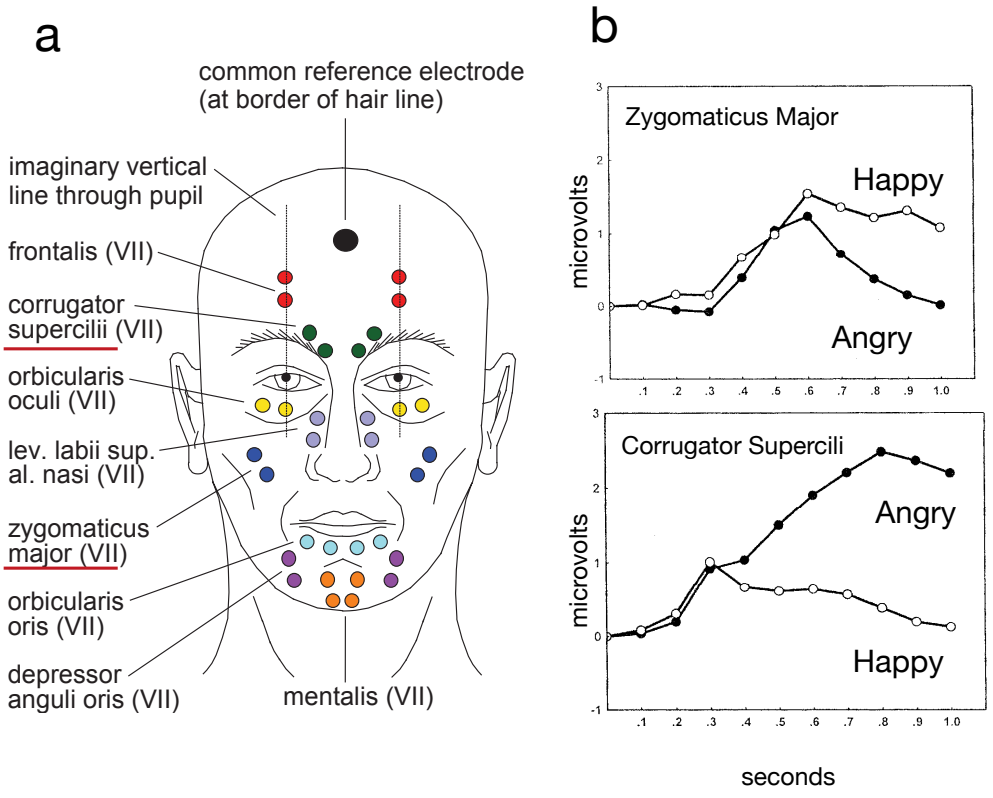
1.2.3 Emotional mimicry: a methodological tool

Definition and examples

One of the most studied imitative behaviors to emotional facial expressions is emotional mimicry. Emotional mimicry is defined as the tendency to imitate the emotional facial expressions of others (Hess and Fischer, 2013), and is, as such, closely related to empathy. This phenomenon has been studied for decades, as it is thought to be a privileged tool to study how we process emotions in others (Hatfield, Cacioppo, and Rapson, 1993).

Emotional mimicry is often measured by means of facial electromyography (EMG), which is an electrophysiological measure of muscle activity, e.g. in the zygomaticus major, used to smile, or the corrugator supercili, used to frown (Figure 1.4; Van Boxtel, 2010). Even if the facial reactions can be very small and difficult to see with naked eyes, surface electrodes are often enough to measure them. Figure 1.4-a shows examples of electrode placement on the human face, and figure 1.4-b presents prototypical zygomatic and corrugator reactions to angry and happy facial expressions. Note that zygomatic and corrugator muscles often have antagonic reactions. The zygomatic tends to increase for positive expressions which involve smiling, whereas the corrugator tends to decrease for these same expressions.

FIGURE 1.4: Facial reactions measured with facial EMG (a) EMG electrode placement examples (adapted from Boxtel 2010) (b) prototypical facial reactions to emotional facial expressions (adapted from Dimberg 2002)



The role of emotional mimicry: cause or consequence?

There are two main theories for the role of emotional mimicry. First, embodied cognition theories suggest that simulating others' emotions is a way to access their affective states. For such theories, we access a target's mental states via an inner simulation, which in turn is used to infer what the target affective state is (Niedenthal et al., 2010). Second, the facial feedback hypothesis suggests that the sole production of a facial expression can change one's emotions (Strack, Martin, and Stepper, 1988; but see also Wagenmakers et al., 2016). And thus, emotional mimicry would be a way to share the emotions of others. Although experimental psychologists have put a lot of effort to know if emotional mimicry follows any of these two roles, findings are still inconclusive and often contradictory.

Studies trying to understand the causal implication of mimicry have developed ways of blocking facial reactions. In Oberman, Winkielman, and Ramachandran (2007), researchers asked participants to bit a pen or chew gum, impairing to some extent, the recognition of emotional expressions. In the same vein, in Neal and Chartrand (2011) researchers manipulated participants' ability to imitate facial expressions either using Botox injections (which reduces muscular feedback from the face) or a dermall filler (an intervention that does not reduce feedback). Supporting embodied theories, dampening and amplifying facial feedback signals respectively impaired and improved people's ability to read others' facial emotions.

Conversely, in Blairy, Herrera, and Hess (1999), successful mimicry manipulation did not affect the accuracy of emotion recognition. And it is clear that muscle activity is not needed for the recognition of emotions, as even individuals with moebius syndrom (an extremely rare congenital neurological disorder which is characterized by facial paralysis) can recognize emotions, even though they are not able to imitate facial expressions (Rives Bogart and Matsumoto, 2010). The implication of mimicry in emotion recognition is not straightforward.

Alternative hypotheses suggest that mimicry does not change the recognition rate of emotional facial expressions, but rather facilitates it, which is translated by diminished reaction times during tasks. In Stel and Knippenberg (2008), women were slower to recognize the affective valence of briefly displayed facial expressions when prevented from mimicking them. Similary, in Niedenthal et al. (2001), researchers measured the speed at which

participants could identify a change in emotion expression. During the experiment, participants watched dynamically changing facial expresions going from happiness to sadness and vice versa while their facial expressions were either blocked (by using a pen in the mouth, the TMS of the field) or not blocked from mimicking. Participants were faster in detecting the change in emotional expression when they were allowed to mimic.

In sum, the role of mimicry is still not clear, and the field needs better and more controlled experimental methods. There is no convincing evidence that emotional mimicry is causally involved in the recognition of emotions, although it may be involved in interactions with processes underlying emotion recognition.

Emotional mimicry: dependency on the task and social context

Although the exact role of emotional mimicry has not been clearly identified yet, these mimetic reactions are known to be modulated by several factors. First, emotional mimicry is strongly modulated by the presence of an affective task (Hess, Philippot, and Blairy, 1998). For instance, in Murata et al. (2016), researchers used emotional tasks as well as other non-emotional tasks to probe the extent to which mimicry is modulated by the active extraction of emotional information from targets. The authors show that facial mimicry is more elicited when partcipants have to explicitely extract emotional cues. These findings have been thoroughly corroborated in the literature (Hess and Fischer, 2016). Second, emotional mimicry is modulated by social context. Facial reactions are modulated by social cues, like social group membership (Bourgeois and Hess, 2008).

Emotional facial mimicry to sound

Emotional mimicry is not only ellicited with facial stimuli. There is a limited number of studies studying cross-channel mimicry: and notably facial reactions to emotional vocal expressions. To my knowledge, the first published study of this phenomenon was Hietanen, Surakka, and Linnankoski (1998), where researchers showed, that hearing an anger vocalisation can increase corrugator activity —although no zygomatic reactions were found. In Hawk, Fischer, and Van Kleef (2012), crying and laughing sounds ellicited congruent facial reactions. In the same line, in Verona et al. (2004) participants

responded with more zygomatic muscle activity to pleasant sounds (baby laugh, female erotic moan, and crowd cheer) than to unpleasant sounds (baby cry, female attack sound, male attack sound), and more corrugator muscle activity to unpleasant sounds than to pleasant sounds.

Unconscious, spontaneous or automatic mimicry?

Two studies in the literature of emotional mimicry shed light on an interesting characteristic of such facial reactions, namely, their link with consciousness and perception.

First, Dimberg, Thunberg, and Elmehed (2000), showed that emotional facial mimicry can be elicited even when presenting stimuli outside of conscious awareness. The researchers used a backward-masking technique to present emotional faces in a subliminal fashion and showed that facial reactions could be triggered even when the stimuli was not consciously perceived.

In the same line, Tamietto et al. (2009), presented emotional facial expressions to two patients who had unilateral destruction of the visual cortex. Impressively, although participants could not perceive the emotional facial expressions because of their visual impairment, their muscles reacted congruently with the facial expressions presented.

These two experiments highlight the fact that conscious recognition or perception of the stimuli is not needed to trigger emotional mimicry, and have contributed to the recent description of neural mechanisms used in the unconscious processing of emotional signals (Tamietto and De Gelder, 2010).

1.2.4 Section summary

Empathy and facial mimicry

I have presented in this section how empathy allows humans to co-regulate emotions. I presented empathy as our capacity to understand the inner (e.g. affective) state of an individual, and to appraise that state. Empathy has three key components: experience sharing, mentalizing and prosocial behavior (Zaki and Ochsner, 2012). The *experience sharing* component of empathy, is supported by specific brain regions, such as the Anterior Insula, that responds both to seeing and observing actions, and which activity correlates with empathy ratings and is modulated by social variables such as in/out-group.

I then presented behaviors related to such systems. Imitation, which is present very early in humans, has been shown to be essential to child development, increasing cognitive capacities and accelerating motor learning. Such imitatory behavior can also be seen in verbal humans, in which mimicry serves as social glue, increasing rapport and affiliation during social interactions.

I finished this section by describing the framework of emotional mimicry, a research paradigm studying facial reactions to emotional facial expressions. Emotional mimicry represents a well framed and well defined motor reaction elicited during the processing of emotion expressions, even when stimuli are presented outside of conscious awareness. This paradigm will be used in chapter 4 to probe the emotional processing of auditory smiles.

1.3 Auditory processing of articulation

As presented in the general outline of this introduction, the aim of this dissertation is to study how facial expressions of emotions, specifically smiles, can be perceived through the voice. To this aim, I will review in this section how humans produce and interpret sound. I will review the main models for speech production and perception, highlighting its physiology and the corresponding acoustic properties. Moreover, because facial expressions of emotions are motor gestures, I will detail with particular care the theories on how motor areas are recruited by vocal perception processes.

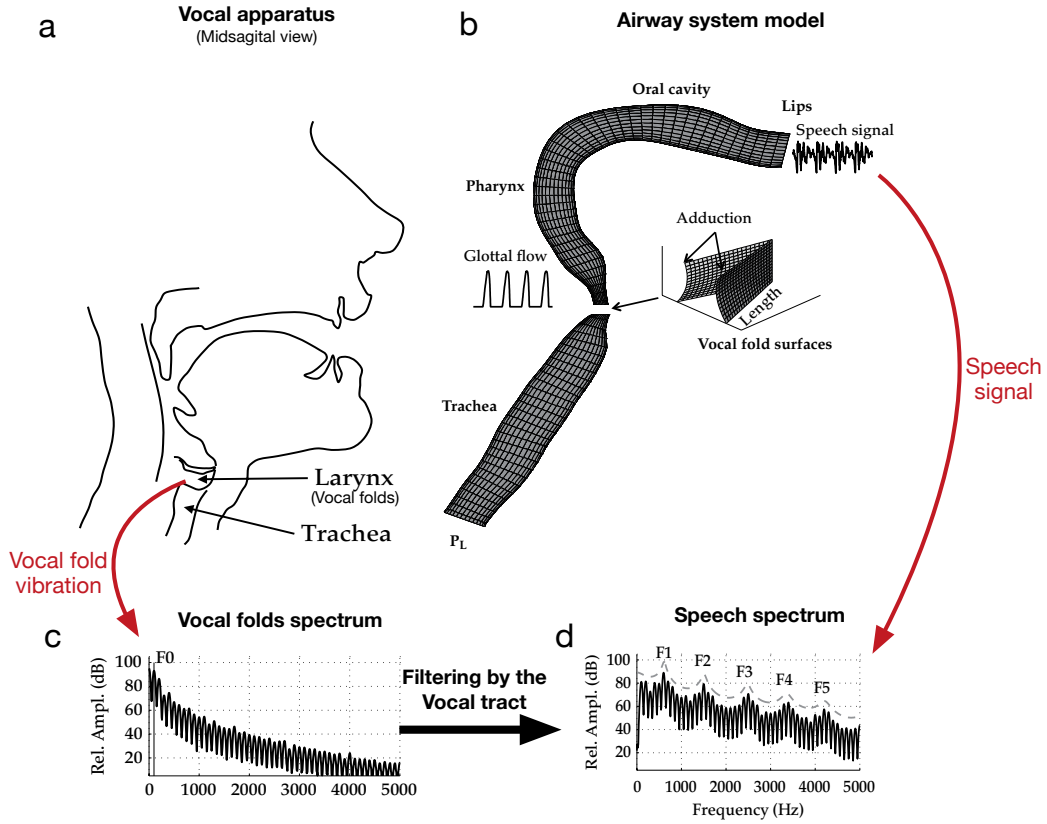
1.3.1 Producing and retrieving articulation

Voice production: vocal apparatus

To understand how speech is perceived, it is important to understand how the motor control of the vocal apparatus generates phonetic information, and how this information, in the form of acoustic features, is represented and encoded in the brain.

The vocal apparatus can be modeled as a source-filter model. The source is the sound generator: the vocal folds (in the case of harmonic sounds). The vocal folds convert air flow coming from the lungs into a series of flow pulses by periodically opening and closing the air space in the larynx, which provides excitation to the vocal tract. The filter is the component that shapes the spectral content of the sound coming from the source (e.g. the spectral shape of a specific phoneme). The filter is composed of the pharynx, the oral cavity and the mouth, which form change dynamically over time. Figure 1.5-a presents the vocal-apparatus with the vocal folds (source) and the filter (vocal tract). Figure 1.5-b presents the schematics of the airway system, or how the pressure is transformed into sound, going from the glotal flow all the way to the speech signal. Figures 1.5-c-d present the spectral content of the source and speech (filtered source) respectively. (Figure adapted from Story, 2015).

FIGURE 1.5: Vocal apparatus (a) Midsagittal view of the upper portion of the trachea, larynx, and vocal tract (b) Schematic diagram of the airway system for speech production. The vocal folds, which are shown magnified, produce the glottal flow signal shown in the left. The flow pulses excite the vocal tract and produce the output (radiated) pressure waveform shown at the lips. (c) Glottal flow spectrum (d) Output spectrum. Adapted from Story 2015



Formants

Note in figure 1.5-d the frequency peaks after the filtering by the lips. These resonances are called formants, and are numbered F1 to F5. Note that formant frequencies are different from the fundamental frequency (Figure 1.5-c), which is called f0 (pitch of the voice).

Formants can be seen in a spectral envelope by looking at its spectral peaks. They are the resonant frequencies of the oral cavity, which are determined by its shape, length and area. As the length and shape of the cavity change with articulation the formant frequencies also change. As such, formant frequencies are enough to reconstruct intelligible speech, as they convey the shape of the vocal tract during production (De Boer and Kuhl, 2003).

Formants are the most important phonetic feature in the human voice. The first two formants are enough to disambiguate vowels. Figure 1.6-a presents the main English vowels in relation to F1 and F2 frequencies. This plot is called the vowel space. Figures 1.6-b-c present examples of how the vowel space changes depending on whether speech is hypo or hyperarticulated (note how the vowel space density increases when speech is hyperarticulated; Story and Bunton, 2017). In a developmental perspective, other studies suggest that, mothers'vowel space density is significantly correlated with both infants' and machines' speech discrimination performance (Liu, Kuhl, and Tsao, 2003; De Boer and Kuhl, 2003).

Because the vocal apparatus is very similar across animal species, animal vocalisations often have formant frequencies (Fitch, 1997). In addition, because formant frequencies are a direct consequence of the resonant properties of vocal tracts, the length of the vocal tract can be estimated acoustically by measuring its formant frequencies (Stevens, 2000). This way, it is possible to infer, from formant frequencies, the length of the vocal tract, and to a certain extent, the vocalizer's body size. For example, formant dispersion correlates with body size in rhesus macaque monkeys (Fitch, 1997), and provides a reliable cue to body size in American alligators (Reber et al., 2017). In a recent study considering 91 mammalian species, the Dominant Frequency of the vocalisation (defined as the frequency of maximum amplitude in the spectrum), was found to inversely correlate with body size (Bowling et al., 2017). This empirical evidence suggests that animal-call's spectrum serves an adaptive function, which is mediated by formant frequencies, and may have played a critical role in the evolution of vocal communication (Fitch, 2000), namely, to communicate the sender's strength and formidability, and avoid both the sender and the receiver the harm of unnecessary conflict.

The importance of formants can also be seen in the human auditory cortex. In a now modern-classic study (Mesgarani et al., 2014), researchers recorded the cortical activity from six human subjects implanted with a multi-electrode array as part of their clinical evaluation for epilepsy surgery. Researchers asked subjects to listen to 500 sentences, spoken by 400 different people. A data-driven approach combining speech phonetics, acoustic features, and electrocochleography (ECOG) recordings revealed how phonetic representations are encoded in the human Superior Temporal Gyrus (STG). The optimal projection of vowels in formant space was the difference F2 minus F1 (first principal component, dashed line, Figure 1.6-c-d). In addition, the sensitivity

to F1 and F2 was negatively correlated across all vowel-selective sites (Figure 1.6-d), suggesting that single STG sites show an integrated response to F1 and F2. The researchers also found high decoding accuracies of f0, F1, F2, F3 and F4 (Figure 1.6-e), which suggests fundamental and formant variability is encoded in neural populations in the STG. Finally, Multi-Dimensional Scaling (MDS) analysis revealed that the relational organization between vowel centroids in the acoustic domain is preserved in neural space (Figure 1.6-f). This data suggests that F1 and F2 frequencies are encoded jointly in the human auditory cortex and that there is a multidimensional feature space encoding the acoustic parameters of speech sounds, where formants, defined by distinct articulations, are the strongest determinants of selectivity.

In sum, formants are important acoustic features encoded early in the human auditory cortex, essential for phonetic communication and used across the animal kingdom as a cue of body size.

1.3.2 The role of the motor system in speech perception

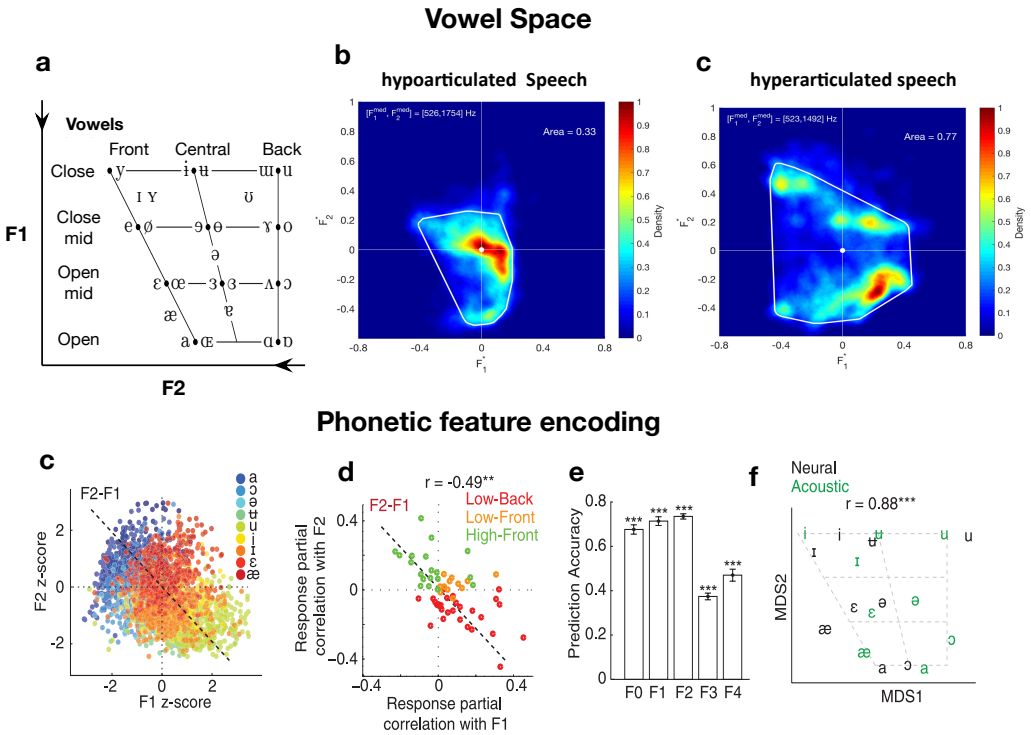
Dual stream model of speech perception

As seen previously, phonetic features (formants) representing both articulation and vocal tract shape are encoded in the auditory cortex. What are the subsequent neural pathways involved in the processing of such acoustic features? How do motor areas support this processing?

Hickok and Poeppel (2007) propose a dual stream model of speech perception in which the motor system is importantly implicated. This model proposes two functionally and anatomically distinct neural pathways to process speech and language information. These two pathways are thought to be highly specialised, and are somehow analogous to the "what?" and "where?" pathways which separate perception and action in vision (Goodale and Milner, 1992). The general overview of this two-stream model is presented in figure 1.7.

The first pathway is the ventral stream, which processes speech for comprehension and is thought to interface sensory and phonological networks with conceptual-semantic systems. This stream involves structures in the superior and middle portions of the temporal lobe (e.g. in the STG presented above).

FIGURE 1.6: (a-c) Linking speech performance to the Vowel Space Density (VSD). (c-f) Formant encoding in the human STG. (a) Schematic view of the vowel space as a function of formants and lip posture (b-c) plots of a male talker producing several minutes of (b) hypoarticulated speech and (c) hyperarticulated speech. (d) Formant frequencies, F1 and F2, for English vowels (F2-F1, dashed line, first principal component) (e) F1 and F2 partial correlations for each electrode's response. Dots (electrodes) are color-coded by their cluster membership. (f) Neural population decoding of fundamental and formant frequencies. Error bars indicate SEM. (g) Multidimensional scaling (MDS) of acoustic and neural space. (a) adapted from the international phonetic alphabet 2018 (b-c) Adapted from Story 2017; (c-g) adapted from Mesgarani 2014



The second pathway, which is more relevant in the context of this dissertation, is the dorsal stream. The dorsal stream is thought to map acoustic speech signals to frontal lobe articulatory networks. In this sense, the dorsal stream, which is left-hemisphere dominant, interfaces the sensory and phonological networks with motor-articulatory systems. The dorsal stream involves neural structures in the frontal lobe and the motor cortex and is involved in translating acoustic speech signals into articulatory representations, which is essential for speech development and normal speech production. In short, the dorsal stream has an auditory–motor integration function.

In the following I will detail the underlying processes of speech perception recruiting motor areas.

The motor theory of speech perception

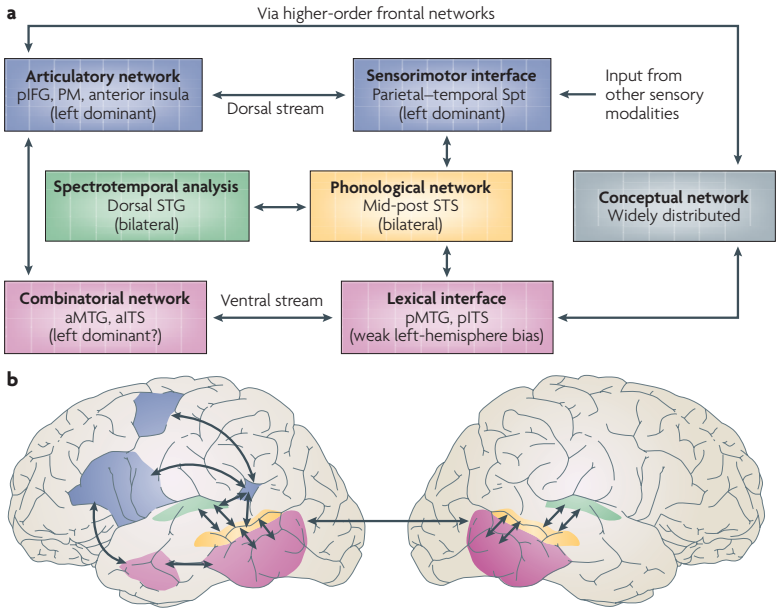
The dual stream model of speech perception suggests motor areas in the dorsal stream are recruited during speech perception. But, when hearing speech, what exactly do we perceive? Do we perceive the gesture, the acoustics, or both? In other words, is the brain representing the acoustic information (e.g in the form of formants) or the gestures that generated the speech sound? (e.g. in the form of motor patterns). These questions have been at the core of speech perception research since the *Motor Theory of Speech Perception* (Liberman et al., 1967).

The simplified hypothesis of Liberman's theory posits that speech recognition is the process to identify the articulatory gestures that are used to produce the speech signal. The motor theory of speech perception takes the form of the following three claims: (Galantucci, Fowler, and Turvey, 2006)

- (1) speech processing is special,
- (2) perceiving speech is perceiving gestures, and
- (3) the motor system is recruited for perceiving speech.

Today, speech perception research has advanced largely. Although the main claims of the motor theory of speech perception turned out not to be completely correct (see Galantucci, Fowler, and Turvey, 2006 for a review), some of its key ideas have received considerable experimental support, mainly inspired from claims 2 and 3. In the following section I review the literature

FIGURE 1.7: The dual stream model of speech perception (a) Earliest stages of cortical speech processing involves spectrotemporal analysis. Phonological-level processing involves the middle to posterior portions of the superior temporal sulcus (STS) bilaterally. Subsequently, the system diverges into two streams, a dorsal pathway (blue) that maps sensory or phonological representations onto articulatory motor representations, and a ventral pathway (pink) that maps sensory or phonological representations onto lexical conceptual representations. (b) Approximate anatomical locations of the dual-stream model components. Green regions depict areas on the dorsal surface of the superior temporal gyrus (STG) that are proposed to be involved in spectrotemporal analysis. Yellow regions in the posterior half of the STS are implicated in phonological-level processes. Pink regions represent the ventral stream. Blue regions represent the dorsal stream. The posterior region of the dorsal stream corresponds to an area in the Sylvian fissure (area Spt), which is proposed to be a sensorimotor interface. The more anterior locations in the frontal lobe correspond to portions of the articulatory network. aITS, anterior inferior temporal sulcus; aMTG, anterior middle temporal gyrus; pIFG, posterior inferior frontal gyrus; PM, premotor cortex. Legend and figure adapted from Hickock and Poeppel 2007



supporting claim 3, and more generally, recent neuroscientific findings underpining the roles of the motor system in speech (and auditory) processing.

Behavioral evidence for motor system recruitment in speech perception

First, the development of the *"motor theory of speech perception"* began when Liberman and colleagues realised that the same phoneme was not pronounced exactly in the same way when preceded by two different phonemes. This phenomenon, called coarticulation, is the fact that when speaking we temporally overlap the phonetic gestures of speech. In the original experiment researchers found that second-formant transitions can signal [d] in the syllables [di] and [du] whereas they can signal [p] and [k] before other vowels (Cooper et al., 1952). This led researchers to suggest that what is perceived when hearing speech is in fact the gesture, since exactly the same sound can be interpreted differently depending on the articulatory context.

The second line of evidence comes from audiovisual speech integration, specifically from what is called the McGurk effect. The McGurk effect is an auditory illusion of the speech signal depending on the co-ocurring visual cues. In the original experiment called *"Hearing lips and seeing voices"* (McGurk and MacDonald, 1976), participants were asked to look at videos of a person vocalising the syllable [ba], superimposed with the video of that same person saying [ga]. Participants reported hearing [da]. This effect highlights how visual articulatory information can influence the (once-thought only auditory) process of speech recognition and how gestures interact with the auditory representations, even in 5-month-old infants (Rosenblum, Schmuckler, and Johnson, 1997). This effect is known to depend on the temporal alignment between visual and auditory flows, suggesting that articulatory information is perceived simultaneously using both visual and auditory information (Munhall et al., 1996).

The third line of arguments for the implication of the motor system in speech perception comes from audiovisual speech intelligibility studies. Research has shown that visual articulatory cues improve the intelligibility of speech in noise due to the contribution of visual information to the extraction of acoustic cues (Schwartz, Berthommier, and Savariaux, 2004). More impressively, evidence also suggests that visual cues derived from the dynamic movements of the face during speech production interact with time-aligned auditory cues to enhance sensitivity in auditory detection. (Grant and Seitz,

2000). In this experiment, the influence of visual cues depended on the auditory-visual temporal coherence, which was measured by the degree of correlation between acoustic envelopes and visible movement of the articulators, while a control condition recreating the same movements in another object (not a face) did not show such effect. This highlights how it's not only the dynamic information, but also the auditory-visual temporal coherence between face movements and sound, which are responsible for the sensitivity enhancement.

In sum, these examples show how speech processing can be modulated by articulatory cues, whether they are heard (in the case of the first co-articulation example), or seen (examples two and three). These examples shed light on how articulatory cues, whether or not purely visual, can modulate speech perception.

Motor activity when processing phonemes

Let's now dive deeper into the dorsal pathway to understand why and how motor areas are recruited during speech processing, and how this activity supports articulation-related processes.

The primary motor cortex is involved in phoneme perception. In Pulvermüller et al. (2006) subjects were asked to produce and listen to phonemes produced with the tongue ([t]) or with the lips ([p]) while their brain activity was recorded with fMRI. Researchers report that distinct motor regions in the precentral gyrus (primary motor cortex) are differentially activated in a somatotopic manner: motor activity distinguished between lip-related and tongue-related phoneme perception. This study highlights how motor circuits can support the articulatory features of speech during perception.

Other studies using TMS have found similar evidence. In Watkins, Strafella, and Paus (2003), TMS was applied to the face area in the primary motor cortex to elicit motor-evoked potentials in the lip muscles (motor evoked potential are rapid muscular reaction measured with EMG to repetitive-pulse stimulation of the brain). In the study, the size of the motor-evoked potentials was compared between four conditions. Two involving speech (hearing and seeing speech), and two control conditions (viewing eye and brow movements or listening to non-verbal sounds). In both speech conditions (hearing and seeing speech), the size of the motor evoked potentials were enhanced compared to control conditions. Similarly, in D'Ausilio et al. (2009), researchers

report that TMS applied to the lip area can improve task performance for lip-related phonemes such as [b] and [p] but not tongue-related phonemes such as [d] and [t], and vice-versa. In the same vein, TMS to the lip representation in the motor cortex can impair subjects' ability to discriminate between synthetic speech sounds that are lip-articulated (e.g. [ba] vs. [da]) but not sounds that are not lip-articulated (e.g. [ka] vs. [ga]) (Möttönen, Dutton, and Watkins, 2012). These studies highlight that listening or seeing speech can enhance the excitability of the motor units underlying speech production, and how these can be selectively disrupted with TMS.

Although the data presented above suggests a strong recruitment of the primary motor and premotor cortex in speech perception, even in the absence of a specific task, clinical data suggests these interpretations have to be nuanced. In their review *What happens to the motor theory of perception when the motor system is damaged?* (Stasenko, Garcea, and Mahon, 2013), the authors present a large amount of studies reporting that patients who have lesions in motor areas can have profound impairments in speech production, but spared speech recognition. For instance, patients who have chronic stroke, which severely impairs speech production, can still perceive speech sounds (Hickok, Houde, and Rong, 2011) and individuals with Broca's aphasia, typical for language impairment, are indistinguishable from control subjects on auditory word comprehension tasks, although Broca's aphasia often involves damage to the motor cortex (Moineau, Dronkers, and Bates, 2005; Hickok et al., 2011).

Recent research projects using original methods are still trying to understand the role of the motor system in speech perception. COSMO (Communicating Objects using SensoryMotor Operations) is an integrative model based on Bayesian programming, which allowed researchers not only to compare motor or auditory implementations of speech perception, but also to test the gain of efficiency provided by the Bayesian fusion of auditory and motor information (Laurent et al., 2017). In other words, this model simulates how speech is learned, during development, using bayesian probabilities, and can be used only with auditory information, only motor or with their fusion. The ongoing results seem to indicate that (1) auditory recognition is more efficient than motor recognition when dealing with learned stimuli, while motor recognition is more efficient in adverse conditions, and (2) auditory cues may be more efficient for vowel decoding and motor cues for plosive articulation decoding (Laurent et al., 2017). This specialisation of the motor system for

phonetic coding is an interesting possibility, and is in line with other findings (Pulvermüller et al., 2006).

In sum, the motor theory of speech perception is probably not true in its broadest sense. Deactivation of the motor system doesn't seem to impair speech recognition —at least when measuring with simple speech discrimination tasks. Nevertheless, TMS and fMRI findings reporting motor related activity during speech perception are numerous and compelling, some models suggest motor information is more efficient in adverse conditions, or for certain types of sounds (e.g. plosives; for detailed reviews on these topics see Hickok, Houde, and Rong, 2011; Stasenko, Garcea, and Mahon, 2013; Scott, 2016)

An action/perception link in the motor cortex

As seen in the previous section, the activity in the motor cortex in response to articulatory cues does not seem to trivially help perception. Its role may thus be orthogonal to the direct perception of speech, but still triggered by the acoustic stimuli. One intriguing possibility to explain part of the motor related activity during speech perception is that motor responses, instead of being an active part of perceptive processes, serve to control/modulate motor/behavioural actions in response to those sounds.

In an fMRI study on the contagious aspects of auditory yawning, researchers found that activity in the posterior inferior frontal gyrus (pIFG; a part of the articulatory network) not only was greater to yawns as compared to other non-contagious sounds, but also that the pIFG activity was higher for trials in which the participant explicitely reported a desire to yawn in response to the yawning sound, as compared to trials in which the participant didn't report like yawning after hearing a yawn, and to trials in which the participant did report like yawning but that was preceded by a control sound (Arnott, Singhal, and Goodale, 2009).

Similarly, a study on the contagious aspects of laughter highlights how the passive perception of a purely auditory laughter can modulate neural activity in a network of premotor cortical regions involved in the control of facial movement. The experimental design, which combines fMRI and facial EMG, demonstrates how facial mucle reactions are not simply an overt laugh but are rather motor response patterns engaging preparation for responsive orofacial gestures (Warren et al., 2006).

In the same line, several studies highlight the overlap in motor and premotor areas during speech perception and production both during silent repetition of bisilabic pseudowords (Buchsbaum, Hickok, and Humphries, 2001), during the production of meaningless monosyllables (Wilson et al., 2004), or when comparing heard and produced real-life speech (Glanz et al., 2018).

These results, are very similar to the ones I presented earlier on the neural substrate of empathy (section 1.2.1). The dual stream model states that the posterior inferior frontal gyrus, the premotor cortex, and the anterior insula, all regions associated with empathic behavior, are part of an articulatory network (Hickok and Poeppel, 2007).

Roles of Supplementary Motor Areas in Auditory Processing

The supplementary motor area (SMA), another key structure of the motor cortex, also plays an important role in sound perception. Although SMA has been primarily investigated in relation to its motor functions, SMA is also consistently reported in fMRI studies on auditory processing. In a recent review Lima, Krishnan, and Scott (2016), report that auditory stimuli, both heard and imagined, activate SMA and pre-SMA. Importantly, SMA and pre-SMA are not recruited only for vocal sounds, but for a wide variety of sound sources, including musical sounds. Rhythmical patterns seem to be important to explain SMA activity, as e.g. SMA is more active for rhythmic sequences compared to random sequences. Because of the wide variety of SMA responses to sound, and its non specialization to speech/articulatory cues, its activity does not seem to fall in the scope of this dissertation. Nevertheless, the large amount of studies on SMA activity confirm the important involment of the motor system in sound processing (Lima, Krishnan, and Scott, 2016).

1.3.3 Section summary

Auditory processing of articulation

In sum, the auditory processing of articulation is a multi-headed phenomenon. Its production is based on the vocal tract which creates specific spectral resonances (formants) to transmit articulatory information in the form of speech. These spectral structures are transmitted from the cochlea to the auditory cortex where they are encoded in the STG. Part of their information is subsequently transmitted via the dorsal stream to a large panoply of motor and articulatory areas, which do not seem to serve a direct role in perception per se but may underlie action and response preparation, for instance, in the form of orolabial gestures.

1.4 Voice and face: audiovisual integration

Humans and non-human animals are constantly confronted to external events which have to be processed optimally to assure the individual's survival (in the case of a threat), or social well-being (in the case of a social stimuli). Such information is often transmitted via several sensory modalities. In order to minimize the uncertainty of ambiguous stimuli, the optimal way to process these sensory inputs is to process them in parallel and jointly, in the case of co-occurring auditory and visual stimuli —this information coupling is termed *audiovisual integration* (Gerdes, Wieser, and Alpers, 2014).

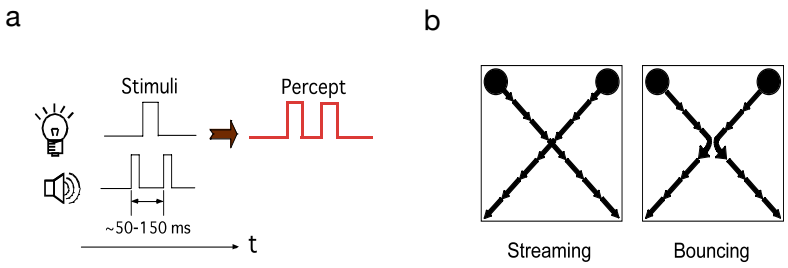
In this dissertation, the cognitive integration of visual and auditory smiles will be studied in chapter 5 using an eye-tracking experimental paradigm. As such, I am going to review in this section the literature of audiovisual integration, specifically, how emotion information from voice and face are integrated into a shared amodal emotional concept representing the information from both sensory inputs. I will detail with particular care recent eye-tracking and pupillometry findings on this topic as an introduction to the study in chapter 5.

1.4.1 Low-level audiovisual integration

Thousands of research papers have been published on multisensory and audiovisual integration, ranging from very low-level perception processes, to very high level (e.g. emotional) interactions. Although the lower-level aspects of audiovisual integrations are not in the scope of this dissertation, their influence in the field as well as general methodologies are important, so I will review very briefly key findings of this literature.

Among the big wave of articles published at the begining of the 2000s on audiovisual interactions, one of the most influential reports is provocatively entitled *What you see is what you hear* (Shams, Kamitani, and Shimojo, 2000). In this experiment, researchers report a perceptual visual illusion induced by sound: when a single visual flash is accompanied by multiple auditory beeps, the single flash is incorrectly perceived as multiple flashes (Figure 1.8-a). This illusion sheds evidence on how low-level visual tasks (perceiving flashes) can be modulated by low-level auditory stimuli (beeps). In the same vein, studies have also found that sound can alter visual emotion perception. Specifically, Sekuler (1997) showed that the perceived trajectory of two identical visual

FIGURE 1.8: (a) The sound-induced flash illusion reported by Shams et al. 2000 When a brief visual stimulus is accompanied by two brief sounds it is often perceived as two flashes. (b) The stream-bounce illusion reported by Sekuler et al. 1997. Two identical visual objects approach and move away from each other on a screen. In the absence of sound the two objects are often perceived as streaming through each other. However, when a brief sound is presented around the time of visual coincidence of the two objects, the probability of perceiving a bouncing motion is increased; Figure adapted from Shams et al. 2010



objects moving towards each other can either be perceived as if they stream or bounce at each other when arriving to the crossing point, depending on a co-occurring sound (Figure 1.8-b).

These studies show how auditory information can interfere with the processing of otherwise unambiguous visual information. They also highlight the importance of the temporal dynamics in audiovisual integration.

1.4.2 Audiovisual speech integration

Another line of evidence reporting integration of audio and visual streams during perception is audiovisual speech. Interestingly, this line of research reports often the opposite trend as the evidence reported above; namely, the visual information triggers auditory enhancement/misdetections.

There are several examples of such misdetections in the literature. For instance, speech can be mislocated from its source depending on visual information (Driver, 1996), and phonemes can be misperceived depending on concurrent and non-congruent visual articulation (McGurk effect, see also 1.3.2). Importantly, as for low-level audiovisual interactions, the McGurk effect is known to depend on the temporal alignment between visual and auditory streams (Munhall et al., 1996).

Visual information in speech can also enhance the decoding of auditory information. For example, visual articulatory information is enough to allow an individual to perceive non-native phonemic contrasts in a non-foreign language (Navarra and Soto-Faraco, 2007). In the same line, seeing speech articulation can also increase speech intelligibility. In such experiments, the more ambiguous the auditory cues (e.g., if they are embedded in noise), the more visual information enhances intelligibility (Sumby and Pollack, 1954). Again, these effects strongly depend on the temporal concordance of the sensory information, as higher correlation between the acoustic envelope and the articulators' visible movements implies higher effect size (Grant and Seitz, 2000).

Electrophysiological data sheds light on the temporal integration of congruent/incongruent articulatory cues. In Kaganovich, Schumaker, and Rowland (2016), participants first heard a word referring to a common object (e.g., pumpkin) and then decided whether the subsequent visual silent articulation matched the word they had just heard. Incongruent articulations elicited a significantly enhanced N400; congruent articulations elicited significantly larger Late positive components. Importantly, researchers found a significant correlation between the amplitude of the N400 and individuals' improvement on Speech-In-Noise Task (SIN: measure of improvement in speech intelligibility when adding visual information to auditory speech). This ERP component is known to reflect information mismatch, here the articulatory audiovisual gesture; which shows that speech units are processed sub lexically as fast as 400 ms post stimulus, during a co-evolving information stream.

1.4.3 Audiovisual (face-voice) emotion integration

Better and faster: behavioral effects when processing audiovisual emotions

As communication in general, emotion expressions rely on different modalities, and the joint processing of visual and auditory emotional information is essential to better understand e.g. a person's intentions, and react in consequence.

Not surprisingly, studies report that emotion classification is improved when individuals have access to bimodal cues, as opposed to unimodal (Kreifelts et al., 2007). Among these emotional dimensions, prosody, semantics and face

articulation are known to additively improve emotion recognition in the case of congruent emotional cues (Paulmann and Pell, 2011). In a similar fashion, incongruent emotion cues (prosody and face) are known to ambiguate an emotion concept, both for static and dynamic faces (Baart and Vroomen, 2018). Participants are not only more precise to recognise emotions in congruent audiovisual situations but are also faster (Föcker, Gondan, and Röder, 2011).

Such bimodal emotional integration also seems to be to some extent automatic. For instance, De Gelder and Vroomen (2000) show that not only the identication of the emotion in the face is biased in the direction of the simultaneously presented tone of voice, but more importantly, that this effect occurs even when participants are instructed to base their judgement exclusively on the face. In that study, the reverse effect was also demonstrated : when presenting an audiovisual stimuli and asking subjects to rate only the auditory information, their response is biased towards the visual emotion. These results suggest the existence of a pseudo-automatic bidirectional link between affect detection in vision and audition.

Such automaticity is known not to be limited to the attentional resources portrayed to the stimuli. Vroomen, Driver, and De Gelder (2001), asked subjects to judge whether a voice expressed happiness or fear, while trying to ignore a concurrently presented static facial expression. In addition, subjects had to add two numbers together rapidly (experiment 1), count the occurrences of a target digit (experiment 2), or judge the pitch of a tone as high or low (experiment 3). Face had an impact on voice emotion judgments in all the experiments, independently of the cognitive load induced by the different concurrent tasks.

During emotion categorisation tasks, the emotional information of both sensory inputs is integrated to create an overall emotional concept. In Collignon et al. (2008) researchers used dynamic emotional facial expressions accompanied by non-linguistic vocal affect expressions to investigate how several intensities of audiovisual emotional expressions are combined together. In the incongruent situation, participants preferentially categorised the affective expression based on the visual modality, demonstrating a visual dominance in emotional processing. However, when the emotion information in the visual stimuli was ambiguous, participants categorised incongruent bimodal stimuli preferentially via the auditory modality.

Taken together, these findings suggest that the audiovisual integration of emotional cues helps both to improve detection rates, and to be faster. These interactions seem to be, to some extent, automatic, as explicit instructions to ignore one modality does not prevent the integration from happening.

Physiology: pupillometry as an index of audioviusal emotion processes

A more implicit measure than the pure behavioural effects seen previously are physiological measures. The physiological changes induced by the processes underlying audiovisual emotion recognition are numerous. In this dissertation I am going to focus on a recent line of interesting evidence reporting significant shifts in both pupil dilation and face exploration strategies (eye gaze) when perceiving congruent/incongruent audiovisual emotions.

Pupillometry, the dynamic measure of pupil's size, is relevant because of its automatic nature. The underpinings of pupil dilation are mediated by the sympathetic nervous system. Under isoluminance conditions, pupil dilation is almost exclusively promoted by norepinephrine release from the locus coeruleus (Wel and Steenbergen, 2018). As such, it is very difficult to consciously control it —other than by non-direct methods such as thinking of a very bright/dark situation or doing complex calculations.

One line of evidence suggests pupillometry indexes cognitive load. In a recent review (Wel and Steenbergen, 2018), researchers report that pupil dilates across a large amount of tasks depending on the amount of cognitive load. These variations of pupil size seem to be triggered by high cognitive demands whether the task depends on inhibiting, shifting or updating mental representations (Wel and Steenbergen, 2018): The higher the cognitive load, the higher the pupil dilation induced by the task.

However, in the emotion literature, pupil size has also been used as indexing arousal. In an influential study, Bradley et al. (2008) monitored participants' pupil size during emotional and neutral picture presentations. Pupillary changes were larger when viewing emotionally arousing pictures, regardless of whether these were pleasant or unpleasant. These changes covaried with skin conductance change, supporting the interpretation that sympathetic nervous system activity modulates these changes in function of the affective context. Of note, these pupillary changes in reaction to affective stimuli do not depend on the conscious perception of the stimuli. In an impressive study, Tamietto et al. (2009) recorded pupillary changes during the

presentation of emotional faces to two participants with unilateral destruction of occipital visual cortex and ensuing phenomenal blindness over one half of their visual fields (i.e., these patients could see with one eye but not with the other, althoug their retina was in good conditions for both eyes). Researchers presented facial and body expressions to both eyes to probe the reactions to "seen" and "unseen" affect expressions. Impressively, both "seen" and "unseen" conditions triggered similar affective reactions as measured by pupil dilation. The researchers conclude that these affective reactions may be mediated by evolutionary ancient visual affective pathways which bypass cortical vision while still transmitting and triggering reactions to emotional information.

Researchers have also studied the development of these automatic emotion reactions during infancy (Jessen, Altvater-Mackensen, and Grossmann, 2016). In this study, researchers probed whether or not pupil reactions to emotional facial expressions are present even when faces are presented subliminally. Results were similar than for "unseen" trials in blind-sight participants, happy facial expressions triggered larger pupil dilation as compared to fearful, regardless of conscious perception. Accordingly, this autonomic mechanism in response to affective stimuli seems to be, to some extent, independent of the conscious perception of the affective stimuli and develop early during infancy.

Another line of evidence linking autonomic arousal (here pupil size) to emotion processing is the findings on audiovisual emotion integration. Pupil dilation has been consistently reported as indexing congruency in bimodal stimuli. For instance, when perceiving incongruent audiovisual stimuli like a dog meow or a cat bark, pupil dilates more than for the congruent stimuli (Renner and Włodarczak, 2017). Such effects are no different in the case of incongruent emotions. In Hepach and Westermann (2013), 10- and 14-month-old infants' were shown video clips in which happy or angry actors performed either a positive or a negative action. 14-month-old infants, but not 10-month-old, showed selectively greater pupil dilation during the incongruent scenarios when compared to the congruent scenarios.

How then to reconcile the evidence reporting that pupil dilation indexes cognitive load, arousal, and the processing of bimodal cues? One possibility is that these reactions are mediated by different mechanisms. Indeed, the tasks used to link pupil dilation to cognitive load or to affective processing are significanly different, and involve very different cognitive demands.

Another intriguing possibility is that pupil dilates more when processing emotionaly incongruent stimulus due to the higher cognitive demands subholded by its processing. Accordingly, pupil dilation was found to reflect the underlying decision processes involved in emotion recognition (Oliva and Anikin, 2018). In this study participants heard human nonverbal emotional vocalizations and rated the emotional state of the target as soon as posible. The results showed that during emotion recognition, the time course of pupil response was driven by decision-related processes, and that pupil responses revealed properties of the listeners' choices, such as the perceived emotional valence and their confidence levels.

Eye gaze strategies during audioviusal emotion processing

Another interesting line of work explores how audiovisual emotion integration influences face exploration strategies.

For example, Rigoulot and Pell (2012) asked participants to look at a series of pictures. Each picture was accompanied by an emotional non-sense sentence pronounced with a specific emotional prosody. Results show that when the emotion and the voice match in their emotional direction, the gaze is implicitly guided towards the face.

Similarly, Paulmann, Titone, and Pell (2012) evaluated whether emotional prosodic cues in speech have a rapid and mandatory influence on eye movements to an emotionally matched face. During each trial, participants viewed six emotional faces while listening to instructions spoken in an emotionally congruent or incongruent prosody. Results showed that participants fixated more frequently the faces with emotionally congruent audiovisual content.

Interestingly, similar patterns of eye gaze activity seem to be already present in preverbal infants (Jessen, Altvater-Mackensen, and Grossmann, 2016). In this study, 7-month-old infants showed a significantly longer fixation duration time for happy compared to fearful facial expressions.

Finally, although perhaps coincidentally, gaze activity seems to be disrupted in certain pathologies known for their impaired emotion processing. For instance, individuals with Autism Spectrum Disorder (ASD) show unusual patterns of face exploration, be it during passive tasks or during explicit emotion extraction tasks (Pelphrey et al., 2002). Results from eye-tracking studies

objectively demonstrate abnormalities in the processing of social information in ASD (Zilbovicius et al., 2013; Wang et al., 2015).

1.4.4 Neural resources responsible for voice-face integration

As seen previously, voice-face integration can be observed at the behavioral (e.g., audiovisual emotion processing, visual speech perception), physiological (e.g eye tracking and pupillometry) or electrophysiological level (e.g. N400). These data suggest that visual and auditory information from voice and face merge together creating shared amodal representations. In the case of emotional information, this can even happen in a pseudo-automatic way (De Gelder and Vroomen, 2000). In the following I'm going to succinctly review the anatomical candidates responsible for the underpining of these integration processes.

The amygdala processes emotion independent from the sensory input

One first key candidate mediating the integration of voice-face emotional information is the Amygdala. The amygdala receives input from all senses, and is known to process emotional signals independently from their sensory modality. For instance, the amygdala is known to be highly activated when processing facial (Pessoa et al., 2005) or vocal (Fecteau et al., 2007; Arnal et al., 2015) fearful stimuli. One relevant study reported how the bimodal presentation of a fearful voice facilitates recognition of fearful facial expressions. Results highlight how the amygdala and fusiform gyrus response to fearful faces are modulated by the concurrent emotion in the voice, specifically, whether or not the voice shares the face emotional tone. (Dolan, Morris, and Gelder, 2001)

The pSTS a region involved in voice-face integration

The literature reports another key structure when dealing with voice-face integration, the pSTS (posterior Superior Temporal Suclus). In a recent fMRI study on identity perception from voice and face (Hasan et al., 2016b), activity in the pSTS was enough to classify identity from both faces and voices. More in the focus of this dissertation, the activity in the pSTS is also related to the emotional congruence between auditory and visual stimuli (Watson

et al., 2014). In that study, fMRI signal in the right pSTS showed an activity reduction in response to congruent emotional stimuli. Moreover, another study controled whether the activity in the pSTS in response to emotional mismatch is not just a confound of task difficulty (incongruent stimuli would be harder to categorize than congruent stimuli). Researchers found that activation in the superior temporal region in response to incongruent emotional information could not be explained by task difficulty, but rather by the detection of an emotional mismatch in the sensory inputs (Watson et al., 2013). In a complementary fashion, other lines of research suggest that the pSTS activity is enough disambiguate articulation from emotional facial expressions, as its activity can decode specific facial Action Units (Srinivasan, Golomb, and Martinez, 2016).

Speech perception research also reports that the STS is involved in audio-visual speech comprehension. TMS in the STS interferes with the McGurk effect (Beauchamp, Nath, and Pasalar, 2010) and in Nath and Beauchamp (2012), activity in the STS was significantly correlated with the likelihood of perceiving the McGurk effect.

1.4.5 Section summary

Voice and face: audiovisual integration

Voice and face signals are integrated at different levels. First, visual information is used during speech perception, increasing speech intelligibility, or helping to discriminate phonemes. Second, audio and visual emotion information are also processed jointly from voice and face as e.g. audiovisual emotional stimuli is recognised better and faster when the two channels are merged. These interactions are to some extent automatic, as explicit instructions to rate only one modality are not enough to disambiguate the information from the sensory inputs.

The physiological evidence reporting audiovisual emotion integration are numerous. Among these, a new line of evidence highlights both pupil size, and eye gaze changes during these processes. Pupil dilation is linked to the processing of (in)congruent audiovisual emotions, as it dilates when individuals are processing incongruent emotional stimuli. Similarly, face exploration strategies, change depending on audiovisual emotional congruency. Congruent voice-face emotions redirect gaze towards the eyes, increasing both the number of fixations and fixation duration time. From a neural perspective, two key structures are thought to be implicated in the processes underlying these sensory integrations: The amygdala, which processes emotional signals independently of the sensory input, and the pSTS, which can e.g. decode specific AUs.

Chapter 5 will study how visual and auditory smiles interact during perception.

1.5 The particular case of the smile

Smiling, the bilateral contraction of the zygomatics major muscles, lies at the intersection of the fields of emotion research, empathy, motor-speech interaction, and audiovisual integration reviewed above. It is one of the universal action-unit patterns used across cultures to express emotions; it triggers facial mimicry in observers; and smiling while speaking has recognizeable acoustic consequences on vocal timbre.

This last part of the introduction reviews the specificities of the smile. I will describe its physiology and neural substrate, how it develops, as well as the theoretical links between its form, function and origin. Finally, I will describe the current state of the research on the acoustic consequences of smiles in speech, will introduce the term *auditory smiles*, and present the main research questions addressed during this PhD.

1.5.1 Form and function

Physiology

Smiles are one of the four basic action patterns used across cultures to communicate emotions (Ekman, Sorenson, and Friesen, 1969; Jack et al., 2016) and, as such, have interested scientists for centuries (Darwin, 1872).

One of the first scientists to be interested in smiles was Duchenne De Boulogne (1806-1875). His most relevant book on emotional facial expressions entitled *The mechanism of human facial expression* (1862) presents an original procedure to isolate specific facial muscle movements. Duchenne used in-skin electrodes to stimulate facial muscles with electrical current and trigger their contraction. With this method, he characterized the visual consequences of contracting individual facial mucles, and in particular to depict the "emotional role" of certain muscles. Figure 1.9-a shows one of his iconic participants (*"the simple old man"*, who suffered from a loss of sensation in the face and thus was the ideal participant for Duchenne's experiment. Figure 1.9-b, presents the facial expression when applying current to one of the zygomaticus major muscles —notice the unilateral lip stretch on the subject and the contralateral neutral face. Figure 1.9-c presents the bilateral stimulation of zygomaticus major muscles.

One of the main observations of Duchenne, was that the bilateral stimulation of zygomaticus major was not enough to trigger a prototypical joy expression. For Duchenne, the sole stimulation of zygomatics triggered what he qualified as a *"false smile"*; he found that, for his subjects to communicate a genuine expression of joy, he had to also stimulate the Orbicularis oculi eye muscle (Figure 1.9-d): *"the emotion of frank joy [is] expressed on the face by the combined contraction of muscle zygomaticus major and the inferior part of muscle Orbicularis oculi. The first obey the will, but the second (the muscle of kindness, of love, and of agreeable impressions) is only put in play by the sweet emotion of the soul. Finally, fake joy, the deceitful laugh, cannot provoke the contraction of this latter muscle"*.

This was the first distinction between different forms of smiles, illustrating how a smile's communicative value can be modulated by small but specific facial patterns. The presence of Orbicularis oculi activity during smiling is so-called the Duchenne marker (Ekman and Friesen, 1982); several classic studies have demonstrated that this feature is present when a smile occurs spontaneously during states of happiness and is lacking when a smile is displayed deliberately, especially in attempts to mask negative feelings (Ekman and Friesen, 1982).

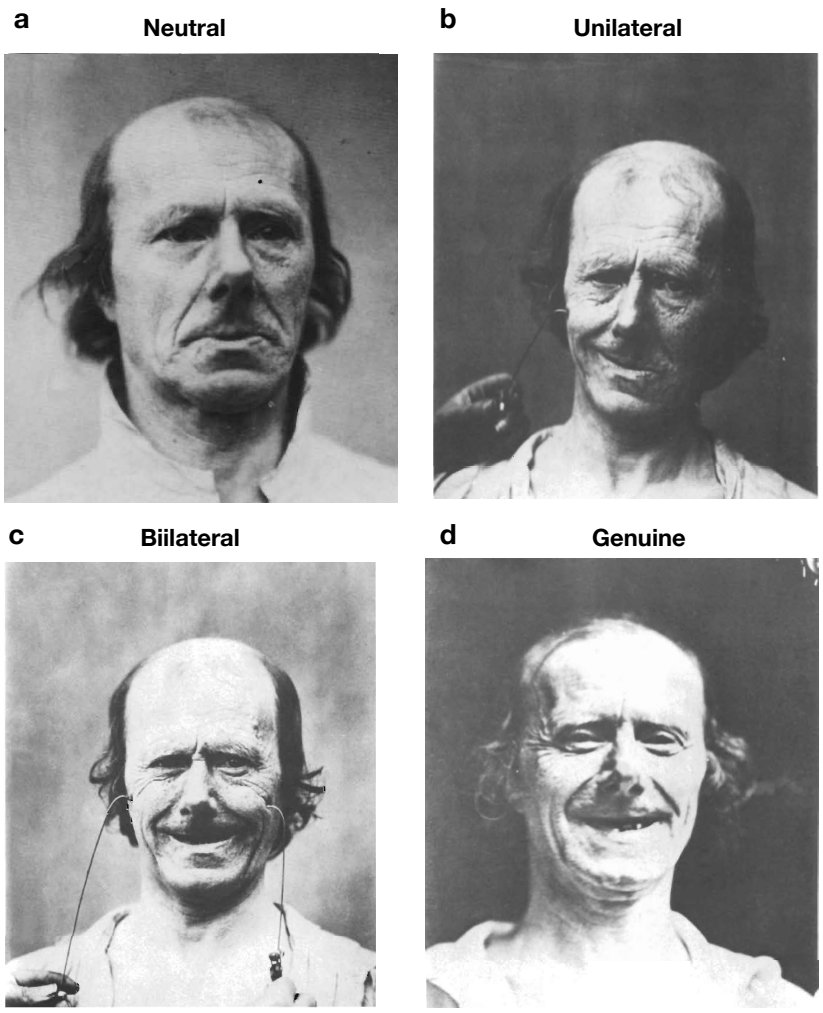
The facial configuration that Duchenne found conveys the *"frank expression of joy"* uses exactly the same facial muscles (AU12 + AU6) as the happiness action-unit pattern illustrated in figure 1.2, and is today still considered one of the four action-unit patterns universally used across cultures (Jack et al., 2016).

Function : an affiliative social tool

Smiles are one of the most important behaviours of social affiliation. A wealth of studies from very different fields highlight how smiles are used in social interactions to create and reinforce social bonds.

For instance, smiling people are more likely to be ascribed positive traits such as kindness, humor, intelligence, or honesty than their non smiling pairs (Hess, Beaupré, and Cheung, 2002; Reis et al., 1990). The more we believe that a stranger's smile reflects sincere and positive feelings, the more cooperative we are likely to be towards that person (Krumhuber et al., 2007). Smiling faces are also easier to remember (Cromheeke and Mueller, 2016), judged

FIGURE 1.9: (a) Neutral expression of *"the simple old man"*, who suffered from facial anaesthesia and was thus an ideal subject for Duchenne's experiments. (b) On the right side of the face, electrical excitation of the zygomaticus major, showing development of the fundamental and secondary lines of this muscle: (false joy or laughter); on the left side, a relaxed face (c) Slightly stronger electrical excitation of both zygomaticus major muscles: development of the same fundamental and secondary expressive lines of joy, with mild contraction of some fibers of the muscle called the sphincter of the eyelids (false laughter) (d) The same subject as in a and b, showing a natural laughter expression constituted by the association of zygomaticus major and the inferior part of orbicularis oculi. Adapted from Duchenne, *The mechanism of human facial expression*, 1990



as more attractive (Golle, Mast, and Lobmaier, 2014) and appear more familiar (Baudouin et al., 2000).

A recent study even suggests that scientists that smile in their social profile pictures are more cited. Using 440 research profiles from ResearchGate the researchers looked at the relationship between the smile intensity, portrayed by researchers' profile pictures, and their citations number. Smile intensity was statistically related to the number of citations and the number of followers. Although the causality of the relation is undetermined (researchers may also be smiling more because of the citations), smiling does seem to portray a status of professional achievement (Kaczmarek et al., 2018).

Although perhaps anecdotal, the ubiquity of smiles can also be seen in text interactions. Indeed, the tears of joy emoji 😂 was selected as *Word of the Year* in 2015 by Oxford Dictionaries. This pictogram was the most-used emoji, and among the most-used expressions, in the world in 2015. This shows how much we need smiles during social interactions to better understand and co-regulate affective states.

Because of the important role played by smiles in social interactions to shape social bonds, it is not surprising that smiling also constitutes a much-researched part of the behavioral repertoire of virtual embodied agents (El Haddad et al., 2016; Yu, Garrod, and Schyns, 2012). Smiles are routinely integrated in human-computer interactive systems, where their role is again to facilitate communication and improve outcome. Avatars with smiling faces are judged more attractive and positive (Oh et al., 2016) and, like smiling humans, trigger physiological reactions in human observers (Partala and Surakka, 2004; KräMer et al., 2013). More than a feature that can be turned on and off, avatar smiles can be synthesized gradually (Ku et al., 2005) and with temporal dynamics (Ochs, Pelachaud, and Mckeown, 2017), allowing researchers to experiment with how and when an avatar should smile to improve the quality of a virtual interaction. Avatar smiles were found to have a positive impact on ongoing interactions (Yee, Bailenson, and Rickertsen, 2007) and on its later outcomes, including better learning (Meij, Meij, and Harmsen, 2015; Maldonado et al., 2005; Kim, Thayne, and Wei, 2016) and problem solving (Partala and Surakka, 2004).

Smile as a multi-purpose signal

Smiles, however, do not only serve an affiliative purpose. As seen earlier with Duchenne's work, subtle changes to the prototypical smiling motor configuration can signal different social contents, making smiles a multi-purpose social signal (Martin et al., 2017).

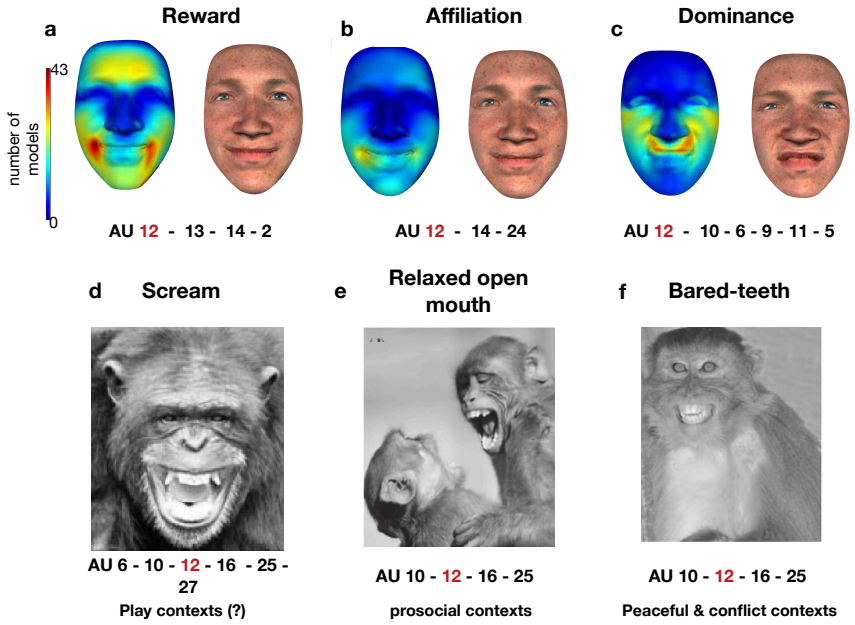
First, humans contract their zygomaticus major muscles when experiencing a variety of emotions, in a variety of contexts, ranging from pleasant to unpleasant. For example, spontaneous 'smiles' have been reported when experiencing distress (Ansfield, 2007) or during painful stimulation (Kunz, Prkachin, and Lautenbacher, 2009) —although the latter may include only the oblique raising of the lip.

Second, a recent data-driven approach has documented how smiles can communicate reward, affiliation, and dominance (Rychlowska et al., 2017). By using a large number of computer-generated dynamic facial expressions (2400 per participant; 48 participants), researchers modeled the different facial configurations needed to express affiliation, dominance and reward. Reward smiles were found to be symmetrical and accompanied by eyebrow raising (figure 1.10-a), affiliative smiles involved lip pressing (figure 1.10-b), and asymmetrical dominance smiles contained nose wrinkling and upper lip raising (figure-1.10-c).

Interestingly, the study of facial displays in monkeys also highlights how primates use the bilateral contraction of AU12 in different contexts to convey different social messages. On the one hand, the relaxed open-mouth face (figure 1.10-e) occurs mainly in positive, prosocial situations. On the other hand, the bared-teeth display (figure 1.10,-f) occurs both in positive and negative contexts like play and threat. This display is suggested to be homologous to the human smile, meaning that they descend from the same ancestral orolabial behavior (Parr and Waller, 2006; Van Hooff, 1972). Research suggests this display may communicate submissiveness, subordination, appeasement, or even the willingness to play, depending on the context and the species (Parr and Waller, 2006; Bliss-Moreau and Moadab, 2017).

In sum, while the prototypical smile gesture is associated with positive and prosocial behavior, subtle variations of its muscle configuration can drastically change the conveyed message, and extend its meaning to high-level social attitudes. Such a rich use of the AU12 in different social contexts is reminiscent of literature on monkey facial displays, which highlights how

FIGURE 1.10: (a-c) Facial expression models of three types of smile. Color-coded face maps show smile type (a) *reward*, (b) *affiliative*, and (c) *dominance* with their corresponding AUs that are shared across participant models (red indicates a high number of models, maximum = 43 models, see colorbar to left). The same facial expression patterns are also displayed on a single face identity to the right. (d-f) Monkeys facial expressions using AU12 with their correspondent action units: (d) Scream display (possibly positive) (e) Relaxed open mouth (prosocial behavior) (f) Silent Bared Teeth (peaceful and conflict contexts) (A-C) Adapted from Rychlowska et al., 2017 (D) Adapted from Parr & Waller 2006 (E-F) Adapted from Bliss-Moreau and Moadab, 2017



monkeys also feature AU12 in a wide variety of displays, species and contexts, essentially used to convey submissiveness.

1.5.2 Smiles across cultures and development

Cultural and universal aspects of smiling

As seen in section 1.1.2, facial expressions of emotions have, to some extent, an universal character. But not all facial expressions are equal. When looking back at the first studies on the cross-cultural recognition of facial expressions of emotions, recognition rates for happy displays were by far the strongest and most compelling (Ekman, Sorenson, and Friesen, 1969), with a mean recognition rate across cultures of 92%, compared to fear (62%), disgust

(54%), anger (64%), surprise (49%) and sadness(56%) which are consistently confounded in some cultures, e.g., people from New Guinea confounded anger displays with fear.

While recent data-driven approaches have challenged this view of universality (section 1.1.2), these studies still find the happy expression to be the most invariant and consistent pattern across cultures (Jack et al., 2012). However, culture-specific variations in how smiles are produced and perceived do occur, and depend on social context and culture. For instance, in a number of studies, Paul Ekman's collaborator David Matsumoto found that Japanese participants use smiles to mask their negative feelings in the presence of higher-status pairs (Matsumoto and Kudoh, 1993) and that, perhaps consequently, Americans judge smiling faces more intensely than Japanese participants who do not necessarily associate smiles with positive emotions (Matsumoto and Ekman, 1989). More generally, recent theories have associated a society's reliance on smiles with its history of heterogeneous or homogeneous migrations (Rychlowska et al., 2015) —the more members of a society have felt the pressure to affiliate with diverse populations, the more prevalent smiles are in that society.

The development of smiling

Not only are smiles used and recognised across cultures but also develop very early. The facial muscles involved in smiling specialize progressively from birth until being able to portray a protypical bilateral smile.

Smiles are thought to first develop in a purely motor fashion. A few days old neonates show more unilateral than bilateral spontaneous smiles, whereas for two months old infants almost all spontaneous smiles are bilateral (Kawakami et al., 2006). The specialization of facial muscles involved in smiling plausibly happens during sleep : in Messinger et al. (2002) twenty-five full-term, healthy neonates (mean age = 55 hours) were videotaped during six minutes of sleep. One-half of the neonates showed bilateral Duchenne smiles among which one-quarter showed bilateral smiles at a mature level of intensity. A baby's first smiles are endogenous : they have an internal cause, rather than represent a reaction to an external stimuli. Researchers studying these first smiles interpret them as being "pre-emotional" because they involve no cognitive evaluation and/or because they are not associated with

stimuli linked to pleasure or positive feelings (Messinger et al., 2002; Camras et al., 2016).

Social smiling (e.g. smiling during social interactions, or smiles triggered by external stimuli) emerges during the second month of life. Infants progressively show more occurences of Duchenne smiles in the context of interactions after their second month of life. For instance, in Wörmann et al. (2012), researchers studied the development of the social smile in a cross-cultural and longitudinal fashion. Researchers compared mother-infant interactions in Germany and Cameroon when infants where both 6 and 12 weeks old. Investigators found that at 6 weeks of age, mothers and their infants from both cultural communities smiled at each other for similar amounts of time —although quite infrequently. Interestingly, infants and mothers portrayed more social smiling when infants were 12 weeks old as compared when infants were 6 weeks old, in both cultures. Moreover, german mothers and their infants smiled and imitated each other more than did the Cameroon mothers and their infants. These findings show that social smiling appears more or less at the same time across cultures although it is latter strongly influenced by sociocultural factors. It is impressive that long before infants know how to talk they already know how to use the smile gesture in a social way, with the aim of communicating and bonding with their caregivers.

1.5.3 Processing smiles

Smiles play an important role in shaping social interactions. How exactly do we perceive and process our own and others' smiles? What underlying mechanisms support their high-level social roles ?

Facial-feedback hypothesis

One of the historically important findings on how smiling can affect our own mood is the facial feedback paradigm (Strack, Martin, and Stepper, 1988). In this study, the authors instructed participants to rate the funniness of cartoons either while holding a pen with their teeth (inducing a "smile") or with their lips (inducing a "pout"). Participants who were covertly asked to smile found cartoons funnier than controls. This important study, which was to some extent the foundation of embodiment theories, was reproduced in 2016

by 17 independent laboratories who could not successfully replicate its results (Wagenmakers et al., 2016). A more recent replication suggested that both the replication and the original study were in fact correct, but that video recording in the 2016 replications may have drastically influenced the results (Noah, Schul, and Mayo, 2018).

While not directly replicating the original Strack finding, a number of studies nevertheless give experimental support to the idea that smiling automatically triggers positive mood changes/affective judgement. For instance, participants may report more positive mood when they are associated a smiling virtual avatar (Neumann and Strack, 2000). Similarly, manipulating participants' own voice to sound more smiling/happy can induce positive emotions in the speakers themselves (Aucouturier et al., 2016).

Neural bases

As seen for emotions in general, amygdala activity seems to support smile perception. When measuring amygdala activity in human volunteers during rapid visual presentations of smiling and neutral faces the amygdala responded preferentially to smiles versus neutral faces (Breiter et al., 1996). Reward regions such as the striatum also appear to be activated both during the perception and the production of affiliative smiles, with the idea that smiles can function as a social reward for both adult and infants (Niedenthal et al., 2010). One the most notable study describes how Deep Brain Stimulation (DBS) in the right and left nucleus accumbens, can trigger the production of an asymmetric smile (Okun et al., 2004). The patient in the study spontaneously reported a simultaneous feeling of "giddiness" and an "urge to laugh" (which according to the authors she did on several occasions). When asked why she laughed, the patient wasn't able to report a precursor to her smile, felt embarrassed and would even attempt to suppress her smile.

Motor regions are also involved when processing smiles. Indeed, the first recording of mirror neurons in humans used the smile gesture to study motor activity to both viewing and producing a specific gesture (Mukamel et al., 2010), thus providing evidence of motor-cortex activity during both smile perception and production. In this study, researchers recorded extracellular neural activity in 21 patients while they executed and observed facial emotional expressions (frowning and smiling) and hand-grasping actions. Significant proportions of cells responding to both perception and action were

found both in supplementary motor areas (SMA) and the medial temporal lobe. Another interesting fMRI study highlights the role of the motor cortex and mirror system during facial mimicry to smiling faces (Likowski et al., 2012). In this study, researchers asked 20 female participants to view emotional facial expressions, while recording brain responses as well as zygomatic and corrugator activity. Results show prototypical patterns of facial mimicry as described in previous findings, but also that these facial reactions correlate with activations in the inferior frontal gyri, SMA and cerebellum.

The SIMS (Simulation of Smiles) model

One of the only models of smile perception is the SIMS (Simulation of Smiles) model. This model takes into account findings about facial feedback, embodied cognition and neuroimaging to suggest how smiles are processed in the brain (Niedenthal et al., 2010).

According to SIMS, different mechanisms underlie the perception and processing of different kinds of smiles. For instance, for enjoyment/reward smiles, processing begins with the detection of uncertainty, generating amygdala activity which, in turn, directs gaze toward the eyes (i.e. triggering eye contact), followed by the generation of reward in the basal ganglia, motor mimicry, and corresponding somatosensory experience.

For affiliative smiles, SIMS posits that the orbitofrontal cortex (OFC) and other prefrontal regions would also be involved, selectively supporting the distinctive positive feeling of seeing an individual smile with whom one has a close relationship. Findings show how e.g. the OFC differentiates the sight of one's own smiling baby from the sight of an unknown smiling baby (Minagawa-Kawai et al., 2008).

When processing dominant smiles, which the SIMS model suggests hide negative intentions, the experience of negative affect would induce right-lateralized activation. Frontal regions would also be involved in their processing as these are central to processing social status (Niedenthal et al., 2010). Like for other types of smiles, activation of cortical motor regions and subsequent facial mimicry also occur, producing somatosensory experiences associated with the feeling of "being dominated" (e.g., that the smile is experienced as “superior” or “condescending”).

Generally, the SIMS model places facial mimicry as a causal component of the cognitive processing of smiles, and supports the idea that eye gaze is a critical modulating factor of this imitatory behavior.

1.5.4 Auditory smiles : perceiving smiles in the voice

Known form; unknown origin

I have introduced several key concepts about smiles, such as their function and ubiquity across cultures and development. But one essential aspect was left aside: their origin. Why do smiles look the way they do? Why do we stretch our lips, and raise our cheeks, when we are in a good mood or we want to communicate affiliation?

Darwin was first to suggest that the visual features of emotional expressions have functional roles. The morphology of a facial display, e.g. closing or opening the eyes or the mouth, may first serve an adaptive role in, e.g., modulating sensory inputs to the organism; it is only latter that these displays adopted the social communicative roles they are widely known for (Darwin, 1872). Such functional origins, proposed Darwin, explain why some displays are found independently of culture.

Recent research have comforted this view. Fear displays for instance have been found to widen the visual field, enhance eye movements during target localisation and increases nasal volume and air velocity during inspiration (Susskind et al., 2008; Lee and Anderson, 2016). All these effects support the notion that fear displays enhance sensory acquisition, probably to optimise the individual reactions in threat situations.

Antagonic evidence is reported for the disgust display. In line with the idea that disgust is an emotion that is associated with sensory rejection, e.g. during the exposure to potentialy poisonous stimuli, the facial expression of disgust (squinted eyes, wrinkled nose) was found to reduce the visual field, and more importantly, decrease nasal volume and air velocity during inspiration (Lee and Anderson, 2016).

No such evidence, however, has been reported for smiles. To my knowledge, there is only one theoretical view suggesting why smiles have the shape they have. In *The acoustic origin of the smile*, John Ohala (1989) suggested that the smile may have originally served a vocal function.

Ohala's argument is the following. Smiles change drastically the shape of the vocal resonator. Indeed, the stretching of the lips shortens the vocal tract, which, as we saw in section 1.3.1, results in raising the voice's formants. Importantly, voice formants are one of the main acoustic cues for signaling body size in the animal kingdom (Bowling et al., 2017), and body size information, specifically smallness, is used across species to signal appeasement, submission and the absence of threat (the so-called "size code hypothesis", Briefer, 2012). Ohala therefore hypothesised that the smile facial gesture may have served the communicative role of reducing an individual's perceived body size (Ohala, 1980) not by virtue of its visual appearance, but of its acoustic consequence.

It is important to note that, today, the function of emotional displays may have changed. Social animals may very well have begun to use facial expressions for their desirable properties of e.g. sensory enhancement/rejection, but latter their social and communicative role eventually surpassed their functional purpose (Susskind et al., 2008). Smiles, says Ohala, may have evolved for their sound, but would have become ritualized and can now be used as a purely visual display, without necessarily having to vocalize.

Smiled speech and auditory smiles

While much research on smiles has been centered on its visual perception, the acoustics consequences of smiling are therefore theoretically important .

The field of phonetics has investigated how different articulators, such as lip stretching, change formant frequencies and phonemes. A simplified version of the consequences of lip stretching in the phonetic literature is to go from phonemes on the right of the vowel space (Figure 1.6-a) to those in the left, for instance going from [u] (u like in round) to [e] (e like in sheep). Models of speech production suggest that lip rounding increases vocal tract length, resulting in lower formants, while smiling decreases vocal tract length and reduces formant frequencies. Lip stretching therefore has the eventual consequence of going from back vowels (rounded lips) to front vowels (unrounded lips).

With smaller amplitude, the contraction of the zygomaticus muscles associated with smiling does not change phonemes but rather modifies their timbre (Tartter, 1980; Basso and Oullier, 2010). Bell Labs psychologist Vivien Tartter was, to my knowledge, the first to study this phenomenon. In 1980 (the same

year Ohala published *the acoustic origin of the smile*), she asked participants to produce sentences with or without smiles, and found that these were rated as more positive and happier than their neutral counterparts. Acoustic analysis revealed that smiled sentences had higher formant frequencies, but also higher fundamental frequency, amplitude and sentence duration.

Subsequent research in "smiled speech" investigated how these acoustic cues relate to smiles, but also how they are used to communicate positive affect. In Barthel and Quené (2015), smiling was found associated with an increase of the second formant (F_2) for words with the round vowel /o:/, of intensity as well as F_0. In Podesva et al. (2015), smiling was associated with an increase of F_2, in El Haddad et al. (2015a) and El Haddad et al. (2017) with an increase of formants and F_0. Higher F1 and F2 dispersion are also reported (Drahota, Costall, and Reddy, 2008). Overall, the literature suggests smiled speech is caracterised by higher mean pitch, higher intensity and higher formants (Quené, Semin, and Foroni, 2012; Barthel and Quené, 2015; Lasarcyk and Trouvain, 2008).

The causal link between the motor act of smiling and the different acoustic features of smiled speech is not straightforward. For instance, Tartter showed that smiles can be recognised in whispered speech, i.e. even in the absence of clear fundamental frequency (Tartter and Braun, 1994). In the same line, smiling can be recognised in individual phonemes (Barthel and Quené, 2015). This demonstrates that pitch and prosody are not necessary to recognize smiles in speech, but more plausibly reflect the positive affective states displayed when asked to vocalize with a smile. Despite years of research on the acoustics of smiled speech, it is not yet clear what features of speech necessarily result from smiling —or simply co-occur with it.

During this dissertation, I introduce the term *auditory smiles* to describe the direct acoustic consequences of smiling, separating them from the more general emotional cues that may co-occur with their production in ecological speech. Auditory smiles can be defined as *"the acoustic consequence of the contraction of the zygoamtic muscle in speech"*, and are defined mainly by changes in sound spectrum.

1.5.5 Section Summary

The particular case of the smile

Smiling is one of the most important gestures in the human emotional repertoire. They are recognised across cultures, develop early, and are used in a great variety of social contexts, usually serving an affiliative function. However, despite their ubiquity, we still do not know why smiles look the way they do. In one notable theory, Ohala (1980) suggested that the origin of the smile is acoustic: by reducing the perceived length of the vocal tract, smiling signals a smaller body size and, thus, that the signaller is less of a threat to the observer.

While they may hold the key to why and how we smile, the acoustic consequences of stretching lips while speaking have not been studied thoroughly. First, it has been difficult to disentangle which acoustic features directly result from zygomatic contraction, rather than simply co-occur with it. Second, because auditory smiles have not been yet clearly characterised acoustically, the mechanisms underlying their perception are also unknown.

1.6 Thesis Overview

The aim of this thesis is to investigate the mechanisms that underlie the perception of "auditory smiles" and question the depth of their cognitive processing which, if theories like Ohala's are correct, may well be as sophisticated as that of their more-widely-studied visual counterparts.

This thesis uses a variety of behavioural, electrophysiological and computational methods to study auditory smiles' emotional and perceptive processing. In chapter 2, I present two studies attempting to characterize the acoustic fingerprint of auditory smiles. The first study investigates the acoustic features caused by smiling during production of smiled and non-smiled phonemes; the second study uses the paradigm of reverse correlation to study how smiles are mentally represented by listeners.

Based on these results, I then present, in chapter 3, a digital audio algorithm developed to recreate the acoustic consequences of smiles in speech. This algorithm, based on the phase-vocoder technique, was designed to change only the specific acoustic cues linked to smiles, while leaving other speech dimensions unchanged —a crucial feature allowing us to have precise experimental control of sound stimuli throughout the rest of the thesis.

Chapter 4 uses stimuli generated with this algorithm in an emotional mimicry experiment. As presented in section 1.2.3, emotional mimicry is related to empathic processing and is at the basis of models of smile perception (SIMS). I will show that, just as visual smiles, auditory smiles can trigger facial reactions which are, to some extent, unconscious.

Chapter 5 presents an eye-tracking study showing how smile information from voice and face are jointly integrated. Using a visual algorithm to recreate smiles in the face, we created congruent and incongruent audiovisual smiles and tested how these influence face exploration strategies. I will show that pupil dilation and eye gaze index how visual and auditory smiles are integrated during social cognition.

In Chapter 6, I will show that the motor reactions involved in the cognitive processing of auditory smiles are even present in congenitally blind participants, who have never seen a facial expression, thus suggesting that the visual experience of a smile is unnecessary for the facial imitation mechanisms to develop and operate.

Finally, I will discuss how these results challenge the current models of smile perception, and how they shed light on the nature and task-dependance of emotion processes. I will conclude this dissertation with a critical discussion of the interdisciplinary methodology used in this thesis, i.e. that of combining the two fields of audio signal processing and cognitive science to study concepts like auditory smiles and vocal emotions. This methodology holds a lot of potential for causal inference in experimental research but, I would contend, also holds important challenges that transcend the typical *savoir-faire* of either fields, and can therefore be challenging in the context of a PhD.

2 The acoustic fingerprint of the smile

As seen in the introductory chapter of this thesis, stretching lips while speaking changes the shape of the vocal resonator, reducing the length of the vocal tract, and thus transmitting filtered frequency content from the glottal pulses compared to normal speech. The aim of this chapter is to investigate (1) how smiling causally changes the spectral features of the voice (Study 1), and (2) whether these spectral features are used during auditory smile perception and internally represented (Study 2).

Study 1, which relies on the recording of a new corpus of sounds, and its acoustic analysis, was published as part of the article *"Realistic transformation of facial and vocal smiles in real-time audiovisual streams"*, (Arias et al., 2018). Study 2, which uses the paradigm of psychophysical reverse-correlation, was published in *"Uncovering mental representations of smiled speech using reverse correlation"*, (Ponsot, Arias, and Aucouturier, 2018).

2.1 Study 1 - Auditory smiles as they are produced

2.1.1 Introduction

The aim of this first study is to clarify the acoustic fingerprint of auditory smiles, and investigate to what extent smiles have similar acoustic consequences across a wide variety of voiced and unvoiced phonemes, produced at different pitchs and by different individuals.

2.1.2 Methods

We recorded a dataset of French phonemes, uttered with and without smiling, and conducted an acoustical analysis of the recordings. We asked N = 8 (male: 6) participants to pronounce 9 types of phonemes (5 voiced: *a, e, i, o, u* [a,ə,i,ɔ,y] and 4 unvoiced: *s, h, j, f* [s, ʃ, f, ʒ]), each with and without stretched lips. Phonemes were recorded three times each, each time at 3 different pitches. The dataset was recorded at sampling rate 44.1kHz, in a sound-proof booth using a high quality microphone (DPA 4088 F). In the following, we analyse the recordings with phonological analysis software to measure the impact of smiling on three aspects of sound spectrum: formants, spectral envelope, and spectral centroid.

2.1.3 Results - acoustic analysis

Consequences of smiling on formants

We analysed formant frequencies for all the smiled and non-smiled voiced phonemes using the PRAAT software (Boersma, 2002). Statistical analysis showed a significant increase of mean F_1 between the non smiled and the smile condition (a 5% increase from M=483 Hz to M=507 Hz; paired t-test t(7)=3.5, p=.008), and a marginally significant increase of F_2 (4% from 1572 Hz to 1634 Hz; paired t-test t(7)=1.9, p=.09), see Figure 2.1-a.

Consequences of smiling on the spectral envelope

We analysed the spectral envelope of the recordings using the adaptive true envelope technique (Villavicencio, Robel, and Rodet, 2006; Röbel and Rodet,

2005). Spectral envelopes measured while smiling have more energy in the high-frequency regions, both for voiced and unvoiced phonemes (Figure 2.1-b). For voiced phonemes, the main difference between the smiled and non-smiled envelopes is found between 700 and 4000 Hz, corresponding to a shift of frequency and boost of amplitude of the region around F1-F3. For unvoiced phonemes smiling affects higher frequencies, creating both resonances and antiresonances in the spectral envelope.

Consequences of smiling on the spectral centroid

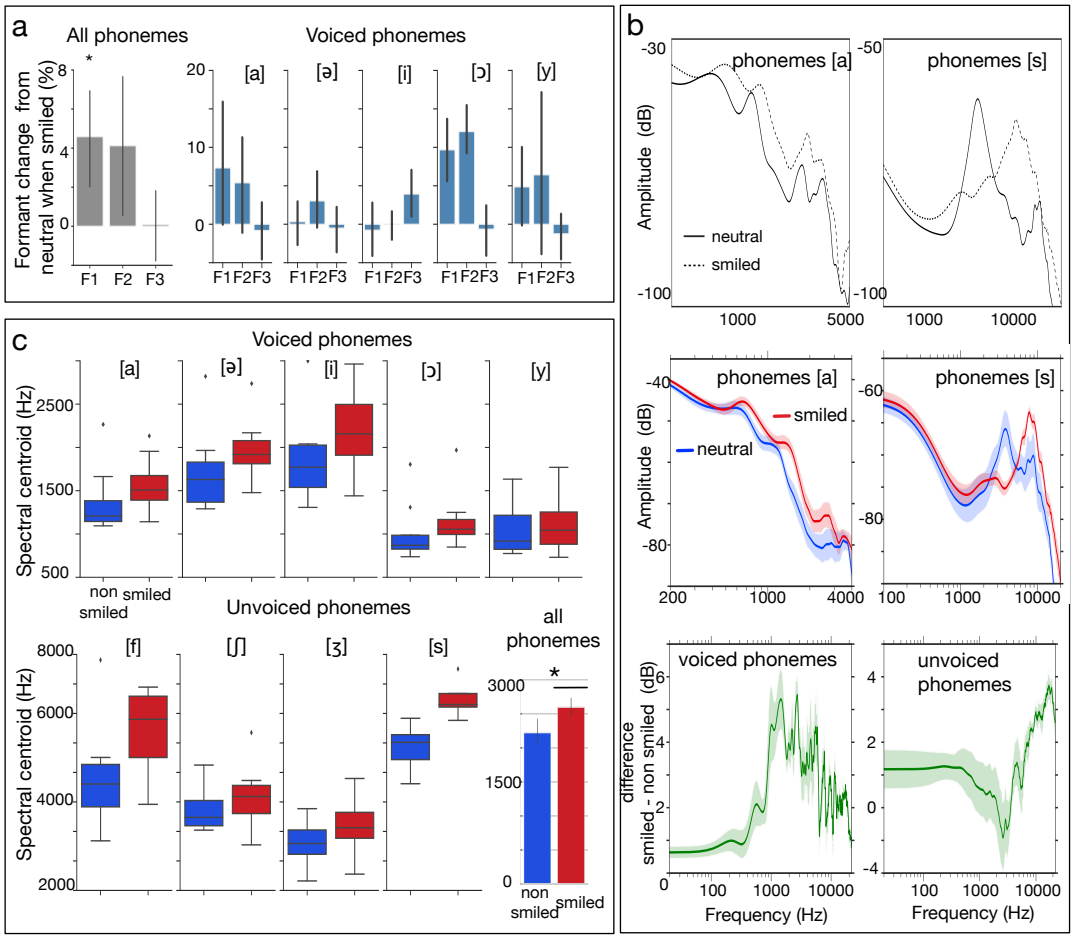
Finally, we analysed the spectral centroid (or "center of mass" of the spectrum, a measure related to perceived brightness) with a custom Python script for all the phonemes of the database (Figure 2.1-c) and found that the mean spectral centroid increases for every phoneme of the database when smiled, regardless of whether the phoneme is voiced, unvoiced, opened or closed. The overall effect is statistically significant (paired t-test $t(7) = 6.2$, $p=.0004$).

2.1.4 Discussion

These results are in line with previous literature, which suggested that smiling increases formants in speech (El Haddad et al., 2015a; El Haddad et al., 2017; Drahota, Costall, and Reddy, 2008; Quené, Semin, and Foroni, 2012; Barthel and Quené, 2015; Lasarcyk and Trouvain, 2008). The consequence of the smile gesture across the phoneme inventory is to increase formants and high frequency content, although not identically across phonemes. In particular, the spectral centroid of all phonemes increased when smiled, independent of the nature of the phoneme (i.e. opened or closed). In addition, all phonemes presented at least one formant increase, and no formant decreases across phonemes.

Taking account of these results, together with previous findings (El Haddad et al., 2015a; El Haddad et al., 2017), we conclude that the average acoustic consequence of smiling on sound spectrum, across phonemes, is to increase the frequency of both formants F1 and F2 by 5-10% (Figure 2.1-a) and to increase the amplitude of high frequency energy in both voiced and non-voiced phonemes (Figure 2.1-b).

FIGURE 2.1: Smiled speech corpus analysis. (a) Consequences of smiling on formants: Mean frequency shift of the first three formants, expressed in percentage of the non-smile utterance, averaged for all phonemes (left) and for each voiced phoneme (right) in the corpus. (b) Consequences of smiling on the spectral envelope. Top: Time-averaged spectral envelope of a single utterance of a French phonemes *'a'* and *'s'*, pronounced with and without smile. Middle: Averaged spectral envelope for all *'a'*s and *'a'*s of the corpus in smile and non-smiled conditions. Error bars represent standard errors. Bottom: Mean spectral envelope difference (smile minus non-smile) for all voiced and unvoiced phonemes of the corpus. (c) Consequences of smiling on spectral centroid. Mean spectral centroid for voiced (top), unvoiced (bottom left) and all (bottom right) phonemes in the corpus. Error bars represent 95% confidence intervals on the mean.



2.2 Study 2 - Auditory smiles as they are perceived

2.2.1 Introduction

While the acoustic signature of smiles can be examined in corpus analyses, it is a different question altogether whether and how these physical characteristics are used as cues during perception. How are the sensory characteristics of auditory smiles internally represented by the listeners ? In this second study, this question was addressed using the experimental paradigm of reverse-correlation.

The general idea of reverse-correlation is to present a system (here, the human observer) with a slightly perturbed stimulus over many trials. This perturbation can be created by directly adding white noise to a stimulus or by manipulating higher-level dimensions using random deviations around baseline. Perturbated stimuli will, on different trials, lead to different responses of the system. The tools of reverse-correlation can be used to infer the functional properties of the sensory system (i.e. here, the listener) from the pattern of stimulus noise and their associated responses. The technique was first used by psychophysicists to characterize human sensory processing (e.g., detection of tones in noise; Ahumada Jr and Lovell, 1971; discrimination of frequency distributions; Berg, 1989) but it is also a powerful tool to characterize higher-level perceptual or cognitive processes, for which it can uncover the "optimal stimulus" (or "mental representation") that is driving participant responses.

In vision, reverse-correlation was applied to derive observers' mental representations of, e.g., what makes a face happy (Mangini and Biederman, 2004), how emotional facial expressions differ across cultures (c.f Section 1.1.2; Jack et al., 2012) or even what makes Mona Lisa appear to be smiling (Kontsevich and Tyler, 2004). A few recent studies have started to use the approach for auditory tasks such as speech intelligibility (Varnet et al., 2016; Venezia, Hickok, and Richards, 2016), speech prosody (Ponsot et al., 2018) and musical instrument recognition (Thoret, Depalle, and McAdams, 2016). In particular, Owen Brimijoin et al. (2013) have used reverse-correlation to uncover the internal representations of a whispered vowel by presenting random-spectrum static noises to human listeners. Their results showed that humans possess strikingly fine spectral mental representations of a vowel, with spectral weights aligned to the formant frequencies of real whispered vowels.

In the present study, we used reverse correlation to characterize the perceptual filters employed by humans to infer whether a person is smiling. Listeners were presented with hundreds of pairs of utterances (of a single vowel, [a]), with randomly manipulated spectral characteristics and asked to indicate, in each pair, which was the most smiling. We examined how participants internally represented the sound of a smile, and assessed the robustness of this representation by quantifying its internal noise.

2.2.2 Methods

Ethics

The protocol of this experiment was approved with an IRB given by the "Institut Européen d'Administration des Affaires" (INSEAD).

Subjects

Ten participants (5 women, 5 men; age 18–29 yrs) were recruited for the experiment. None reported having hearing problems. In accordance with APA Ethical Guidelines, participants gave their informed written consent prior to the experiment and were debriefed about the true purpose of the research immediately after. Participants were paid for their participation.

Stimuli

We recorded an utterance of the phoneme [a], pronounced with constant pitch (122Hz) by a single male speaker with a neutral facial expression, and selected a 500–ms stationary part of the sound to create a stimulus with constant spectral energy (SI audio 1: audio file of the original /a/ vowel pronounced with a neutral facial expression). We then produced many spectral variants of this baseline stimulus by manipulating its spectral characteristics using a random frequency equalizer composed of 25 linearly interpolated, log-separated frequency points spaced between 100 and 10 000 Hz, with gain values (in dB) drawn from Gaussian distributions [standard deviation (SD) = 5 dB clipped at 62.5 SD].

Apparatus

All stimuli were mono sound files at a sampling rate of 44.1 kHz with 16-bit resolution using MATLAB. They were presented diotically through headphones (Beyerdynamic DT 770 PRO, 80 ohms) at the same level for all participants (70dB sound pressure level). Sound levels were measured using a Brüel & Kjær 2238 Mediator sound-level meter placed at a distance of 4 cm from the right (left) earphone. A DPA 4066 omni-directional microphone was used to record the voice of the male speaker employed to create the stimuli.

Procedure

The experiment consisted of a single 1 h experimental session including 6 blocks of 100 trials. Using a two-alternative forced choice (2AFC) procedure, participants were presented pairs of randomly-filtered voices (hear a trial example in SI Audio 2) and asked in each pair which of the two appeared to have been produced with the greatest smile. Since there were no correct or incorrect answers, participants did not receive trial-by-trial feedback. Trials presented in the first 5 blocks were all different, but the 100 trials of the sixth block were the same as those presented in the fifth block (in the same order). This double-pass procedure was used to evaluate the percentage of agreement and thus the level of internal noise for each observer in the task. None of the observers noticed this repetition.

Data Analysis

One reverse-correlated frequency filter (a 25-points vector) was computed for each subject as the mean filter of the voices classified as smiling from which we subtracted the mean filter of the remaining voices that were not chosen as smiling by the participant (the data collected during the sixth block, i.e., the same trials as in the fifth block, were not used to derive these filters).

Formant frequencies and bandwidths were computed using Praat (Boersma and Weenink, 2017). The spectral envelopes were extracted using the true envelope implementation of IRCAM's Super-VP tool (Villavicencio, Robel, and Rodet, 2006). Formant gains were estimated as the values of the spectral envelope at the formant frequencies.

2.2.3 Results

Perceptual filters and mental prototypes

The averaged reverse-correlated frequency filter underlying the evaluation of smile in the [a] vowel used in the task is plotted in Figure 2.2-a. The structure of this filter is clearly (yet entirely agnostically) aligned with the formant frequencies and bandwidths of the original phoneme. The filter also shows an overall amplification of the high frequencies compared to the low frequencies. Because the reverse-correlation technique only allows us to compute participants' internal filters with amplitudes that are proportional to the SD of the external noise used in the experiment, we generated prototype stimuli for smiling and non-smiling voices by applying the filters to the base stimulus with a gain of 650 (Figure 2.2-b), a value chosen to qualitatively match the average spectral-envelope differences of the stimuli presented in the experiment. These prototypes have fair intra-individual consistency and appear to implement distinctive operations on the formants: (i) formants F1 and F2 are represented with increased frequency in the smiling prototype (in red, SI audio 3: smiling audio prototype derived from the perceptual results), compared to the non-smiling prototype (in blue, SI audio 4: Non-smiling audio prototype derived from the perceptual results.) and (ii) formants F2, F3, and F4 are represented with increased amplitude. Figure 2.2-c presents the difference between the spectral envelopes computed over these prototypes: it is virtually identical to the raw filter plotted in figure 2.2-a, showing that the filter profiles represent the real physical changes that occurred on spectral envelopes. Overall, as summarized in figure 2.2-d, the spectral transformations required for our participants to correctly perceive the phoneme as smiling consist primarily of a frequency increase of F1 and F2 and an amplification of F2, F3, and F4.

Observer's consistency

The double–pass methodology was used to assess observers' consistency from a measure of internal noise relative to the external noise added to the stimuli: the last two blocks were identical so that all observers were presented the same 100 trials twice. All but two participants (who had percentages of agreement of 4% and 51%) performed well above chance over these repeated blocks: when these two participants were removed, the average

percentage of identical responses over repeated blocks was 68.1% (SD = 5.8). We estimated the amount of internal noise for each of the remaining eight subjects using a signal detection theory model with late additive noise (Neri, 2010). We found an average internal noise level of 1.2 (SD = 0.9), as expressed in units of external noise SD. There are no available report of internal noise measures in facial emotion or visual smiles tasks to which we can compare our present estimate, but it is of note that previously reported values in other high-level visual processing tasks are generally higher than our estimate; an internal noise level of 2 is found for human biological motion discrimination (running vs walking) (Boxtel and Lu, 2015), levels higher than 2 for the evaluation of ensemble average size (Im and Halberda, 2013) and between 1 and 4 for face identification (Gold, Sekuler, and Bennett, 2004). The internal noise level found here (1.2) rather corresponds to levels generally seen in different low-level auditory and visual tasks (Neri, 2010), suggesting that the high-level auditory filtering of smiles in speech relies on a fairly stable processing, i.e., that observers possess robust and stable auditory representations of what is smiled, and what is not.

2.2.4 Discussion

This study examined what spectral filtering underlies the auditory processing of smile in the human voice, using behavioral reverse-correlation. We show that humans rely on robust mental representations that allow them to tell whether a voice is smiling or not, and that these internal representations include increased F1 and F2 frequency, increased F2, F3 and F4 amplitude, and an overall enhancement of the high frequencies compared to the low frequencies. The structure of these filters demonstrates a delicate ability of the auditory system to parse amplitude and frequency changes by formant. In addition, internal representations were fairly consistent across participants, demonstrating that auditory consequences of articulatory gestures associated to smiling are accurately available even to naive participants.

This study focuses on the special case of phoneme [a]. For the particular phoneme, our results are in line with Keough et al. (2015), which report an increase of F1 and F2 during production. As shown by Study 1 and previous literature on the acoustics of smile production (Barthel and Quené, 2015; Fagel, 2010; El Haddad et al., 2015a; Keough et al., 2015), even if the consequence of smiled speech on formants depends on the vowel, these always

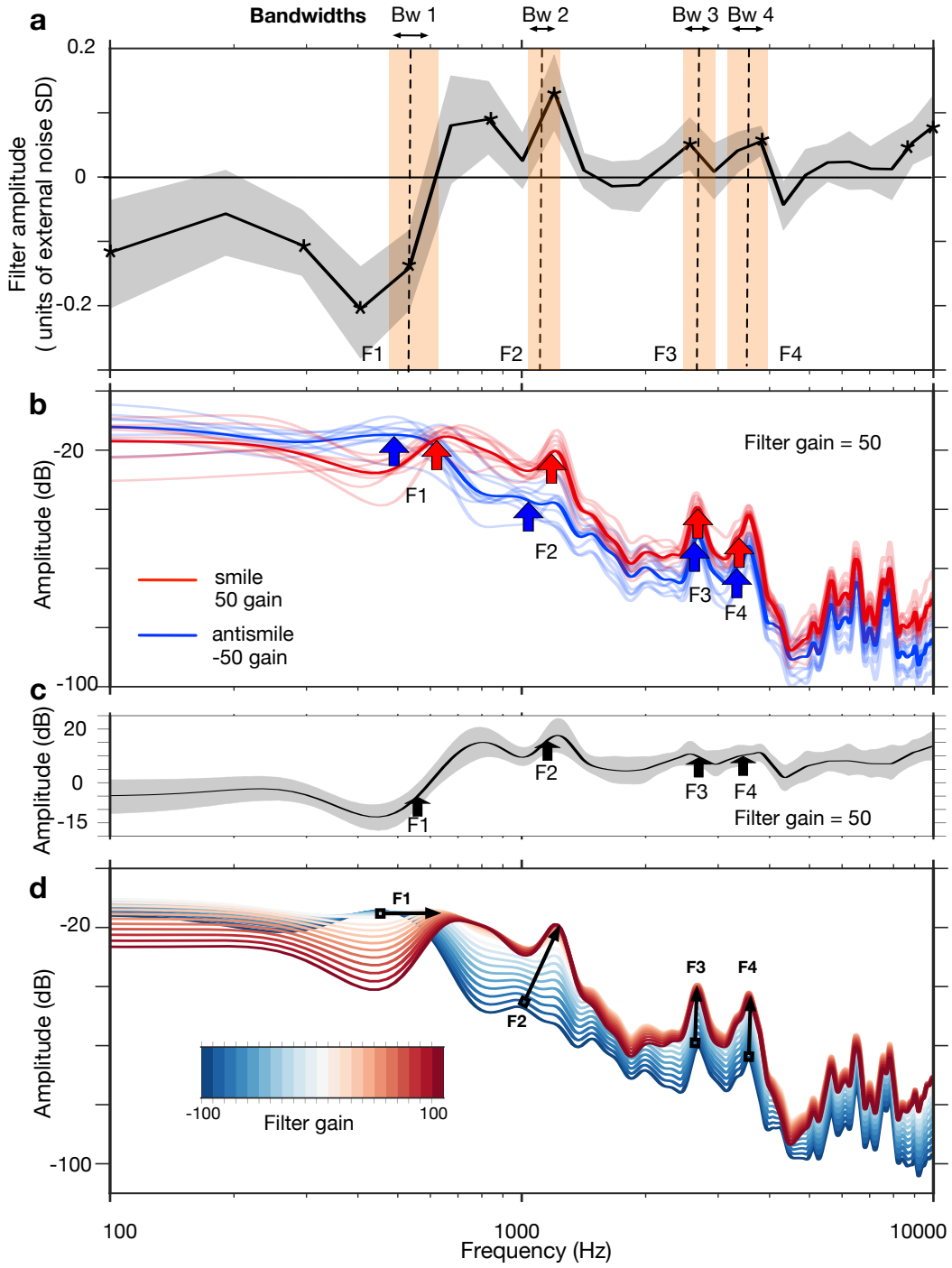


FIGURE 2.2: (Caption next page.)

FIGURE 2.2: (Previous page.) Reverse correlation analysis (a) Averaged filter underlying the judgment of the vocal smile, as derived with reverse-correlation. Asterisks indicate significant differences from 0 (two-tailed; paired-sample t-tests, $p < 0.05$). Vertical shaded areas indicate how the first four formants of the voice align with the structure of the filter. (b) When this filter or its opposite is applied (here with a gain of 50) to the original voice, it reveals the internal auditory representations of a smiling (SI audio 3) and a non-smiling voice (SI audio 4). Shaded lines represent the corresponding spectral envelope for each participant's internal filter. (c) Mean spectral envelope difference between smiling and non-smiling sounds for a filter gain of 50. (d) Mean spectral envelope across participants for different filter gains, highlighting the overall transformation over the spectral envelopes as one goes from mental representations of a strongly non-smiling voice, to that of a strongly smiling voice. Shaded areas represent 95% confidence intervals computed with a bootstrap procedure.

exhibit an overall increase in frequency. Thus, it can reasonably be assumed that the filters returned with our method for other vowels and/or speakers would implement changes on the formant structure of the tokens, although in details these may be specific to the considered phoneme. If such is the case, the listeners' ability to recognise smiled phonemes would be a remarkable mechanism, as the acoustic consequences of smiling are non-linear across the spectrum and across phonemes.

2.3 Chapter conclusion

The acoustic fingerprint of the smile

This chapter presented two studies aiming to measure the acoustic and auditory fingerprint of auditory smiles. Study 1 showed that smiling mainly increases F1 and F2 formant frequencies and high frequency content. Study 2 showed that, when tasked to discriminate smiled and unsmiled phonemes, participants rely on robust spectral representations that specifically implemented vowel formant modifications. These findings demonstrate the causal role played by formants in the perception of auditory smiles and shed light onto the remarkable abilities of the human auditory system to use the acoustic features of voice to infer the oro-labial/articulatory gestures with which it was pronounced.

In order to study auditory smiles' cognitive processing, we first need to control these smile specific features in experimental situations. To do so, the next chapter will introduce a digital audio algorithm to control auditory smiles in running speech, with the aim of generating stimuli for subsequent studies.

3 Modelling auditory smiles

As seen in the introductory chapter of this dissertation, several emotional cues in speech co-ocurr with smiles during positive affective states, but are not directly caused by the contraction of AU12. In order to study the cognitive processing of auditory smiles, one needs a way to control for those acoustic dimensions. One typical way of dealing with this problem is to use actor vocalisations as stimuli. Although actor vocalisations have the advantage of being natural stimuli (produced in artificial situations), actors usually can not parametrically vary only one dimension in speech leaving all others unchanged. Because such level of acoustic control is crucial in the study of auditory smiles, this thesis will instead use a computational approach to control for acoustic feature variation in stimuli.

The aim of this chapter is to present a digital audio signal processing algorithm able to recreate the acoustic changes linked to smiling in running speech (identified in chapter 2), while leaving unchanged other emotional and non-emotional dimensions in speech, such as pitch contour, speech rate or content. This chapter presents the technical details of this 'smile transformation' algorithm, as well as three additional studies (Study 3, 4 and 5) attempting to validate the effectiveness of the effect on listeners' perception. Work reported in this chapter has been published as part of *"Realistic transformation of facial and vocal smiles in real-time audiovisual streams"* (Arias et al., 2018) and *"Auditory smile trigger unconscious facial imitation"* (Arias, Belin, and Aucouturier, 2018).

3.1 Algorithm design

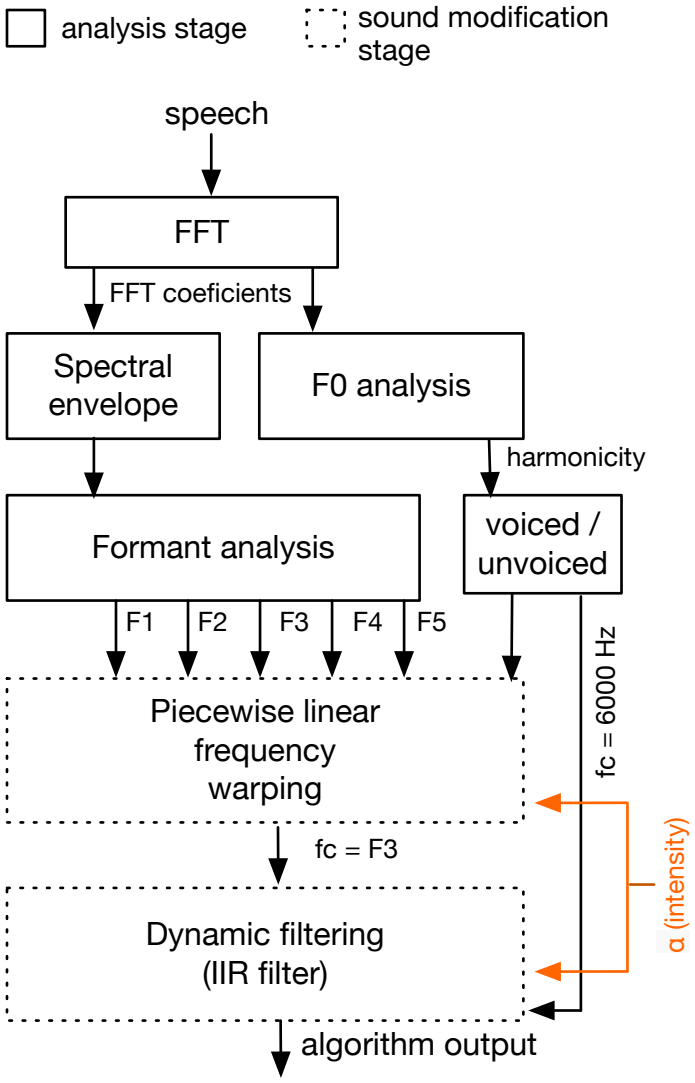
To simulate the acoustic changes caused by smiling on arbitrary spoken input, we designed a two-stage signal processing algorithm, which first transforms (or *warps*) the vocal spectral envelope, and then filters the reconstructed signal adaptively. Both stages are informed by a prior detection stage which tracks the positions of the formants. Figure 3.1 shows a general view of the algorithm.

This approach is different from the literature in several ways. First, compared to Quené, Semin, and Foroni (2012), Barthel and Quené (2015), El Haddad et al. (2015a), and El Haddad et al. (2015b), what we implement here is a transformation (i.e., operating on real speech input, and preserving its identity, prosody and content) rather than a synthesis technique (i.e., which generates speech from scratch). Second, by operating only on the spectral envelope and preserving the harmonic partials of the original voice, we avoid artifacts caused by the synthetic glottal impulses found with other formant re-synthesis approaches. Finally, although we don't use that feature in this thesis, the frame-by-frame architecture of the system makes it suitable for real-time processing, i.e. transforming speech as it is produced, with a latency adapted e.g. to a telephone conversation.

3.1.1 Piecewise linear frequency warping

In order to model how smiling transforms the vocal tract filter, we use a spectral envelope manipulation technique, called frequency warping, which does not only transform the local peak resonances (formants) but also the acoustic details besides these local peaks, e.g anti-resonances. Frequency warping was introduced to normalize vocal tract differences across speakers in order to improve the performance of recognition and categorization algorithms (Lee and Rose, 1996). More recently, it was applied to make speakers unrecognizable (a process called speaker de-identification; Magariños et al., 2016), transform a speaker's voice into another speaker's (voice conversion), or transform a voice's apparent sex (Tian et al., 2014; Erro, Navas, and Hernaez, 2013). Here, we use frequency warping to shift the spectral envelope (with its formants) either high or down with the aim of reinforcing or reducing the smile impression of a voice.

FIGURE 3.1: Overview of the audio smile transformation. The first stage of the algorithm is a transformation of the audio frames to the frequency domain, followed by both spectral envelope and f0 analysis. Spectral envelope analysis allowed us to compute speech's formants and f0 analysis to extract its harmonicity, and to categorize it either as a voiced or unvoiced frame. The two dotted blocks are the sound transformation stage, informed by the formant frequencies and harmonicity parameters extracted in the first stage.

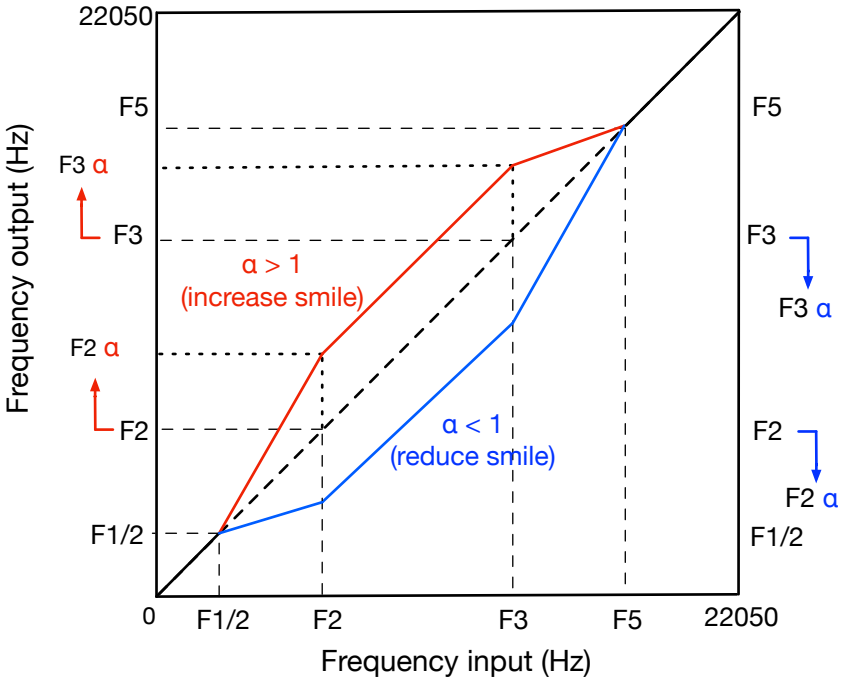


The algorithm operates on a frame-per-frame basis. For each frame, it estimates the vocal spectral envelope (f_{in}), using the 'true envelope' technique (Villavicencio, Robel, and Rodet, 2006; Röbel and Rodet, 2005), and manipulates it using a non-linear change, or warping, of the frequency dimension (f_{out}). The intensity and direction of the warping are controlled by the parameter α, such as $f_{out} = \Phi(f_{in}, \alpha)$. The transformation function Φ, illustrated in Figure 3.2, was heuristically designed to shift the voice's formants by stretching and warping parts of the spectral envelope, and generate similar formant distributions as the ones measured in Chapter 2, increasing F1 and F2 frequencies. Φ is piece-wise linear with cut-frequencies defined as a function of the input signal's formant frequencies F_i: the output spectral envelope is untransformed below $F_1/2$ and above F_5; the segment between $F_1/2$ and F_2 is warped so that the spectral envelope at F_2 is mapped to $\alpha.F_2$ and F_3 to $\alpha.F_3$; eventually, the last segment between $\alpha.F_3$ and F_5 is warped to return to identity after F_5. Finally, we reapply the warped spectral envelope to the harmonic information and resynthesize the signal using the phase vocoder technique (Roebel, 2010; Liuni and Axel, 2013).

Note that, if $\alpha = 1$ then $f_{out} = f_{in}$; if $\alpha > 1$, the algorithm shifts the envelope towards the high frequencies, and the higher α, the higher the shift, which should increase the smile impression in a voice; Conversely, if $\alpha < 1$, the acoustic effect is opposite and the envelope is shifted towards the low frequencies, which should reduce the smile impression.

Because frequency breakpoints follow formant frequencies in the signal, the output of the frequency warping stage is adaptive to the input signal. This can be used at different time scales: at low adaptation rates, if mean formant frequencies are computed for a range of sentences by a given speaker, the algorithm will adapt to speaker characteristics such as sex or body size (e.g., males have lower, more dispersed formants (Evans, Neave, and Wakelin, 2006)); at faster rates, if formant frequencies are computed for each frame, the mapping will change phoneme per phoneme. In the current implementation, mean formant frequencies are computed for each 1-second sentence in the validation set by averaging the formants over all the harmonic parts of the signal. Formant frequencies are estimated by taking the peaks of the 45-coefficient LPC envelope at a window size of 512 samples and hop size 8 samples (2ms), using the superVP software (Liuni and Axel, 2013) —a non real-time alternative would be to use the formant estimation algorithm from the Praat software (Boersma, 2002).

FIGURE 3.2: Piecewise linear warping function mapping the frequency axis of the input envelope to the frequency axis of the output envelope. This function defines how the segments of the input spectral envelope are warped to the segments of the output spectral envelope. For instance, the segment [F1/2 , F2] will be warped to the segment [F1/2, F2α], which will shift F2 either towards the high frequencies if $\alpha > 1$ or towards the low frequencies when $\alpha < 1$. The same logic applies to all the segments of the piecewise linear function.



3.1.2 Dynamic filtering

As seen in Chapter 2, in addition to warping the signal's spectral envelope and shifting the first formant frequencies, smiling also increases spectral energy in the higher-medium frequency range of the signal (between 1 and 4 kHz for harmonic signals), a frequency area typically associated with F3. To simulate this element of smiled speech, in a second stage of the algorithm, we filter the reconstructed audio signal with an adaptive bell IIR filter which cut-frequency follows the third formant frequency. The filter gain is computed as $g = 20(\alpha - 1)$ dB, which for α in the range $[0.75, 1.25]$ varies from -5dB to 5dB. The cut-frequency refresh rate for the filter was chosen heuristically at 15ms, thus low-pass averaging the formant frequencies extracted at a rate of 2ms in the previous stage.

3.1.3 Special case of non-harmonic frames

Unvoiced phonemes, such as [s], don't have clearly defined formants like voiced phonemes, and when they do, not in the same frequency region. To avoid formant estimation errors, we measure the signal harmonicity frame by frame, using the confidence of the pitch estimation algorithm of superVP. Upon reaching a low-harmonicity frame, neither the frequency warping stage nor the filtering stage update their parameters to the estimated formants of the frame, which are poorly reliable; rather, they use the formant frequencies of the last-seen harmonic frame until a new incoming harmonic frame is detected, at which point continuous adaptation resumes with updated formant frequencies. In addition, in order to recreate the type of resonance seen in Figure 2.1-b, non-harmonic frames are processed with a static filter centered at 6000 Hz with a Q of 1.5 and gain $g = 20(\alpha - 1)$ dB.

3.2 Acoustic performance and latency

3.2.1 Acoustic performance: optimal alpha range

We present here an acoustical analysis of the algorithm performance, to choose the optimal α values, and test whether changes of α produce, as intended, formant movements comparable to those observed in Chapter 2.

We analyze the formant frequencies of a set of 15 French speech sentences (mean duration = 2.3s, $F_s = 44100$), for five manipulation intensities (α=0.8, 0.9, 1, 1.1, 1.25) for which we compute the statistical effect on F1 and F2.

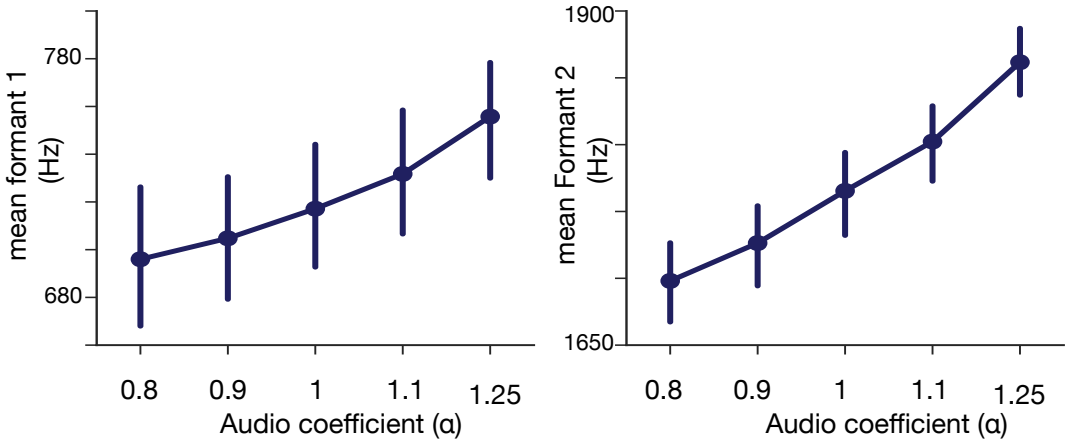
The analysis was done with two one-way, within-sound-files, repeated-measures analysis of variance (RM-ANOVA). Data were analyzed using R (R Development Core Team, 2016), effect sizes are reported as generalized η^2 (Eta-Squared), Greenhouse-Geisser adjustment for sphericity corrections was applied when needed, and corrected p-values are reported along with uncorrected degrees of freedom.

The analysis revealed a significant main effect of the audio coefficient α on F1 (F(4,56)=61.5, p=7.7e-10, η^2 =0.14) and F2 (F(4,56)=137.1, p=4.8e-13, η^2 = 0.5), as illustrated in Figure 3.3, showing that the manipulation shifts formant frequencies as intended. The α value that best simulates the amount of formant movements observed in natural recordings in Chapter 2 (5% for F1 and 4% for F2) is $\alpha = 1.25$, which increased F_1 of 4.8% (from 717 Hz to 756 Hz) and F2 of 3.9% (from 1765 Hz to 1698 Hz). Conversely, for $\alpha < 1$, we observe the opposite acoustic effect—a decrease of formant frequencies—for both F1 and F2. For instance, for $\alpha = 0.8$, F1 and F2 decreased 2.9% and 3.8% respectively (from 717 Hz to 696 Hz for F_1; from 1765 to 1698 for F2). Thus, the range [0.8, 1.25] for α seems to recreate the formant variation range seen in natural recordings.

3.2.2 Latency

As typical for frequency-based digital audio effects, the latency of the algorithm depends on the window size of the FFT. An accurate time-frequency analysis is essential for high quality transformations as it is used to extract both the spectral envelope and the formants in the analysis-resynthesis stage. Here, for a sampling rate of 44100 and for a window size of 1024 samples,

FIGURE 3.3: Formant changes as a function of α. F1 frequency and F2 frequencies averaged over 15 validation sentences for intensities of manipulation α. Error bars represent 95% CI on the mean

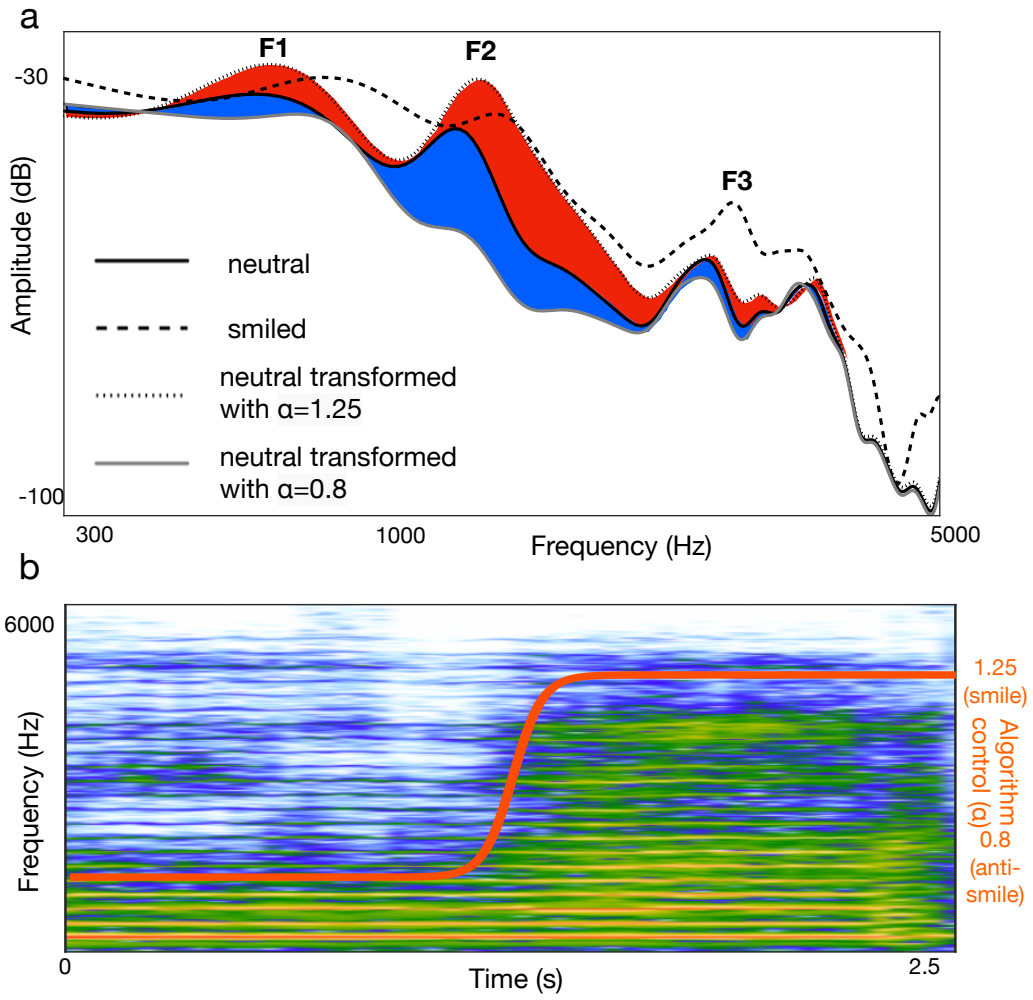


which is suitable for human voice signals, the latency of the algorithm is 75ms. This is satisfactory for real time human-human interactions, but may not be for direct sensorimotor feedback (Aucouturier et al., 2016).

3.2.3 Sound examples

As an example of the overall transformation, figures 3.4-a and 3.4-b present the transformed spectral envelope and spectrogram of an utterance of phoneme [a] for different α values. See SI audio 5 and SI audio 6 for sound examples.

FIGURE 3.4: Examples of the audio transformation. (a) Spectral envelopes of recorded and transformed phonemes [a]: solid bold: original version, pronounced with a neutral tone; dotted bold: original version, pronounced with stretched lips (smiled); dotted light: original version transformed with $\alpha = 1.25$; solid light: original neutral transformed with $\alpha = 0.8$. Red area represents spectral energy added to the neutral spectral envelope when $\alpha = 1.25$; blue area represents energy taken out from the neutral envelope when $\alpha = 0.8$. (b) Spectrogram of a single phoneme [a] transformed with the audio algorithm with a time-varying α (a sigmoid going from 0.8 to 1.25; orange)



3.3 Effect validation (Study 3): Detecting smiles from speech

As a first experimental validation of the algorithm, we conducted a short behavioural study aiming to measure whether naive participants rated transformed stimuli as more smiling than stimuli transformed with the opposite effect.

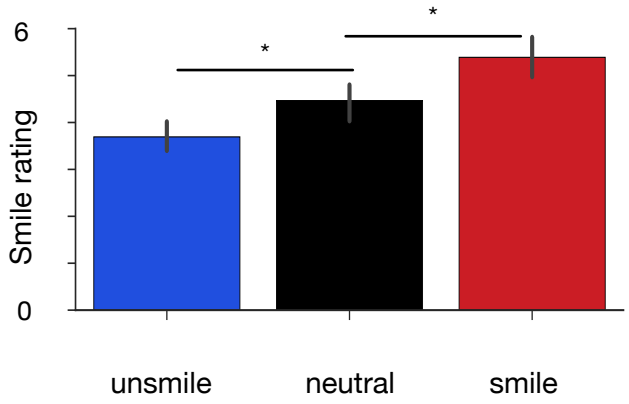
3.3.1 Methods

20 neutral-content sentences adapted from Russ, Gur, and Bilker (2008) were recorded by 10 male and 10 female native French speakers, and transformed using the smile (α = 1.25) and unsmile (α = 0.8) transformations, resulting in 20 neutral, 20 smile- and 20 unsmile-transformed sounds (hear SI audio 5 for stimuli examples; hear SI audio 6 for manipulation examples). Median stimulus duration was 1.86 seconds (SD=0.30 s, min=1.48 s, max=2.60 s). All stimuli were normalized at 70 dbA using the Matlab toolbox Pampalk (2004). N=8 participants (female:8, M=22.5,min=20, max=26) were then presented with the 60 stimuli in pseudo-random order (maximizing the distance of presentation of sentences from the same triplet), and asked to rate "to what extent this sentence [was] pronounced with a smile" using a unipolar continuous scale ranging from 1 ("not smiling") to 10 ("a lot of smile"), with a midpoint at 5. All participants gave written consent and were compensated at a standard rate. The experiment was included into a larger session, in which participants also answered questions about stimulus emotionality and realism, and were tested for EMG reactions in order to pilot the experiments of Chapter 4 —these other tasks are not presented here.

3.3.2 Results

Participants' ratings were analysed with General Linear Mixed Models (GLMMs). We report p-values, estimated from hierarchical model comparisons using likelihood ratio tests (Gelman and Hill, 2007), and only present models that satisfy (1) the assumption of normality (validated by visually inspecting the plots of residuals against fitted values), (2) statistical validation (significant difference with the nested null model) and (3) models which

FIGURE 3.5: Speaker's perceived smiliness is increased by the smile transformation (red) and decreased by the pursed transformation (blue); error bars represent 95% CI on the mean



minimize the Akaike information criterion (AIC, Akaike, 1974). To test for main effects, we compared models with and without the fixed effect of interest. To test for interactions, we compared models including fixed effects versus models including fixed effects and their interaction.

The best-fit model had sound effect (3 levels: smile, unsmile, neutral) as independent variable, and participant, trial number, and sound token as random factors, and was significantly different from the nested null model ($\chi^2(12)=13.0$, p=0.001). As predicted, the smile effect significantly increased perceived smiliness compared to non-modified sounds by about 0.9 ± 0.2 scale points (standard errors; p=0.001; d=1.5; Fig S1-c), while the unsmile effect significantly lowered smiliness by about 0.8 ± 0.2 scale points (standard errors; p=0.01; d=-1.4). Results are presented in figure 3.5

3.3.3 Discussion

Study 3 shows that, as intended, smile-manipulated sounds are perceived as more smiling than neutral and unsmile manipulated sounds. However, the use of an overt explicit task and the lack of a control task do not let us conclude on whether the pattern observed here is simply a demand effect. These limitations are addressed by Study 4 and 5.

3.4 Effect validation (Study 4): Overt imitation

To validate whether listeners spontaneously associate the manipulation to the action of smiling even in the absence of explicit emotional instructions, we asked a separate group of N = 35 right-handed participants (female: 35, M=23, min=18, max=28) to overtly imitate a series of phonemes transformed with dynamically-changing intensities of the smile manipulation, while measuring both their corrugator and zygomatic muscles.

3.4.1 Methods

We recorded 20 phonemes ([a,e,i,o,u]) at a constant pitch, from 4 female speakers, and transformed them using the smile manipulation into two conditions: the "rise" condition shifted from unsmile ($\alpha = 0.8$) to smile ($\alpha = 1.25$) by following a 2.5s sigmoid contour, and the "fall" condition shifted from smile to unsmile following the opposite contour (Figure 3.6-a). All sounds were normalized at 70 dBA.

3.4.2 procedure

Participants listened randomly to each of the 40 phonemes (20 rise and 20 fall) and were asked to overtly imitate them as precisely as they could. Participants were equipped with DPA 4066 omni-directional microphones and Beyerdynamic DT 770 PRO (280 Ω) headphones. Sound was routed via a Fireface UCX USB interface where headphones and microphone were connected. During the imitation, the target sound was played back in order to ensure participants followed the temporal profile of the sigmoid transformation. Critically, no instructions referred to smiling or lip movements and the word "smile" was never used during the experiment. One participant was excluded from further analysis because she didn't complete the task.

3.4.3 EMG recording and pre-processing

Electromyographic (EMG) activity from corrugator supercili and zygomaticus major muscles was recorded during the imitations on the left side of the face at Fs = 1000 Hz. Three filters were used during recording: a high-pass

filter at 10 Hz, a notch filter at 50 Hz and a low-pass filter at 499 Hz. EMG activity was recorded using bipolar electrodes (BIP2AUX adapter), an ActiChamp amplifier and Brainvision recorder software. Synchronization between the stimuli presentation and the recording computer was done via the Cedrus StimTracker serial port. Data was first filtered with a 50Hz high-pass IIR filter and a 250Hz low-pass IIR filter, then segmented into 3.8s epochs (incl. 800ms pre-stimulus baseline). Epochs were rectified and smoothed using a moving average function with a window of 300 ms, and z-score normalized with respect to each trial's baseline. Because participants were explicitly asked to vocalize, no artifact rejection was attempted as the task entailed important muscular activity.

3.4.4 Analysis and results

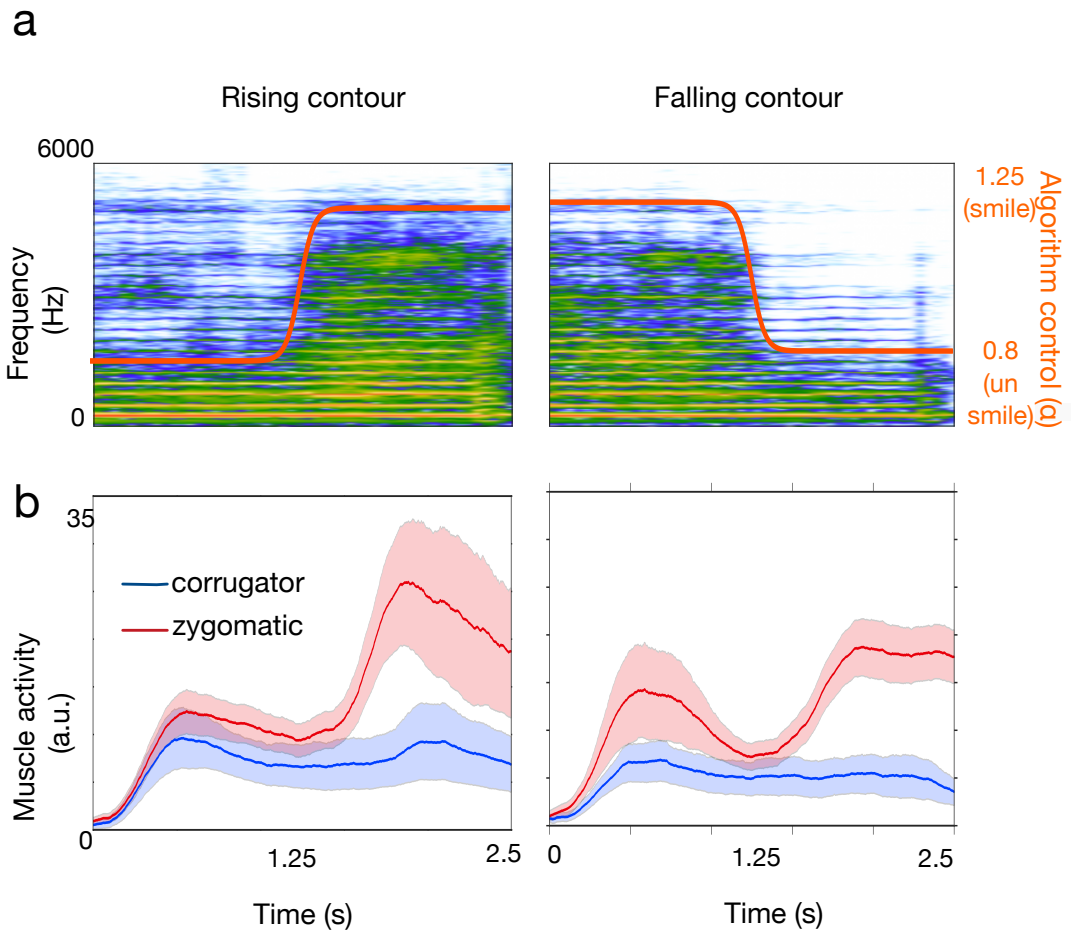
EMG epochs were grouped by muscle, condition (rise/fall) and participant (Figure 3.6-b;). We computed the mean EMG response for both corrugator and zygomatic in three time intervals, corresponding to the start (0.5-0.8s), the middle (1.1-1.4s) and the end (1.8-2.1s) of the sigmoid.

EMG activity was analysed using GLMMs, with participant number, trial number and stimulus as random factors. For the zygomatic muscle, there was a significant main effect of the time interval ($\chi^2(17)$=12.9, p=0.001), but no main effect of condition ($\chi^2(17)$=2.0, p=0.15). Imitators of rise-contour phonemes used significantly more zygomaticus major activity at the end of the imitation (M=24.9) than at its midpoint (M=9.8, paired t(34)=-2.5, p=0.01). Conversely, imitators of falling-contour phonemes reacted with an initial increase of zygomatics major (M=13.7), returning to baseline at midpoint (M=7.5), although the difference was not significant (paired t(34)=1.7, p=0.08). For the corrugator muscle, there was no effect of time interval ($\chi^2(17)$=1.8, p=0.4) not condition ($\chi^2(17)$=2.2, p=0.14, Figure 3.6-b).

3.4.5 Discussion

When asked to overtly imitate sounds manipulated with a rising contour of smile transformation, they produced higher zygomatic activity in the smiled part than in the unsmiled part of the sound. This behaviour is consistent with the idea that the positive part of the transformation simulates the effect of smiling.

FIGURE 3.6: (a) In orange is the α transformation profile used for the rise (left) and fall (right) conditions as a function of time. The parameter α controls the intensity and direction of the voice transformation algorithm; superimposed: spectrogram of a phoneme manipulated with the orange contour. (b) Mean corrugator (blue) and zygomatic (red) activity as a function of time for the rise (left) and fall (right) conditions in the imitation task; shaded areas represent the standard error on the mean.



The converse pattern of zygomatic acitivty was not observed when participants imitated sounds that were manipulated with a falling contour (from smile to unsmile). This difference between how participants imitate sounds in the rise and fall conditions may be linked to how well the starting point of the imitation closely matches the imitator's initial position. While it may be possible for participants to increase 'smiliness' starting from a relatively neutral position (Fig.3.6-left), it may be difficult to decrease smiliness away from an already neutral position in the case of falling contours; in this later case, some participants may e.g. artificially increase smiling at the onset of

the sound so to be able to decrease it after (a possible interpretation of Fig.3.6-right), or use other articulatory strategies than stretching/narrowing lips to attempt to imitate the change.

In sum, at the level of individual phonemes, the smile effect can be perceived as smiling even in the absence of explicit and overt instructions, although some parameters like coarticulation and phoneme play an important role in the perception of the gesture dynamics. It remains unclear from Study 4 whether the smile effect, in certain conditions (e.g. the Fall contours above), may correspond to other patterns of articulation than stretching/narrowing lips. Study 5 will clarify the specificity of the facial action unit (AUs) involved in the effect, by looking at how it is interpreted in terms of facial emotional expressions that include activity inside and outside of the mouth region.

3.5 Effect validation (Study 5): Emotion specificity

To control whether the smile manipulation selectively signals properties of lip stretching/contracting in speech, we asked N=20 participants (female:10, male: 10, M=22.5,min=18, max=28) to rate the same stimuli used in validation 1 (20 non-modified, 20 smile-transformed, 20 unsmile-transformed sentences) on a wider series of 9 emotions and attitudes whose prototypical facial expression involved a variety of facial Action Units (AUs) in the mouth region. Specifically, we selected three emotions all known to involve AU12 (lip stretching), but differing in valence (positive: joy - *French:joie*, interest - *intérêt*; negative: irony - *ironie*), two emotions that involved a deactivation of AU12 / activation of AU15 (lip contraction/depressing) but differed in arousal (low arousal: sadness - *tristesse*; high arousal: upset - *déception*) and four emotions involving other labial gestures such as lip rounding (AU18, skepticism - *suspicion*) and mouth opening (AU25, fear - *peur*, surprise - *étonnement*, anger - *colère*, Ekman and Rosenberg, 1997).

3.5.1 Methods

Stimuli were repeated in 9 blocks, each dedicated to one emotion. The order of the blocks/emotions was randomized across participants. In each block, participants were asked to rate to what extent the stimuli was pronounced with the block's emotion, using a unipolar continuous scale ranging from 0 ("not at all") to 10 ("a lot"), with a midpoint. Instructions about the different scales were given in text, as well as with pictures of prototypical facial expressions.

3.5.2 Analysis

Ratings from the smile and unsmile stimulus categories were normalized by the corresponding non-manipulated stimuli to control for variations of speaker identity, prosody pitch and semantic content. Data was analysed with GLMMs, and post-hocs comparisons comparing smile vs unsmile were conducted with paired t-tests with Bonferroni corrections for multiple measures (alpha=.005). The best fit model, included expression (9 levels) and sound effect (2 levels; smile or unsmile) as independent variables and their interaction. Random factors were participant number and sound token (trial

number was not used as a random factor as each data point was the difference between two trials: the smile/unsmile trial and its correspondent non-modified trial).

3.5.3 Results

We found a significant interaction between emotion and sound effect (χ^2(21)=190.7, p=2e-16). As predicted, the smile effect positively affected ratings of emotions involving AU12 (joy: t(18)=7.0, p=1.3e-6, d=2.2; interest: t(18)=3.5, p=0.002, d=1.1; irony: t(18)=6.5, p=3.8e-6, d=2.0), regardless of their emotional valence. It negatively affected emotions involving AU15 (sadness: t(18)=-4.7, p=0.0001, d=-1.7; upset: (t(18)=-7.34, p=8.1e-7, d=-2.12)) regardless of their degree of control. It had no significant effect on emotions involving AU18 (skepticism: t(18)=-0.4, p=0.6) and AU25 (fear: t(18)=-2.7, p=0.02 $>$ Bonferroni alpha=0.005; surprise: t(18)=-0.5, p=0.6; anger: t(18)=-2.5, p=0.02). Results are presented in figure 3.7.

3.5.4 Photographic credits

Pictures in figure 3.7 adapted from the Extended Cohn-Kanade (CK+) database Lucey et al. (2010): images `S106_001`, `S106_002`, `S106_005`, `S106_006`, `S106_007`.

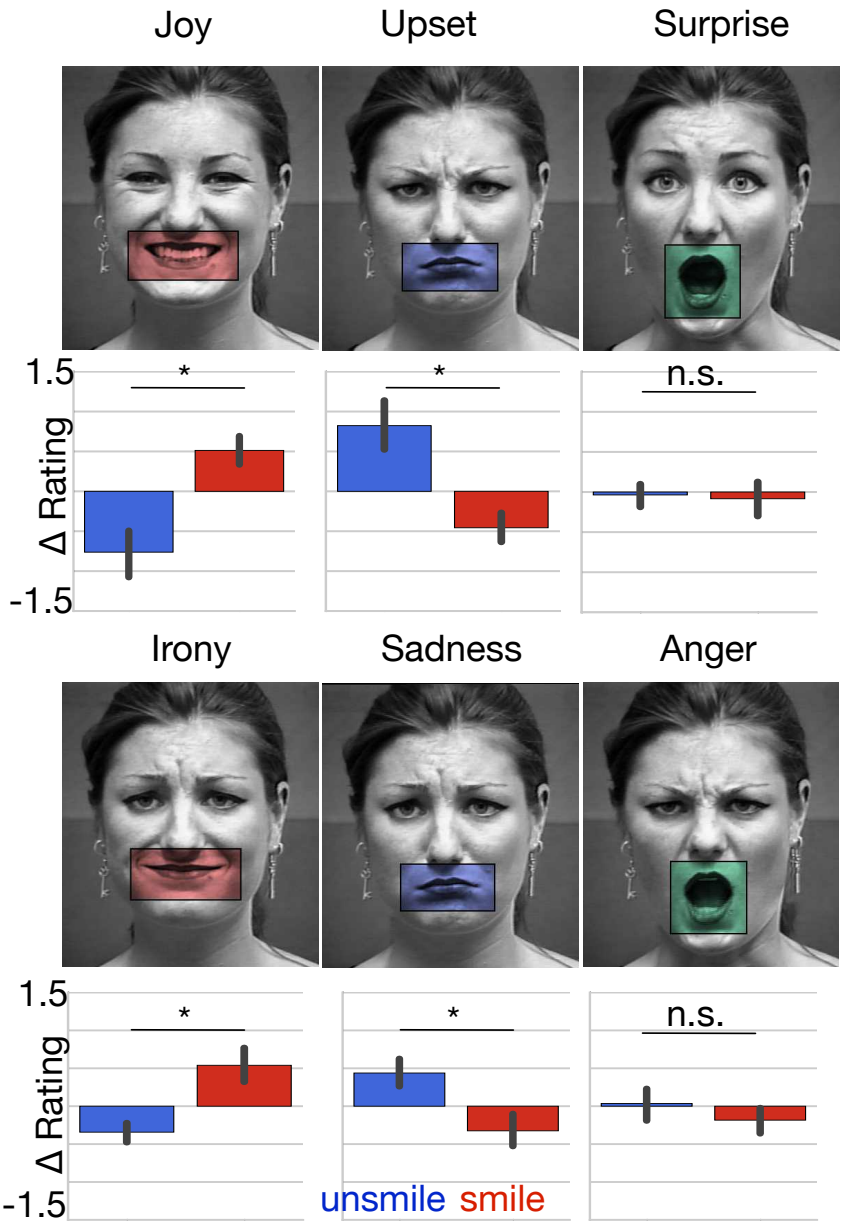
3.5.5 Discussion

These results, and especially the fact that the smile effect increasing perceived irony, demonstrate that the manipulation is not a holistic happy/sad effect, but rather a selective model of the effect of lip stretching/contracting (i.e. facial AU12) while speaking.

Beyond validating the effect, the fact that participants were able to associate a given formantic movement (i.e. here, the specific change of formants implemented by the effect) not to a general gestalt of emotion (positive vs negative, etc.), but to a specific oro-labial configuration shows that the oro-labial characteristics of facial expressions have an important and neglected role in shaping how emotions are signaled vocally. This concerns, here, AU12 as it is used in typical expressions of joy, irony, upset or sadness; recent work has

also suggested that the closing of nasal airflow with AU9 (nose wrinkler) and AU10 (upper lip raiser), typical of the expression of disgust, may also have audible consequences on speech formants (Chong, Kim, and Davis, 2018).

FIGURE 3.7: Mean rating of speaker emotion along six emotional/attitudinal dimensions (joy, irony, upset, sadness, surprise, anger) for smile- (red) and unsmile-transformed (blue) versions of 20 sentence stimuli. Ratings normalized by the corresponding non-modi ed stimuli. Asterisks indicate statistically signi cant differ- ences; error bars are 95% con dence intervals on the mean.



3.6 Chapter conclusion

Modelling auditory smiles

This chapter presents an audio signal processing algorithm able to simulate the sound of smiling in speech stimuli, which we designed as a methodological technique to investigate how participants process auditory smiles in the rest of this thesis. In three validation experiments, we verified that sounds produced with the 'smile effect' had all the desirable properties to be used as stimulus control in the rest of this work: they are recognized as more smiling (Study 3); when asked to imitate them, participants stretch their lips in a pattern ressembling a smile (Study 4); and they increase appraisals of emotions involving stretch lips (joy, irony), decrease those involving narrowed lips (upset, sadness) and leave appraisals of other emotions unaffected.

In the rest of this thesis, we will use this smile effect to investigate whether auditory smiles can trigger facial mimicry (chapter 4), how visual and auditory smiles are cognitively integrated (chapyer 5) and finally, whether reactions to auditory smiles depend on visual experience, in an experiment with blind participants (chapter 6).

4 Embodied mechanisms during auditory smile perception

As seen in chapters 2 and 3, humans rely on robust mental representations to decide whether a voice is smiling or not. These mental representations correspond to specific acoustic changes, which can be modeled and controled computationaly in experimental situations. How are these cues cognitively processed? Are the mechanisms involved in the processing of such features similar to those usually associated to visual smiles?

Embodied mechanisms such as facial mimicry/imitation are considered important components of processing visual facial expressions of emotions. For smiles in particular, these motor reactions are an essential part of current theoretical models of perception (Niedenthal et al., 2010). The aim of this chapter is to investigate whether these facial reactions, which are usually associated to the processing of visual facial expressions, are also recruited during auditory smile perception. The study presented in this chapter has been published in *"Auditory smiles trigger unconscious facial imitation"*, Arias, Belin, and Aucouturier, 2018.

4.1 Study 6: Auditory smiles trigger unconscious facial imitation

4.1.1 Introduction

This study presents an electromyography (EMG) study which aim is to measure facial reactions during the perception of auditory smiles created with the computational model presented in chapter 3.

4.1.2 Methods

Participants

The same N = 35 participants (all female, M=23) as Study 4 (section 3.4) took part in this experiment.

Stimuli and apparatus

Stimuli were identical to the stimuli used in Study 3 (section 3.3), i.e. 20 sentences, each in smile, neutral and unsmile versions. Audio and EMG apparatus were the same as Study 4.

Procedure

Participants rated stimuli along two dimensions, in two blocks. In the first block ("emotion task"), participants heard each of the 60 stimuli and rated "to what extent the sentence's speaker [is] happy?", using a unipolar continuous scale ranging from 0 ("not happy") to 10 ("happy"), with a midpoint at 5. In the second block ("smile task"), participants heard again all 60 stimuli and were instructed to try to recognize the facial configuration of the speaker, specifically "to what extent this sentence [is] pronounced with a smile?", using a unipolar continuous scale ranging from 0 ("no smile") to 10 ("a lot of smile") with a midpoint at 5. Stimulus order was pseudo-random within each block to maximize the distance of presentation between the manipulations of the same recording. Participants who asked about the role of the electrodes were told a cover story explaining that these were sweat sensors. The

placement of the electrodes was verified before the beginning of the first task by asking participants to imitate French phoneme 'i' (for which spreading the lips is necessary), imitating the experimenter gestures, or telling them jokes during the calibration in order to induce smiles. Participants didn't hear the word "smile" until the beginning of the second task and never saw the EMG signals.

EMG recording and pre-processing

EMG recording and pre-processing were the same as in Study 4. Following the method presented in Künecke et al. (2014), we rejected artifacts based on two criteria: first, trials for which the mean activity was larger than three times the standard deviation of the mean activity across trials and participants were discarded; second, trials for which the maximum activity was larger than three times the standard deviation of the mean maximum activity across trials and participants were also discarded. In total there were 8400 EMG recordings (35 participants x 2 blocks x 60 sounds x 2 muscles), from which we discarded 5.4% of the trials (94 and 360 for the mean and max rejection criterions respectively).

Third-party tools

All experiments were developed in Python 2.7 using the Psychopy module Peirce (2008). The MNE package was used for preprocessing and filtering of the EMG data Gramfort et al. (2014). Statistical analyses were conducted using R 3.3.0 (R Core Team, 2015, R Core Team, 2016), using the Mediation package for CMA analysis Tingley et al. (2014). Sound analyses used the Praat software Boersma and Weenink (2017), and sound processing algorithms were implemented in Python using IRCAM super-vp.

Ethics

All experiments were approved by the Institut Européen d'Administration des Affaires (INSEAD) IRB. In accordance with the American Psychological Association Ethical Guidelines, all participants gave their informed consent and were debriefed and informed about the true purpose of the research immediately after the experiment.

4.1.3 Results - Smile task

Rating and acoustic analysis

Formant and rating changes between the unsmile and smile manipulations were compared after normalization by the corresponding non-modified sound using two-tailed paired t-tests between the smile and unsmile distributions (Figure 4.1-a). As predicted, the smile manipulation significantly increased F1 (t(19)=11.6, p=4.6e-10, d=3.6), F2 (t(19)=8.6,p=6.0e-8, d = 2.5) and participants' ratings (t(19)=7.6, p=3.7e-7, d=2.7).

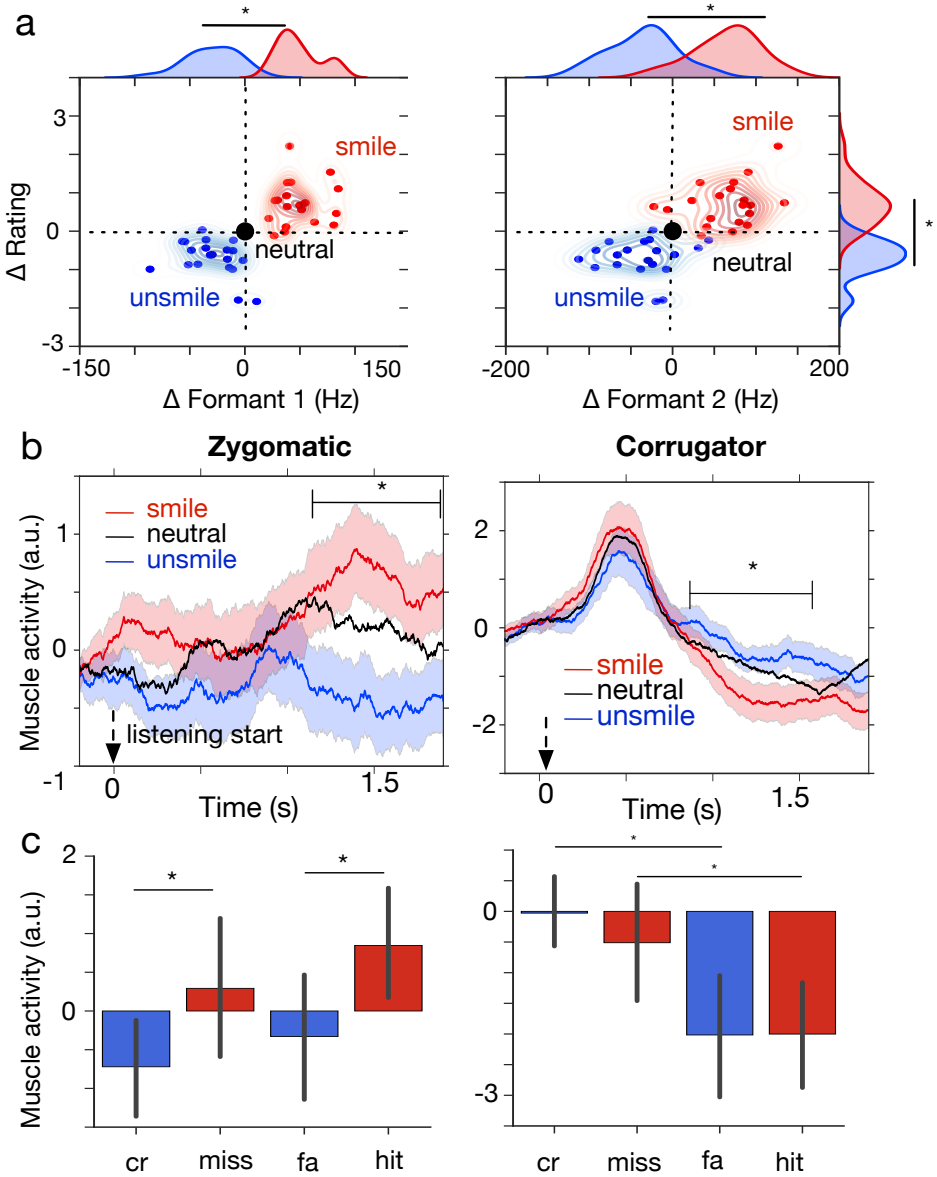
Participants' ratings were analysed using GLMMs following the procedure explained in Study 3 (section 3.3.2). We found a significant main effect of sound effect (3 levels: unsmile, neutral, smile) (χ^2(12)=66.3, p=3.9e-15). The unsmile effect significantly lowered the smile ratings from the non-modified sound by about -0.7 $\pm$ 0.1 (standard errors;p=p=2.9e-11; d=-0.9). Conversely, the smile effect significantly increased the smile ratings, by about 0.7 $\pm$ 0.1 (standard errors; p=9.5e-11; d=1.0) when compared to the neutral sound.

EMG analysis

The difference between smile and unsmile EMG time series (0 to 1.9 seconds) was assessed using cluster permutation tests to correct for non-independence of time-samples and correct for multiple comparisons (Maris and Oostenveld, 2007). For each participant and for each muscle we computed the difference between the mean smile and mean unsmile time series. Clusters were constituted by the consecutive samples that passed a specified threshold of significance (p-value of 0.05; one sample t-test against zero). For each cluster, the sum of the t-values of all the samples was then computed, and compared with the maximum cluster statistics of 5000 random permutations. Significance was assessed using a threshold monte-carlo p-value of 0.05. Effect sizes were computed using Cohen's d, using two-tailed t-tests comparing the average activity in the clusters of interest.

Data is presented in Figure 4.1-b. Smile-transformed sentences triggered larger zygomatic reactions than unsmile-transformed sentences in the second-half of the sentences (t=1.1-1.9sec.;p=0.001; d=0.52), whereas activity

FIGURE 4.1: Results smile block. (a) Mean rating of speaker smiliness for smile and unsmile transformations, displayed as a function of computer-generated changes of first and second formant frequencies (all values normalized by corresponding neutral stimuli). Asterisks indicate statistically significant differences. (b) Participants' corrugator and zygomatic EMG activity while rating speaker smiliness for neutral (black), smile- (red) and unsmile-transformed (blue) stimuli, displayed as a function of time. Asterisks indicate time clusters showing statistically significant differences between smile and unsmile conditions; shaded areas represent the standard error on the mean. (c) Mean zygomatic and corrugator EMG activity while rating speaker smiliness, grouped by signal-detection categories: cr, correct rejections; fa, false alarms; smile-transformed categories are in red, unsmile-transformed categories in blue; asterisks indicate significant difference between response categories; error bars are 95% confidence intervals on the mean.



for the non-transformed sentences was intermediate. The corrugator supercilii activity also differed between the smile and unsmile conditions (t=0.8-1.6sec.; p=0.008). It reacted with an undifferentiated, sudden increase during the initial 500ms; then, during the 0.8-1.6sec period, smile-transformed stimuli triggered lower corrugator supercilii muscle activity than unsmile-transformed stimuli, while the response for non-transformed stimuli was again intermediate. These temporal patterns of zygomatic and corrugator activity were remarkably similar to those normally observed with visual stimuli (Dimberg, Thunberg, and Elmehed, 2000).

Signal Detection analysis

Ratings over the smile- and unsmile-transformed trials were categorized into 'high' and 'low' responses using the median rating value over all trials (M=4.9). 438 unsmile trials with low ratings were categorized as correct rejections; 260 unsmile trials with high ratings, as false alarms; 256 smiled trials with low ratings as misses; 435 smiled trials with high ratings were categorized as hits. All participants had at least one trial for each response category, with a median number of trials per participant of CR = 13, FA=7, hit=13, miss=7. The average hit rate across subjects was 63% (SD=6%, min=52%, max=77%). In each response category (hit, miss, false alarms and correct rejection), we averaged zygomatic and corrugator activity in the last part of the sound (1.3-1.9s).

Activity across categories was then compared using GLMMS (using the method presented in Validation 1). The data are presented in Figure 4.1-c. For the zygomatic muscle, we found a significant main effect of the sound effect ($\chi^2(11)$=6.0, p=0.01) and no effect of the rating category ($\chi^2(11)$=2.5, p=0.1). The smile effect significantly increased zygomatic activity by 0.8 $\pm$ 0.3 a.u. (standard errors; p=0.01; d=0.5). Conversely, for the corrugator muscle, we found a significant main effect of the rating category ($\chi^2(11)$=14.7, p=0.0001), and no main effect of the sound manipulation ($\chi^2(11)$= 1.4, p=0.2). High-ratings significantly decreased corrugator activity by -1.5 $\pm$ 0.3 a.u. (standard errors; p=2e-5; d=-0.8). The nested null models for main effects had either 'sound effect' (when analysing main effect of 'rating') or 'rating' (when analysing main effect of 'sound effect') as predictors. Random factors were 'trial number', 'participant number' and 'sound token'.

Post-hocs comparisons between categories were done using two-tailed paired t-tests. Zygomatic activity was significantly higher for misses than correct rejections (t(34) = 2.1, p=0.039, d=0.39), and lower for false alarms than hits (t(34) = 2.2, p=0.033, d=-0.45). No significant differences were seen between hits and misses (t(34) = 1.17, p=0.24) nor between false alarms and correct rejections (t(34) = 0.78, p= 0.44). In contrast, corrugator activity was significantly lower for hits than misses (t(34)=-2.26, p=0.03, d=-0.47), and for false alarms than correct rejections (t(34)=-3.27, p=0.002, d=-0.72). There was no difference between hits and false alarms (t(34) = 0.05, p=0.96) and between misses and correct rejections (t(34)=-1.35, p=0.18).

Alternative analysis 1: continuous GLMMs

As control, we performed the same GLMM analysis as above, only using continuous participants ratings instead of categories with a median split. GLMMs used the same random factors as above. Ratings were z-scored when needed for model convergence. The analysis yielded similar conclusions: for the zygomatic muscle, there was a main effect of sound manipulation (χ^2(7)=6.6, p=0.01) and rating (χ^2(7)=4.5, p=0.03). For the corrugator muscle, there was a main effect of rating (χ^2(7)= 27.2, p=1.7e-7) but no effect of the sound manipulation (χ^2(7)=0.88, p=0.35).

Alternative analysis 2: causal mediation analysis

Finally, we also conducted a model-based Causal Mediation Analysis (CMA, Tingley et al., 2014), in order to confirm whether there was a direct effect of the sound manipulation (presence of smile spectral cues) on muscle activity or whether muscle activity was mediated by participants' ratings. For each muscle, we fit two linear regression models, one represents the mediator and the other the outcome. The mediator model modeled participants' ratings of smiliness as a function of the sound manipulation (smile/unsmile). The outcome model modeled muscle activity as a function of both the sound manipulation and participants' ratings. For each CMA, sensitivity analysis was performed to check for the robustness of the analysis to the variability in the sequential ignorability assumption. For zygomatic activity, we found a significant ACME (from ratings to muscle activity) of 0.18 (p<0.01) and a significant ADE (from sound effect to muscle activity) of 0.69 (p=0.02). For corrugator activity, we found a significant Average Causal Mediation Effect (ACME;

from ratings to muscle activity) of -0.56 (p<0.01) but a non-significant average direct effect (ADE; from sound effect to muscle activity) of -0.08 (p=0.83).

4.1.4 Results - Emotion task

Rating and acoustic analysis

Participants' ratings of happiness were analysed using GLMMs (see Validation 1 for details on the procedure). As expected, we found a significant main effect of the sound manipulation (3 levels: unsmile, neutral, smile) (χ^2(12)=55.2, p=1.0e-12). The smile effect significantly increased perceived happiness compared to non-modified stimuli by about 0.63 $\pm$ 0.1 scale point (standard errors; p=5.5e-8; d=0.7). Conversely, the unsmile effect decreased happiness by about -0.63 $\pm$ 0.01 scale point (standard errors; p=1.8e-10; d=-0.7; Fig.4.2-a).

Figure 4.2-a presents the rating difference between the unsmile and smile ratings after normalization by the non-modified stimuli (t(19)=8.1, p=1.2e-7, d=2.7) and the also significant acoustic differences between F1 and F2 in the emotion task (sounds are the same as in Figure 4.1-a).

EMG analysis

For each muscle, EMG time series (0 to 1.9 sec.) were analyzed using cluster permutation tests to correct for non-independence of time-samples and correct for multiple comparisons Maris and Oostenveld (2007) using the same procedure as the smile task. For the zygomatic muscle, we found a marginal difference between the smile and unsmile conditions in the segment [1.6;1.9] (p=0.054, d=0.48;). For the corrugator muscle, we found a significant difference between the smile and the unsmile conditions in the segment [0.7;1.9] (p=0.0004, d=-0.62; Figure 4.2.b).

Continuous GLMM analysis

As for the smile task (see above, alternative analysis 1), we performed a GLMM analysis using participants' continous ratings and sound effect as predictors. Participant number, trial number and sentences were used as

random factors. Ratings were z-scored to ensure convergence of the models. For the zygomatic muscle, we found a significant main effect of rating (χ^2(11)=6.05, p=0.01) but no main effect of sound manipulation (χ^2(12)=1.2, p=0.27). For the corrugator muscle, we found a significant main effect of the rating (χ^2(11)=18.77, p=1.48e-5) and no main effect of the sound manipulation (χ^2(12)=2.6, p=0.10). In other words, while data in the smile task establishes that the conscious recognition of a smile is not a necessary antecedent of imitative behaviour, data from the emotion task further shows that attention and/or context (e.g. paying attention to other emotional cues than smile-related changes) can modulate such unconscious processes. These results are consistent with well-known effects of context in the facial mimicry literature Murata et al. (2016) and Cannon, Hayes, and Tipper (2009) and are not further discussed here.

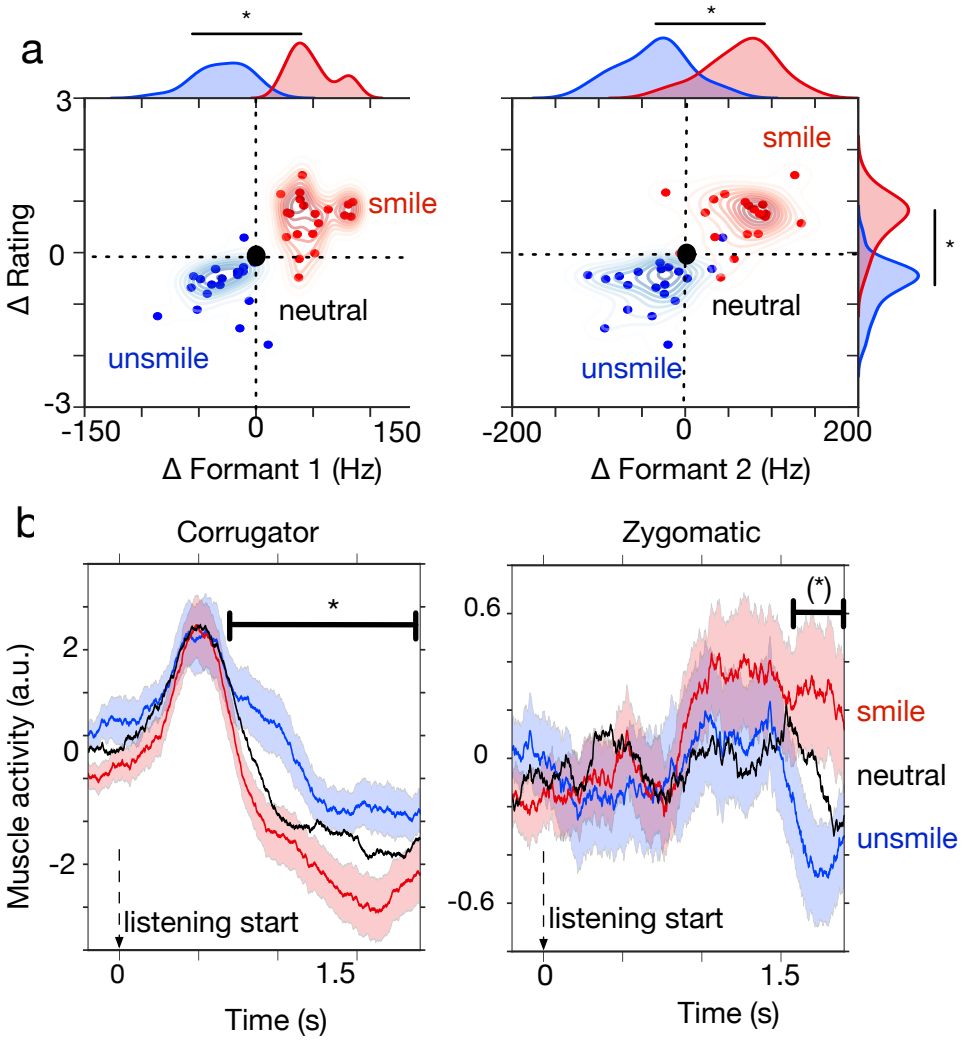
4.1.5 Discussion

Auditory facial mimicry

Using the new voice transformation technique described in Chapter 3, we were able to parametrically control the amount of 'smiliness' in spoken sentences and show that, in the absence of any visual stimulation, these cues were enough to trigger congruent zygomatic (more smile) and corrugator (less frown) activity in adult human listeners. The fact that listeners reacted with more or less smile to stimulus changes designed to reproduce the acoustic consequences of speaking with stretched lips provides a compelling case of audio-based facial mimicry.

Mimicry, the predisposition to mirror a social partner's facial expression and a plausible basis for the human capacity for empathy (Lipps, 1935; De Vignemont and Singer, 2006a), has been almost exclusively studied as a visual-motor process (Niedenthal et al., 2010; Hess and Fischer, 2013). When facial reactions have been observed in response to affective vocalizations (Hietanen, Surakka, and Linnankoski, 1998; Verona et al., 2004; Magnée et al., 2007; Hawk, Fischer, and Van Kleef, 2012), it involved short and stereotypical vocal bursts (e.g. laughter, shouts or cries) which made it difficult to determine whether these reactions were imitations of a facial gesture decoded from the sound, or more simply consequences of the listener's appraisal of the social or emotional significance of the stimuli (Hawk, Fischer, and Van Kleef,

FIGURE 4.2: Results emotion block. (a) Mean rating of speaker happiness for smile and unsmile transformations, displayed as a function of computer-generated changes of first and second formant frequencies (all values normalized by corresponding neutral stimuli). Asterisks indicate statistically significant differences. (b) Participants' corrugator and zygomatic activity, as a function of time, grouped by sound transformation. Muscle activity is measured during listening. Asterisks indicate time clusters showing statistically significant differences between smile and unsmile conditions; shaded areas represent the standard error on the mean.



2012), like smiling when seeing baby animals" as put by (Hess and Fischer, 2013). Here, we exhibit quasi-parametric control over a listener's zygomatic response by manipulating cues that are specific to a smiling gesture in an otherwise unchanged spoken stimulus. In addition, zygomatic activity observed in passive listeners (Figure 4.1) had the same characteristics as the activity produced by speakers when asked to imitate the sounds (Study 4). These results therefore establish that human listeners are able to "reverse-engineer" parts of the articulatory conditions with which speech was produced, specifically here labial gestures based on dynamic spectral cues such as shifts in both first formant positions and in spectral centroid frequencies. This capacity supports "motor-theory" views of auditory perception, which argue that perceiving speech entails perceiving vocal track gestures (Liberman and Mattingly, 1985; Hasan et al., 2016a) and more generally, that listeners are able to situate acoustic signals in a space that captures their distal gestural causes (Kohler et al., 2002; Lemaitre et al., 2017), be it a finger movement, a linguistic gesture or, here, a smile.

The corrugator activity measured here in response to smiled speech is consistent with a positive expression of emotion: the muscle was most desactivated for smile-condition stimuli and least desactivated for the unsmile-condition stimuli. This concurrent involvement of zygomatic and corrugator muscles is typical of most previous studies of facial mimicry in the visual domain (Hess and Fischer, 2013). Similarly, the sudden increase of corrugator response seen here peaking at 200ms post-stimulus was also observed in several previous studies (Dimberg and Thunberg, 1998; Dimberg, Thunberg, and Elmehed, 2000). In these studies, the response was interpreted as an effect of visual stimulation akin to a startle reflex (Dimberg, Thunberg, and Elmehed, 2000), with a plausible source in the orbicularis oculi. The current data shows, as others previously have, that this reflex is in fact amodal and can occur with auditory-only stimulation even at moderate intensity (Blumenthal and Goode, 1991). In addition, because this initial response did not vary between conditions, it seems consistent with the interpretation as an index of general orientation to the stimulus, rather than a valence-specific "probe" response as others have proposed (Vrana, Spence, and Lang, 1988).

There is considerable debate about the cognitive mechanisms underlying mimicry and empathetic imitation (De Vignemont and Singer, 2006a; Hess and Fischer, 2013), and the dissociated behaviors seen here on the zygomatic and corrugator muscles appear difficult to reconcile under any single model.

In the framework of "embodied simulation" for instance (Barsalou et al., 2003; Niedenthal et al., 2010), the processing of a facial expression in one modality (e.g. spectral cues of a smile in the audio modality) is assumed to activate simulations in other modalities (e.g. here, zygomatic-related activation in the motor or somatosensory area), which in turn contribute to the understanding of the original stimuli by "filling in unperceived elements of the original experience" (Hawk, Fischer, and Van Kleef, 2012). If we interpret zygomatic activity on the missed trials as motor simulation which did not accumulate enough evidence to trigger recognition, it appears difficult to explain the lack of zygomatic activity in false alarms —which yet reached recognition threshold with the same mechanism. Rather, the present results raise the possibility that mimicry consists in fact of two mechanistically-separated processes: on the one hand, a high-level cognitive mechanism, which is posterior to the appraisal of the emotional or social significance of the stimuli, and enables motor reactions that were not directly implied by the signal (e.g., here, deactivating the corrugator, a gesture which is not necessary to produce the original stimuli, but congruent with its interpretation as a smile); on the other hand, a lower-level sensorimotor mechanism, which precedes stimulus interpretation and allows the automatic imitation of the gestures specifically involved in the production of the stimuli. The paradigm used in this study, which combines signal detection theory with parametrically-controlled gesture cues in the stimuli, would allow for a confirmation of this dual-pathway in the visual modality, for instance using synthetic facial actions (Jack et al., 2012) instead or in addition to the vocal actions created here.

Unconscious, spontaneous and automatic processes

Even though both zygomatic and corrugator were involved in the response, signal detection and causal mediation analyses both revealed a clear functional distinction between the two muscles: while zygomatic activity was present even when smiles were not consciously detected (responding to missed trials but not to false alarms), corrugator activity was entirely mediated by participants' judgments (responding to false alarms but not to

missed trials). The fact that zygomatic activity continued to track the presence of smile-like spectral cues even when smiles remained undetected suggests that this perceptual process can operate below the threshold for conscious awareness. This behavior reinforces a small but theoretically important set of results showing that facial mimicry in the visual domain can operate on subliminally presented stimuli (Dimberg, Thunberg, and Elmehed, 2000) and even when participants are asked to suppress their own facial reactions (Dimberg, Thunberg, and Grunedal, 2002). The present results show that, even when stimuli are presented consciously and evaluated explicitly, mimicry can still operate on an unconscious level, i.e. take place even if not all of its operations are made available to consciousness (see Lehmann et al., 2004 for a similar pattern of results in the visual domain).

This modularity supports the view that expressive reactions like smiling are fast, innate and automatic "affect programs" (Tomkins, 1962b) which function as the precursors of cognition rather than its consequence (De Vignemont and Singer, 2006a; Heyes and Frith, 2014). That such automatic reactions should result here from auditory-only processing further establishes that there may be nothing primarily visual to such programs. This result could even be taken to support evolutionary theories according to which the smile facial gesture was in fact ritualized on the basis of the adaptiveness of its auditory consequences which, by reducing the length of the vocal tract, would signal smaller size and submissiveness (Ohala, 1980).

It is important to stress that, because participants were explicitly asked to evaluate the amount of smile in stimuli (or, in a second task, to rate their general emotionality), it cannot be assumed that the motor reactions observed in our work occur independently from the task set of attending and processing smile-related acoustic properties of the signal, and that they would occur even in the absence of an explicit task, e.g. pre-attentively. Rather than unconscious, such reactions would more rightfully be described as *spontaneous* (i.e. occurring without any prompt to act upon the stimuli) or *automatic* (i.e. inevitably engaged by the presentation of the stimulus, regardless of any intention on the part of the subject (Hommel, 2007)). In short, our claim is not that the facial reactions observed here occur independently from the *task* of judging smiles, but rather that they occur independently from participants' *judgement* of what is smiling or not. To elucidate whether the unconscious processing of auditory smiles evidenced here is contingent on the pre-established intention or task goal with which it was elicited (something

also called 'conditional automaticity', (Bargh, 1989)) or whether it is truly spontaneous or automatic may require other experimental paradigms than facial mimicry which, in itself, is increasingly considered to depend on the prior establishment of an evaluative context (Hess and Fischer, 2013; Murata et al., 2016). Study 7 below will turn to another implicit measure of processing auditory smiles, eye-tracking, to further investigate this issue.

4.2 Chapter conclusion

Embodied mechanisms during auditory smile perception

Facial mimicry is an important mechanism involved in the cognitive processing of visual smiles, and is at the core of visual smile perception models (Niedenthal et al., 2010). These models suggest, for instance, that facial mimicry reactions are triggered by eye contact. In contrast, we demonstrate here that auditory smiles, even in the absence of any visual stimulus, are enough to trigger facial mimicry that in many points resembles that which is triggered by visual smiles.

Because auditory and visual smiles result from the same motor source (an oro-labial gesture having both visual and acoustic consequences), and because they both trigger highly similar motor reactions during their perception, it is unclear whether visual and auditory cues of smiles are processed by separate unimodal mechanisms, each independly resulting in e.g. mimicry, or whether they are integrated in a joint (multimodal or amodal) 'smile' processing mechanism. To further examine this question, Chapter 5 will use eye-tracking to investigate how the processing of auditory and visual smiles interact during the visual exploration of expressive faces.

5 Processing (in)congruent audiovisual smiles

In the previous chapter, we showed that the unconscious processing of auditory smiles recruits similar embodied mechanisms as the ones usually associated with processing visual smiles. It remained unclear, however, to what extent visual and auditory cues of smiles are integrated in the cognition of ecological situations, and whether these are processed together to create a multimodal or amodal 'smile' concept.

The aim of this chapter is twofold. First, to investigate how smile-cues from the voice and face interact in an explicit emotional rating task. Then, to study whether there are implicit, physiological markers of this cognitive integration.

To do this, this chapter uses the equivalent visual algorithm as the auditory smile effect presented in Chapter 3. Used together with the auditory transformation, this model allows us to create video stimuli that have congruent or incongruent cues in both the audio and visual modality. Study 7 is a behavioral study measuring how such smile-related audiovisual information is integrated in an emotional rating task. Study 8, uses eye-tracking to measure gaze and pupil dilation changes during the perception of congruent and incongruent audiovisual smiles for both emotional and passive tasks. The visual computational model and Study 7 are part of the article *"Realistic transformation of facial and vocal smiles in real-time audiovisual streams"* (Arias et al., 2018).

5.1 Study 7: Emotional rating of audiovisual smiles

Using computational smile effects in both modalities (Chapter 3, and below), we created congruent and incongruent audiovisual smiles in order to study the multimodal integration of these cues in the context of explicit emotional judgements.

5.1.1 Methods

Modeling visual smiles

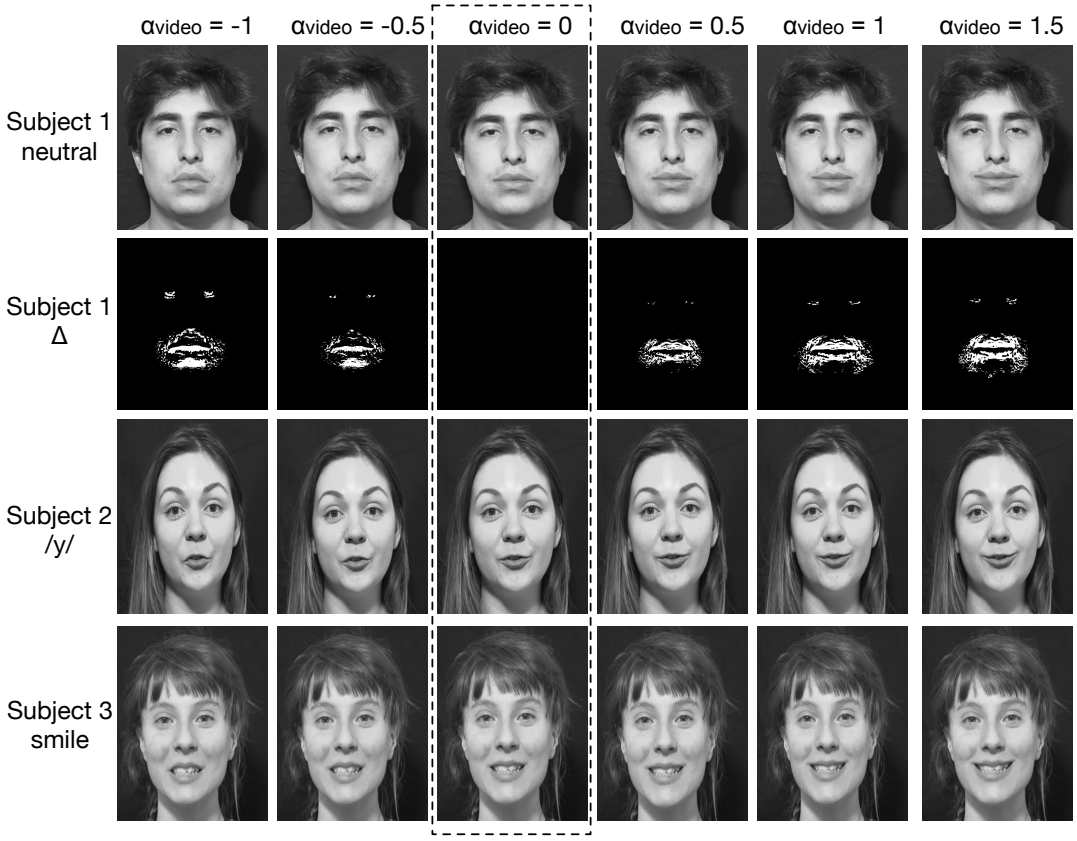
In order to study the multimodal integration of auditory and visual smiles, we collaborated with the FAST team in Centrale Supelec in Rennes, to build the visual equivalent of the audio smile algorithm. The developed visual algorithm is able to transform video of speakers in such a way that they appear more, or less, smiling. This visual algorithm was developed with the same technological constraints as its auditory counterpart: real time, and parametric. Details about the visual algorithm are presented in Appendix A.

Figure 5.1 shows some examples of original ($\alpha_{video} = 0$) and transformed video frames with various positive and negative intensities ($\alpha_{video} = -1..1.5$). The second row shows the absolute difference between the grey-level intensities in the original vs transformed, confirming that modified pixels are found inside the grids around the mouth and eye areas. A detailed quantitative evaluation can be found in (Arias et al., 2018). On a qualitative level, transformations appear plausible even in contexts where the subject is speaking (subject 2) or already smiling (subject 3).

Stimuli

A set of 12 videos (3 males, 9 females) was manipulated using five levels of auditory smile intensity (α_{audio}) and six levels of visual smile intensity α_{video}, conjointly. The audio channels of the 12 videos were transformed at α_{audio} = 0.8, 0.9, 1.0, 1.1, 1.25, for a total of 75 audio stimuli. The video channels were transformed with the visual smile algorithm at 6 levels of intensity α_{video} (-1,-0.5, 0, 0.5, 1.0, 1.5). For each original audiovisual recording, we thus

FIGURE 5.1: Examples of original and modified images with various positive and negative intensities ($\alpha_{video} = -1..1.5$). The original image is either neutral (subject 1) or speaking (subject 2) or already smiling (subject 3). Subject 1 Δ presents the difference between the non-modified and the modified image for subject 1, the black areas show where the image is unchanged, the white areas where the image is transformed



created 30 (6*5) manipulated videos with all the pairs of possible audiovisual manipulations, for a total of 360 rated videos, in which both congruent and incongruent audiovisual smiles are present.

Participants

N=15 participants (M=22, SD=3.6, 8 female, 7 men) took part in this study. Participants were naive to the fact that stimuli may be algorithmically manipulated, gave informed consent and were compensated for their participation.

Procedure

In a single experimental block, participants were presented all 360 stimuli, for each of which they were asked to answer "what is the emotional state of this person?" on a unipolar continuous scale ranging from "negative" to "positive".

5.1.2 Results

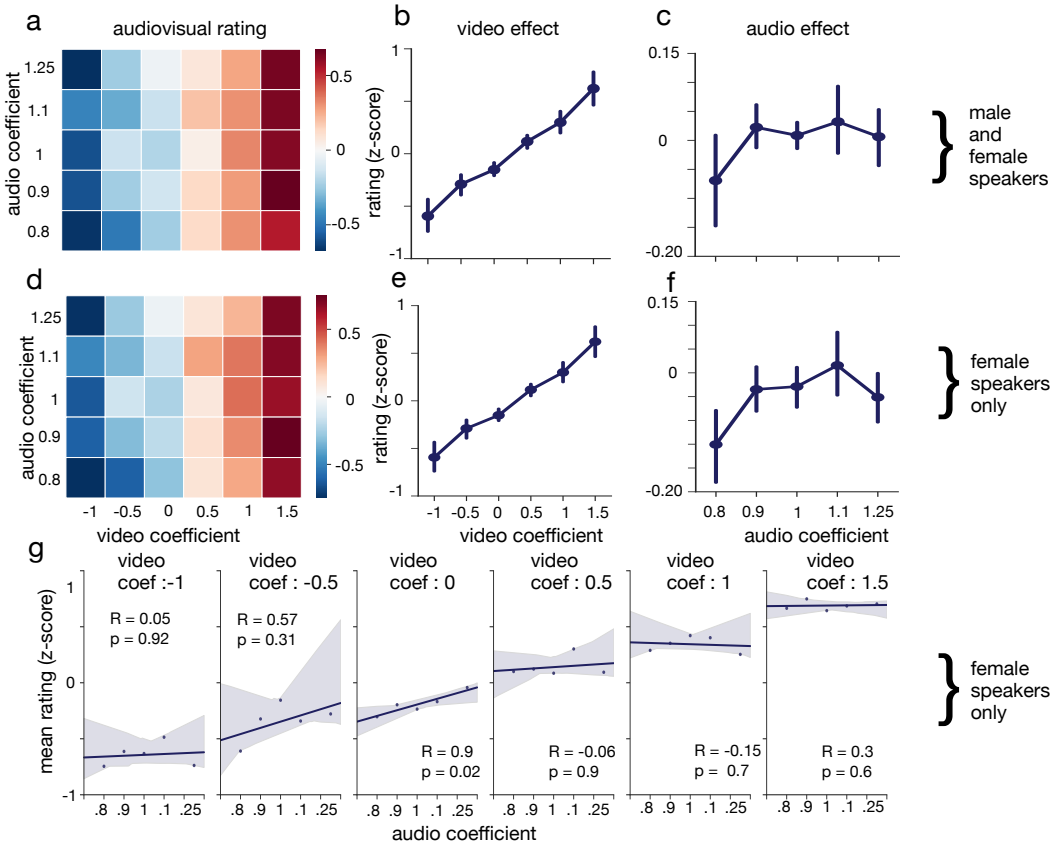
Participants ratings are presented in figure 5.2. Figure 5.2-a presents mean participant rating (z-scored) for each pair of audio and video intensity levels. As can be seen, there was a clear gradient of emotional ratings following the video intensity parameter (horizontal, left to right), but no obvious gradient of ratings following the audio smile intensity parameter (vertical). Figures 5.2-b-c slice through the same data, grouping by separate values of visual smile intensity (a), and audio smile intensity (c). A repeated measures-anova (RM-anova), with two within factors (audio coefficient: 5 levels, and video coefficient: 6 levels), confirmed a significant main effect of the video coefficient (F(5,70)=25.9, p=2.2e-5, η^2 = 0.24) and a non-significant effect of the audio coefficient (F(4,56)=1.6, p=0.19, η^2 = 0.003) on participant rating of the emotional state displayed in these stimuli.

Because the audio smile manipulation was found to operate more strongly on female than male speakers (Arias et al., 2018), we analysed the subset of the current audiovisual data restricted to female stimuli (Figure 5.2d-f). This time, a RM-ANOVA revealed a significant main effect of both the audio (F(4,56)=3.0, p=4.7e-2, η^2 = 0.006) and the video (F(5,70)=23.5, p=8.4e-5, η^2 = 0.3) coefficients on participant ratings of emotion, as well a significant interaction between the audio and the video coefficients (F(20,280)=2.04, p=3.5e-2, η^2 = 0.015). Even restricted to female speakers, the size of the effect of the video transformation (η^2 = 0.3) remained 50 times larger than that of the audio effect (η^2 = 0.006).

5.1.3 Discussion

In sum, while the audio smile transformation is effective in an audio-only presentation (Study 3, 4, 5, 6), its effect is largely overridden by that of the video smile transformation in an audiovisual context. It appears cognitively

FIGURE 5.2: (A) Mean participant rating of speaker emotionality (z-score) in transformed audio-visual recordings as a function of audio and video algorithm intensity (male and female stimuli). (B) Mean participant ratings of speaker emotionality (z-score) in transformed audio-visual recordings as a function of visual algorithm intensity only (male and female stimuli) (C) Mean participant ratings of speaker emotionality (z-score) in transformed audio-visual recordings as a function of audio algorithm intensity only (male and female stimuli) (D-E-F): same data as A-B-C, restricted to female stimuli. (G): scatter plot and linear fit between the audio coefficient and participant ratings, broken down by level of video transformation intensities (female stimuli only). Error bars represent 95% CI on the mean



plausible that visual cues are considered more reliable and salient in judging a given speaker's emotion, and that in some cases, audio cues are only useful when visual cues are ambiguous or otherwise unavailable. The present data supports this interpretation: Figure 5.2-g breaks down the relation between the audio coefficient and participant ratings for the different levels of video transformation intensities for female stimuli. At extreme positive and negative video transformation intensities, the relation between α_{audio} and participant ratings is a flat horizontal line. However, for intermediate α_{video} values (i.e. when positive or negative cues are not, or not as much, available in the visual modality), the correlation between the strength of the audio transformation and the ratings becomes positive and statistically significant (R=0.9, p=.02 for $\alpha_{video} = 0$). This pattern of results, where audio information is used when visual cues are ambiguous, is consistent with previous literature on emotional multimodal integration (Collignon et al., 2008).

The present Study can not conclude whether visual smiles are more salient and important than auditory smiles in the context of emotion rating because auditory and visual algorithm intensities were not perceptually normalised —for instance with Just Noticeable Difference methods. However the data supports the idea that smile-related information from auditory and visual inputs are jointly integrated during explicit emotion tasks. The next study will investigate the extent to which audiovisual integration can be seen at lower levels of cognitive processing, using implicit evaluation methods, namely eye-tracking.

5.2 Study 8: Processing audiovisual smiles, an eye tracking study

5.2.1 Introduction

This study presents an eye-tracking and pupillometry study designed to probe the processing of audiovisual smiles.

Pupil dilation, a sympathetic nervous system response that varies the size of the eye pupil via the optic and oculomotor cranial nerves, is known to index audiovisual congruence during stimuli perception, both for non-emotional (Renner and Włodarczak, 2017) or emotional stimuli (Hepach and Westermann, 2013). Similarly, congruent emotional audiovisual cues trigger significant changes on eye gaze patterns: congruent emotional face and prosody information selectively redirect gaze to the face (Rigoulot and Pell, 2014), and increase face's fixation time (Rigoulot and Pell, 2012; Paulmann, Titone, and Pell, 2012).

The primary aim of this study, following the results of Study 7, is to investigate whether the gaze reactions usually observed during congruent/incrongruent audiovisual emotional processing are also recruited when perceiving (in)congruent audiovisual smiles. Following previous literature, we would predict that (1) pupil dilates when audiovisual smiles are incongruent, and (2) that gaze is redirected towards the eyes when audiovisual emotion information is congruent.

A secondary aim of this study, following the task differences seen in Study 6, is to investigate the extent to which such audiovisual integration processes depend on an overt emotional task. In chapter 4, we found that facial mimicry was recruited in a task evaluative of smiliness, but less so in a task aimed at general emotional evaluation; while results were to some extent independent from judgements (cf. our signal-detection-theory argument), they were not independent from the task. The present study uses implicit pupil dilation measures to probe whether the audiovisual integration of auditory smiles also depends on explicit emotional evaluation tasks.

5.2.2 Methods

Participants

N=30 participants (male = 14, female = 16, mean age = 22, min = 18, max=30) took part in the experiment. All participants gave written consent and were paid at standard rate for their participation. They all reported not having psychological/neurological disorders and no hearing/vision problems.

Stimuli

We asked actors to record a set of sentences in an audiometric cabine, with black background, using a tripod, an external LM400 light, a DPA 44100 omnidirectional microphone and a sony (HVR-Z5E) camera.

Nine videos (min duration=5.2s, max duration=8.0s, mean duration=6.2s) were chosen among the recordings and transformed using the visual and audio algorithms. Each video was transformed with two video transformations: the smile manipulation and the unsmile manipulation. Audio recordings for each transformed video were then transformed with the audio smile and unsmile manipulations. This way, a total of 36 stimuli were generated (9 audio smile, video smile; 9 audio unsmile, video unsmile; 9 audio smile, video unsmile; 9 audio unsmile, video smile), for a total of 18 congruent and 18 incongruent audiovisual articulations (see SI video 1 and 2 for stimuli examples).

To ensure isoluminance conditions across stimuli presentation, the luminosity and colors of the stimuli were kept constant across manipulations. Stimuli across conditions had the same length, the same colors, the same dynamics, the same luminance, the same content, the same prosody, and varied only in the manipulated acoustic/visual dimensions of the voice/face manipulated by the audiovisual algorithm.

Procedures

The experiment was composed of two blocks. The first block was a free viewing paradigm (*passive task*), in which participants were only instructed to watch the stimuli. The second block was a directed attention task (*emotion task*), in which participants were asked to attend to the emotional state of

the person in the videos, although no judgement was required nor collected. Other than in the instructions, block one and two were strictly identical.

The experiment began with a calibration of the eye tracking device (Tobii Pro X3-120), in which participants had to red point moving in-screen. The experimenter explained the first block to the participant, and left the experimental cubicle. After the first block, instructions were displayed on screen to explain the second block. In the second block, participants were asked to attend to the emotional content of the stimuli and were told (deceptively) that questions about the stimuli would be asked at the end of the experiment. Presentation of the stimuli was pseudo-randomised for each participant by maximizing the distance of appearance of stimuli from the same original video. Each stimuli was preceded by a 1.5s black screen with a centered fixation cross.

Participants' dynamic pupil size, number of fixations, and fixation duration were measured using the Tobii Pro X3-120 (sampling rate: 120 Hz). To ensure equal luminance between trials and participants, the experiment was performed in a room without daylight. The artificial lighting was identical and constant for all participants. The experiment lasted around 20 minutes per participant.

Ethics

All experiments were approved by the Institut Européen d'Administration des Affaires (INSEAD) IRB. In accordance with the American Psychological Association Ethical Guidelines, all participants gave their informed consent and were debriefed and informed about the purpose of the research after the experiment.

5.2.3 Pre-processing

Gaze analysis

For each stimuli, we defined 4 dynamic Areas Of Interest (AOIs): eyes, mouth, rest-of-face and background. To distinguish between saccades and fixation we used a I-VT fixation filter (Olsen, 2012). Raw data were exported and analysed in python 2.7. All time points where there was no co-ocurring

presentation of sound and face stimuli (typically, the start and end of each sentence) were removed from the analysis.

The percentage of the number of fixations was computed for each trial as the ratio of number of fixations for a specific AOI, to the total number of fixations during the trial. Similarly, the percentage of fixation duration was computed as the ratio of time fixating one AOI, to total fixation time in the trial. To control for interindividual variability, both numbers of fixation and fixation duration percentages were z-scored (i.e. grouped by participant, AOI and task, substracted with the mean of the group, and divided by its standard deviation).

Pupil size analysis

As for gaze analysis, all data where there was no co-ocurring sound/face signals (at the begining and at the end of the sentences) were excluded from the analysis. Left and right pupil size estimations were recorded at 120 Hz and averaged into a single measure of pupil size. As participants had to actively explore stimuli in dynamic faces, eye movement was important during trials, which added noise to the pupil measure. To reduce this noise, pupil size was downsampled to 1.5 Hz. As above, pupil size measures were z-scored to control for interindividual variablity.

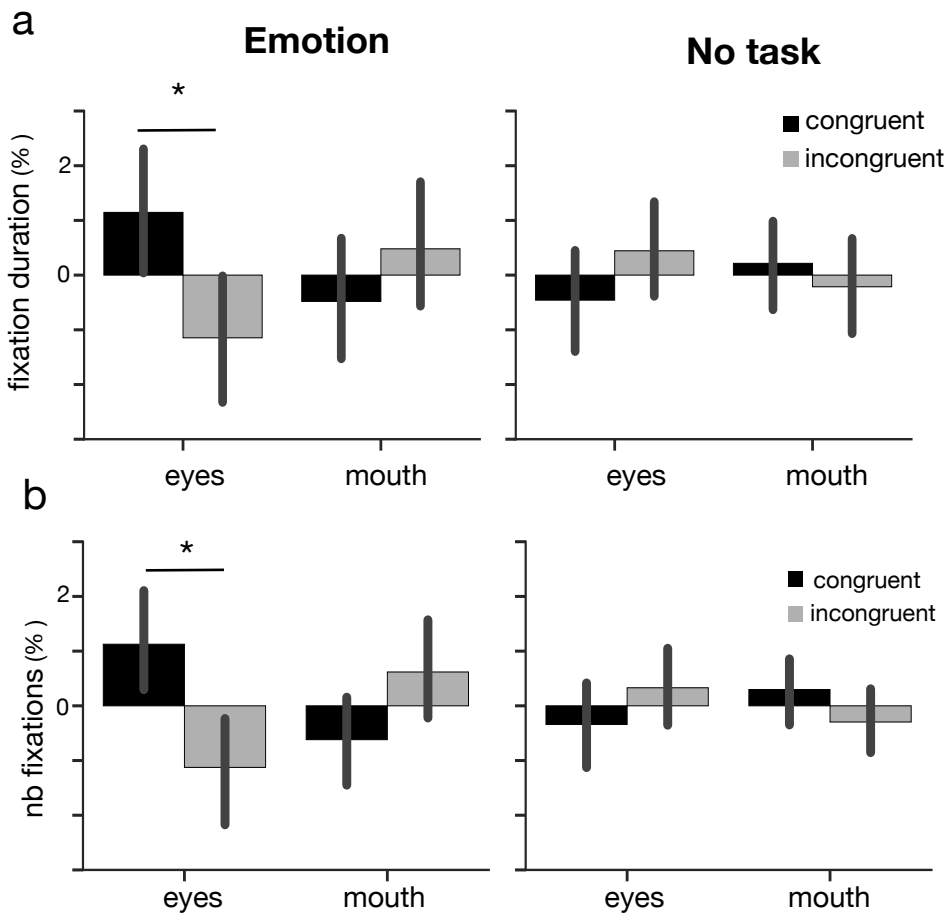
5.2.4 Results

Statistical analyses

Participants' gaze and pupil size were analysed with GLMMs (General Linear Mixed Models). We report p-values, estimated from hierarchical model comparisons using likelihood ratio tests (Gelman and Hill, 2007), and only present models that satisfy the assumption of normality (validated by visually inspecting the plots of residuals against fitted values), statistical validation (significant difference with the nested null model) and models which minimize the Akaike information criterion (AIC, (Akaike, 1974)). To test for main effects, we compared models with and without the fixed effect of interest. To test for interactions, we compared models including fixed effects versus models including fixed effects and their interaction. Post-hoc analyses were computed using two-tailed paired t-tests corrected with

Holm-bonferroni correction for multiple comparisons. We report corrected p-values.

FIGURE 5.3: Gaze Results (a) Fixation duration in congruent and incongruent conditions for each task and each AOI (b) Number of fixations for each task and for each AOI for congruent and incongruent conditions; error bars are 95% confidence intervals on the mean; Asterisks indicate statistically significant differences.



Eye gaze

For each task (emotion and passive tasks) and each gaze measure (fixation duration and number of fixations), we performed GLMM analyses to test for effects of AOI (2 levels: mouth, face) and congruence (2 levels: congruent, incrongruent).

Analyses of the fixation duration in the passive task did not reveal a significant main effect or interactions between Congruence and AOI (Congruence x AOI: $\chi^2(6)=2.2494$, p=0.1337). In contrast, in the emotion task, we found

a significant AOI x Congruence interaction ($\chi^2(6)=7.4154$, p=0.006). Post-hocs analyses revealed that participants eye-fixation time was significantly higher in congruent rather than in incongruent trials (Holm Bonferoni corrected paired t-test; p=0.04; figure 5.3-a).

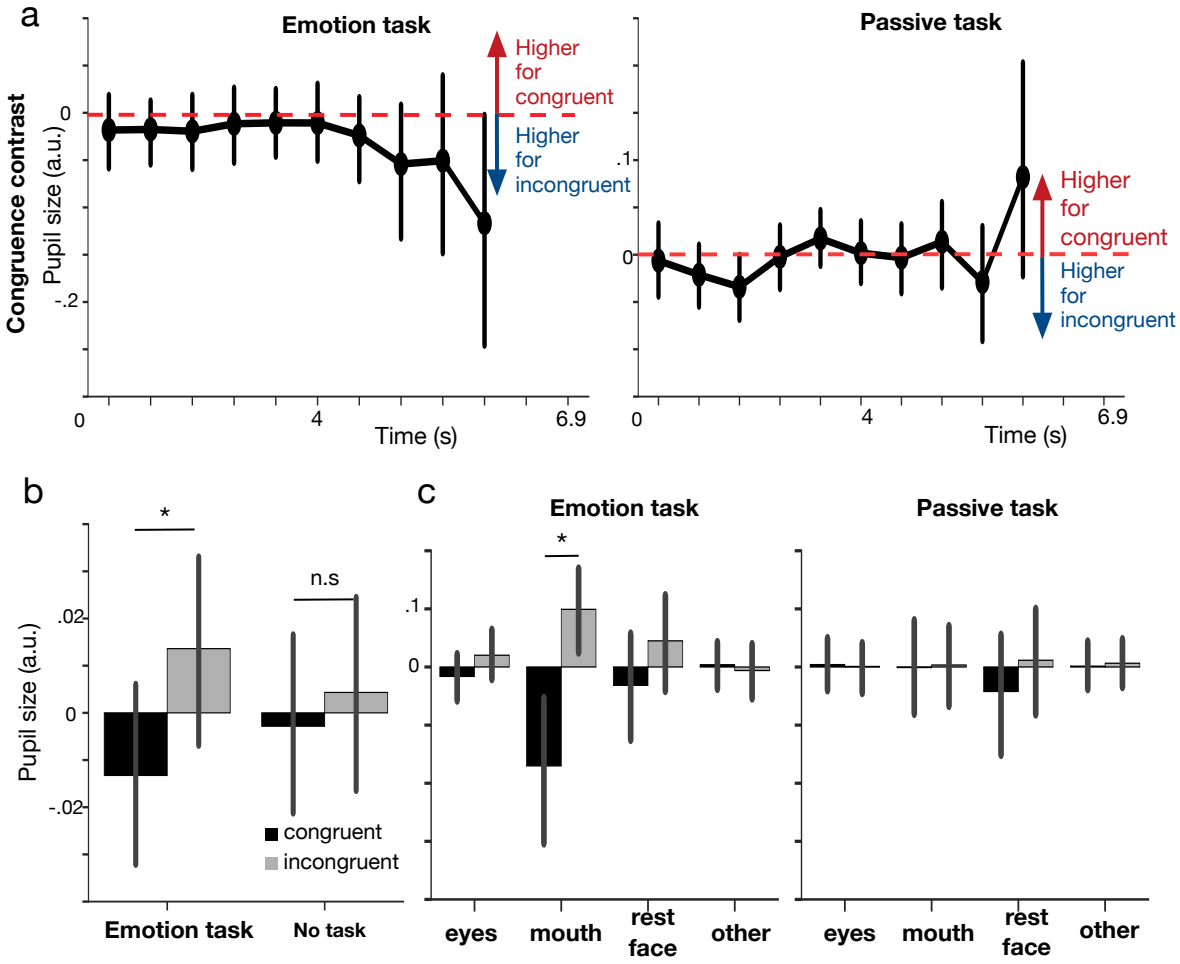
Similarly, for the number of fixations, in the passive task, we found no significant interaction between AOI and Congruence ($\chi^2(6)=3.4825$, p=0.06) whereas this interaction was significant in the emotion task ($\chi^2(6)=13.357$, p=0.0003). Holm-Bonferoni corrected post-hocs revealed that congruent trials significantly increased the number of eye-fixations when compared to incongruent trials (p=0.007; figure 5.3-b).

Pupil size

We computed pupil size congruence contrasts by substracting the time series from congruent and incongruent trials for each group of transformations from the same original token to control for dynamic light and shape variations in stimuli from different videos. Congruence contrasts are presented in figure 5.4-a. Again, GLMM analysis revealed no significant main effect of time in the passive task ($\chi^2(11)=4.84$, p=0.77), but a significant main effect in the emotion task ($\chi^2(11)=16.2$, p=0.04)

After averaging time series across the time dimension, we checked for main effects and interactions between AOI (4 levels: eyes, mouth, rest of the face, other) and Congruence (2 levels; congruent, incongruent). Results are presented in 5.4. Similarly to the gaze analysis, we found no main effect of Congruence in the passive task ($\chi^2(4)=0.38$, p=0.53), whereas this effect was significant in the emotion task ($\chi^2(4)=11.58$, p=0.0007). Further analyses revealed that pupil dilation was affected by congruence dynamically —depending on the area of the face where gaze was positioned. We found a significant Congruence x AOI interaction in the emotion task ($\chi^2(10)=16.09$, p=0.001). Post-hoc analyses revealed that differences in pupil size between congruent and incongruent conditions were stronger when participants' gaze was inside the mouth area (Holm Bonferoni corrected post-hocs; p=0.008; figure 5.4-b).

FIGURE 5.4: (a) congruence contrast: mean congruent minus incongruent conditions time series (z-scores) (b) Mean pupil size (z-scores) for congruent and incongruent trials and for each task (c) Mean pupil size (z-scores) for congruent and incongruent trials and for each task and each AOI incongruent trials for each task; error bars are 95% confidence intervals on the mean; Asterisks indicate statistically significant differences.



5.2.5 Discussion

This study probed the processing of audiovisual smiles using eye tracking, in both a passive task, and a task in which participants' attention was directed to the emotional expression of the stimuli. Although we found no evidence of smile-related audiovisual integration in the passive task, we found significant gaze and pupil changes in the emotion task, which depended on the audiovisual congruence of the stimuli.

As predicted, congruent trials triggered both higher fixation time, and higher fixation duration in the eyes area. These results are consistent with recent litterature on audiovisual emotion integration. In Rigoulot and Pell (2014), for instance, congruent emotional voice-face cues redirected gaze to the eyes; in Rigoulot and Pell (2012), emotional prosody guided visual attention to faces; in Paulmann, Titone, and Pell (2012), participants made longer and more frequent fixations to facial expressions that were congruent with the emotional prosody. Our results replicate these findings, this time with dynamic facial/vocal expressions, and using corresponding smile-related cues in both modalities.

Interestingly, while previous studies have mostly used prosody as the varied acoustic dimension in speech, prosodic dimensions, such as pitch contour, speech rate and content, were here kept constant across congruent and incongruent trials. The only acoustic differences across conditions were spectral manipulations. The fact that similar physiological results are observed both using spectral features or prosody suggest that these gaze reactions are mediated by high-level emotion processing, more than the analysis of individual speech features.

In the emotion task, incongruent trials also triggered stronger pupil dilation. This too is consistent with previous findings. In Renner and Włodarczak (2017), incongruent audiovisual cues (dog meows and cat barks) triggered higher pupil dilation; pupil size also differed when infants observed happy or angry actors performing positive or negative actions (Hepach and Westermann, 2013).

Although pupil dilation indexed audiovisual congruence in the emotion task, this effect disapeared in the passive task. One possible explanation is that pupil dilation in the emotion task reflects cognitive load (Wel and Steenbergen, 2018), indexing the additional cognitive machinery recruited to extract the emotional content in incongruent stimuli compared to congruent stimuli, and that these computations are not involved in the passive task.

The fact that these effects were modulated by the task, even in the absence of explicit participant response, confirms that the physiological reactions that we associate with the processing of emotional facial expression strongly depend on the extraction of emotional cues, and that these reactions do not typically occur spontaneously, but strongly depend on the setup of a previous evaluative task set. This is entirely consistent with the presence of audio-based mimicry in the smile-evaluative, but not the emotion-evaluative

tasks of Study 6, as well as with the previous literature of visual-based facial mimicry (Murata et al., 2016; Hess and Fischer, 2013; Bourgeois and Hess, 2008).

In any case, the present data provides evidence that the specific auditory and visual manifestations of the smile are integrated and processed jointly when extracting audiovisual emotions. The fact that pupil dilation was stronger when gaze was inside the mouth area —exactly where our algorithmic model produces the physical incongruency—is remarkable. This shows how audio-visual articulatory features are dynamically integrated over time during the stimuli presentation.

5.3 Chapter conclusion

Processing (in)congruent audiovisual smiles

This chapter studied how smile information from visual and auditory inputs interact during perception. Using a computational model, we created congruent and incongruent audiovisual smiles. In an explicit rating task, we found that the relative contribution of one sensory input (e.g. audition) is linked to the amount of information of the other sensory input (e.g. vision). Using eye tracking, we then show that congruent voice-face smiles change face exploration strategies, and that incongruent audiovisual smiles trigger pupil dilation, specifically when gaze is directed inside the mouth area.

The processing of auditory smiles, which involves facial mimicry, unconscious processes and an implicit interaction with visual perception, therefore appears to be as deeply-rooted in perception and cognition as that of visual smiles. These results raise the question of the cognitive and developmental primacy of one process over the other. Are cognitive representations of auditory smiles only a by-product of visual experience, learned and made relevant because they co-occur with smiling faces in our usual social interactions ? To investigate this last-pending question, the final chapter of this thesis will investigate whether the auditory perception of the smile can also trigger facial reactions in blind participants.

6 Facial mimicry in the blind

Chapters 3 and 5 demonstrate the depth of the emotional processing of auditory smiles, which in many aspects mirrors the complexity of their visual counterparts. However, it remains entirely plausible from the present results that the cognitive processing of auditory smiles is in fact mediated by visual representations. For instance, in Study 6, participants may activate their zygomatic muscles because they recall what a visual smile would *look* like; incongruent audiovisual smiles may trigger pupil dilation specifically when gaze is in the the mouth region because participants know that the speech they hear does not usually *look* like that. The auditory smile could be, in that respect, a stimulus that remains secondary to the visual smile, only learned because it co-occurs with the smiling face —an epiphenomenon, so to speak. Alternatively, it could also be that auditory smiles are represented independently from visual smiles, and that their cognitive processing develops e.g. autonomously, or in conjunction with the proprioception (but not the vision) of a speaker's own facial muscles.

In this chapter, we recruited participants who have never seen a smile, —and more generally any sort of visual input —, to investigate whether they recognise smiles in voice stimuli, and whether they show the same embodied mechanisms during auditory smile perception as sighted participants. Do blind participants facially imitate smiles heard in speech, despite having never been exposed to a visually smiling face?

6.1 Study 9 - Mimicry in the blind

6.1.1 Methods

Participants

N=14 French-speaking, blind participants (female:5, male:9, Mean age=33.5, min=21, max=58) took part in this experiment. As in Wan et al. (2010), we divided participants in three groups: 5 congenital participants (never had sight), 6 early participants (lost their sight before being 13 years of age) and 3 late participants (lost their sight after being 13 years of age).

Participants were included on the basis of prior medical screening by the chief ophtalmologist (Caren Bellmann, M.D.) of the Institut National des Jeunes Aveugles in Paris, confirming both their vision status and the fact that they had no psychiatric or neurological condition that could interact with the task (such as autism spectrum disorders, a frequent comorbidity with visual impairment; Zafeiriou, Ververi, and Vargiami, 2007). In addition, participants reported having no hearing impairments.

Participant 6 was excluded from all EMG analysis because of technical problems.

Stimuli

40 sentences were recorded by male and female native French speakers, and transformed using the smile and unsmile transformations, resulting in 40 neutral, 40 smile- and 40 unsmile-transformed sounds, for a total of 120 stimulus. Mean stimulus duration was 1.9s seconds (SD=1.4s). All stimuli were normalized at 70 dbA using the Matlab toolbox Pampalk (2004).

Procedure

The experiment consisted of three separate blocks. In the first block participants were presented with the 120 audio stimuli using a beyerynamic DT-770 headphones, and an audio interface. Stimuli were pseudo-randomised (maximizing the distance of presentation of sentences from the same triplet). During this task participants were asked to answer for each sentence *"to what*

extent [was] this sentence pronounced with a smile" using a unipolar continuous scale ranging from 1 ("not smiling") to 20 ("a lot of smile") (*'rating-scale block'*).

In block two participants were presented with the same 120 stimuli as in block one, also in pseudo-random order, but were this time asked to choose, for each sentence, whether the sentence was pronounced with a smile or not in a 2AFC task, which was followed by a confidence judgement ranging from 1 to 4 (the confidence scale ranged from 1: "I am not sure I gave the correct answer" to 4: "I am sure I gave the correct answer") (this block is subsequently called *'detection block'*).

EMG pre-processing

EMG preprocessing was the same as in Study 6 (section 4.1.2).

6.1.2 Results

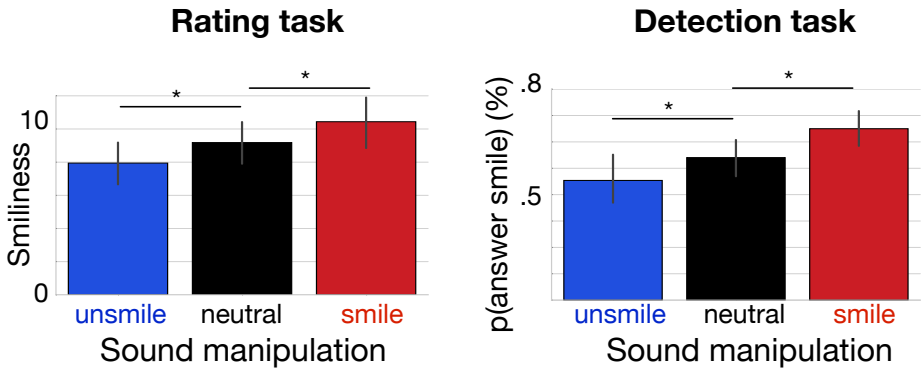
Ratings - group analysis

Participant ratings in both tasks were analysed using GLMMs.

In the rating-scale task, we found a significant main effect of sound effect (3 levels: unsmile, neutral, smile) ($\chi^2(11)=16.46$, p=0.0003). For the best fit model, the unsmile effect significantly lowered the smile ratings from the non-modified sound by about -1.24 $\pm$ 0.30 (standard errors; p=7.85e-05; d=-0.50). Conversely, the smile effect significantly increased the smile ratings, by about 1.26 $\pm$ 0.41 (standard errors; p=0.0089; d=0.47) when compared to the neutral sound. File token and participant number were used as random factors in the GLMM.

In the detection task, we computed the probability of answering "smile" for each transformation category (3 levels: unsmile, neutral, smile). As for the rating task, we found a significant main effect of the sound effect ($\chi^2(5)=35.1$, p=2.38e-08). For the best fit model, the unsmile effect significantly lowered the smile ratings from the non-modified sound by about -0.087 $\pm$ 0.02 (standard errors; p=0.0008; d=-0.55). Conversely, the smile effect significantly increased the smile ratings, by about 0.11 $\pm$ 0.02 (standard errors; p=5.61e-05; d=0.82) when compared to the neutral sound.

FIGURE 6.1: Participant ratings of smile and unsmile-transformed sentences: (a) Smiliness rating for each sound manipulation in the rating-scale task; (b) Smile response probability for each sound manipulation in the detection task. Error bars 95% confidence intervals on the mean.



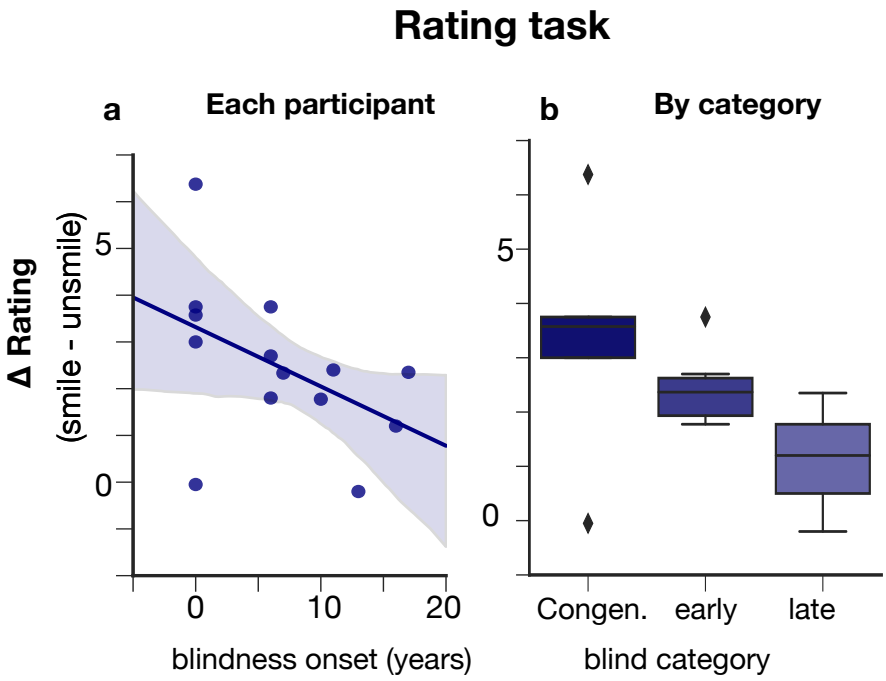
Ratings - individual differences

In order to examine individual differences of ratings within the group, we computed the difference between each participant's ratings of the smile and unsmile effect in the rating-scale task, and correlated it with participant's onset of blindness (figure 6.2-a). There was an apparent negative relation between the rating sensitivity to the effect and the onset of blindness, although the correlation was not statistically significant (p=0.09; r=-0.5). The same data is also represented grouped by participant category in figure 6.2-b.

EMG activity - group analysis

We performed a cluster permutation test for each task and each muscle by taking the mean time series between participants for each condition (figure 6.3). The smile effect had a significant effect on EMG activity in the rating-scale task, where zygomaticus major activity was higher in the smile condition as compared to the unsmile condition (p-value: 0.0416; time: 2.0-2.7; peak: 2.4; Figure 6.3-a). There was no effect on EMG activity in the detection task.

FIGURE 6.2: Influence of onset of blindness on the rating sensitivity to the smile effect. (a) correlation between blindness onset and sensitivity to the smile effect (computed as the rating difference between smile and unsmile) (b) Same data grouped by participant category.



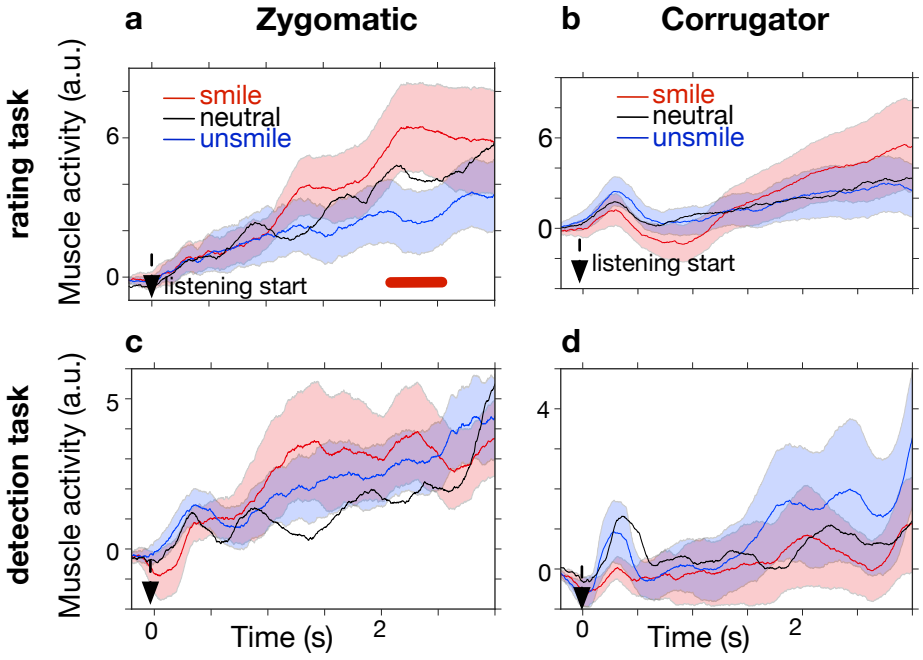
GLMM analysis with continous ratings

As for the data in chapter 4, we performed a GLMM analysis for each muscle using participants' continous ratings and sound effect as predictors. Participant number and sound token were used as random factors.

For the corrugator muscle, we did not find any significant effect. However, for the zygomatic muscle, we found a significant main effect of *rating* ($\chi^2(10)=3.7$, p=0.05), and adding *sound effect* as a predictor to that model significantly improved model's performance ($\chi^2(11)=5.1$, p=0.02).

This shows, as found in chapter 4, that smile specific acoustic cues explain muscle activity even when taking into account participant ratings. In other words, this data replicates the results found in Chapter 4, where zygomatic activity is significantly explained both by rating and by the acoustic manipulation.

FIGURE 6.3: Time series analysis (a) zygomatic and (b) corrugator muscle activity during the sentence listening in the rating-scale task. (c) Zygomatic and (d) corrugator activity during the sentence listening in the detection task. The red line indicates a significant difference between smile and unsmile conditions. Shaded areas represent SEM



EMG activity - individual differences

In order to study individual differences in EMG responses to smiliness within the group, we performed cluster permutation tests for each individual, for both muscles, by considering all 240 trials (both tasks, grouped). The results for the Zygomatic muscle are presented in 6.4, the results for the corrugator are presented in 6.5. All significant clusters are presented in Table 6.1.

For the zygomatic muscle, EMG activity differences between smile and unsmile stimuli were significant at the individual level, in at least one cluster, for 4 participants, all of which were congenitally blind (4/5, 80%). For the corrugator muscle, EMG differences between smile and unsmile stimuli were significant at the individual level in 6 clusters across 4 participants (congenital:1; early:2; late:1). Importantly, all significant clusters across participants and muscles followed the predicted effect direction (unsmile < smile for the zygomatic muscle; unsmile > smile for the corrugator muscle).

FIGURE 6.4: Zygomatic time series for all participants; red bars indicate significant clusters where smile > unsmile; Shaded areas represent SEM.

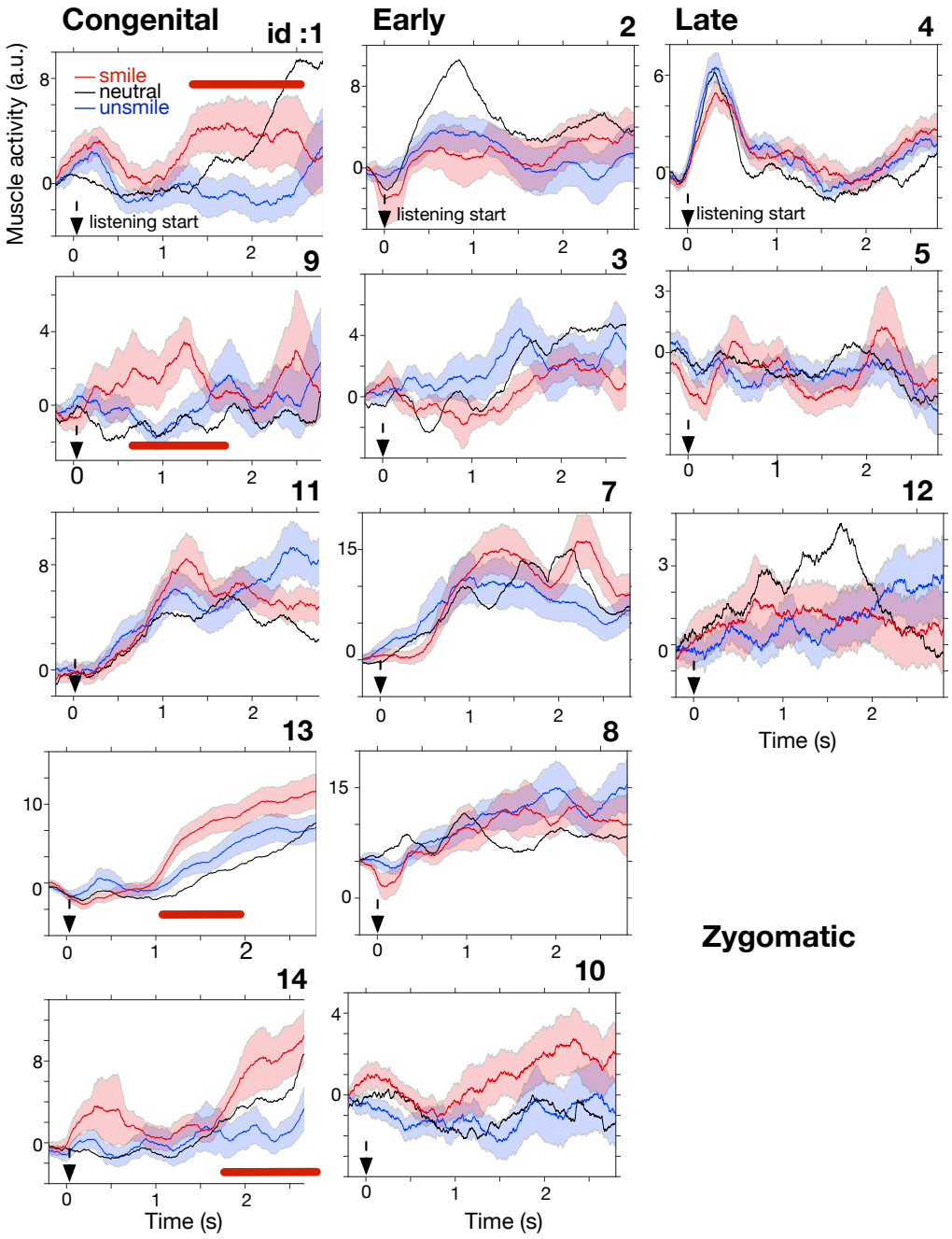


FIGURE 6.5: Corrugator time series for all participants; blue bars indicate significant clusters where smile < unsmile; Shaded areas represent SEM.

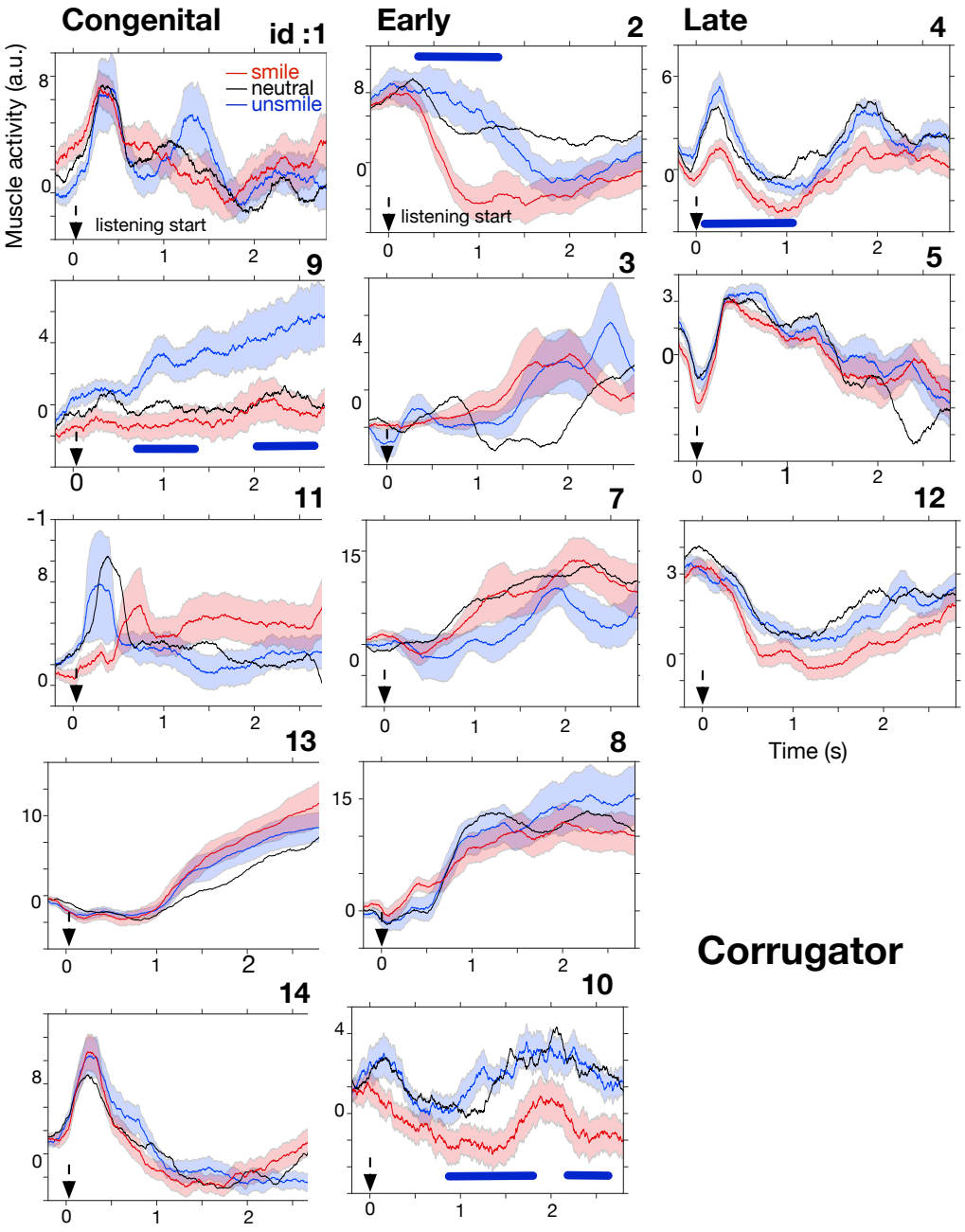


TABLE 6.1: Significant individual cluster permutation tests

Category	Muscle	Participant id	Cluster	P-value	Peak	Direction
Congenital	Zygomatic	1	1.2-2.4	0.02	2.4	correct
		9	0.8-1.4	0.04	1.2	correct
		13	1.1-1.8	0.05	1.6	correct
		14	2.1-2.8	0.03	2.5	correct
	Corrugator	9	0.9-1.9	0.003	1.4	correct
		9	2.0-2.8	0.007	2.4	correct
Early	Corrugator	2	0.5-1.3	0.02	1.0	correct
		10	0.9-2.0	0.002	1.5	correct
		10	2.0-2.8	0.007	2.4	correct
Late	Corrugator	4	0.1-1.1	0.0006	0.2	correct

6.1.3 Discussion

In this study, blind participants recognised auditory smiles both in the rating-scale and in the detection tasks. This confirms that blind participants use facial expressions of emotion as an emotional communication tool, albeit here via the proxy of their corresponding acoustic features. The fact that congenital blind participants recognise smiles in speech is consistent with previous studies showing that both congenital and non-congenital blind athletes spontaneously produce Duchenne smiles when receiving medals or completing intense competitions (Matsumoto and Willingham, 2009). This suggest that blind individuals rely on facial expressions of emotions both spontaneously, and in the context of social interactions, to better understand and communicate affective states, just as for their sighted counterparts, and this even with a congenital impairment that has deprived them of any possibility to learn the visual appearance of a smile.

Second, we observed congruent zygomatic EMG activity during the rating-scale task, indicating that our participants somehow facially imitated the orofacial configuration they heard in the speech stimuli that were presented to them. Because traditional stimuli used in mimicry experiments are all visual, this may be the first experimental evidence of facial mimicry in blind observers, apart from one isolated case of mimicry in a patient with cortical lateral blindness (blindsight; Tamietto et al., 2009). Here, participants had no

visual input to their visual cortices during the experiment and, in the case of congenital participants, had never had any of such input. The theoretical consequences of this result are manyfold, indicating in particular that visual experience is not required to develop the cognitive machinery involved in facial mimicry (Meltzoff and Moore, 1997), and that eye-contact is not a necessary trigger for mimicry (Niedenthal et al., 2010). The precise interpretation of these results, however, is necessary limited by our small sample size. Mimicry was only observed for one out of two muscles, and only in one out of two tasks, and it remains unkwown whether this pattern of result is mechanistically significant, or only the product of weak statistical power.

Despite the low power at group-level, the large amount of trials (240) per subject across both tasks allowed us to perform cluster permutation analyses at the individual level, and therefore treat participants as the experiment replication unit (Smith and Little, 2018). These analyses revealed that four out of five congenital blind participants significantly showed congruent zygomatic activity during auditory smile perception. Moreover, four participants showed significant and congruent corrugator activity during these tasks. Importantly, across muscles, all 10 significant clusters (differences between unsmile and smile conditions) were all in the predicted direction (i.e. muscle reactions were congruent with the auditory condition). These results confirm that blind participants recruit embodied mechanisms during auditory smile perception, despite the information being transmitted by the voice, and despite having limited, or even any, visual experience.

Such individual differences of the recognition and imitation of auditory smiles, linked as they are to the onset of blindness, are particularly telling with respect to the cognitive and developmental primacy of processing visual and auditory smiles. If the cognitive representations and mechanisms involved in processing of auditory smiles develop because of their association with visual smiles, blind participants who have had more exposure to visual displays (later blindness onset time), during a crucial developmental period, should show stronger motor and behavioral effects. In contrast, if embodied reactions to auditory smiles are not learned through visual exposure, but rather develop in parallel or independently to visual social cognition, then embodied mechanisms and discrimination sensibility may be stronger in congenital blind participants, because early blindness onset is typically related to enhanced pitch discrimination (Gougoux et al., 2004), source localisation (Lessard et al., 1998), and overall auditory perception abilities (Wan

et al., 2010). The present data, namely (1) the negative correlation between blindness onset time and smile discrimination abilities, and (2) the fact that the only participants showing significant zygomatic responses at the individual level were congenital (4/5, 80% of the congenital group) provides almost no experimental support to the notion that the processes responsible for auditory smile perception and mimicry need visual experience to develop, and to operate.

6.2 Chapter conclusion

Facial mimicry in the blind

Facial mimicry, the predisposition to mirror a social partner's facial expression, and a plausible basis for the human capacity for empathy and social cognition (De Vignemont and Singer, 2006b), has been almost exclusively studied as a visual-motor process, intrinsically linked to visual attention and eye-contact (Niedenthal et al., 2010). In this work, we have demonstrated that facial mimicry can not only happen with purely-auditory cues indicative of a smiling oro-facial configuration (Study 6), but also that its development and operation does not depend on visual experience (Study 9), because it is present in congenitally-blind participants.

Beyond smiles, these results highlight how important aspects of social cognition, previously thought to be inherently visual, can be shared across modalities, can develop independently from the processes underlying visual emotion processing, and may be accessible to individuals who have totally lacked visual experience.

7 Discussion

7.1 Summary

This thesis focusses on how auditory smiles can be described acoustically, modelled computationally, and how they are cognitively processed in both sighted and blind participants.

1. Using acoustic analyses and psychophysical reverse correlation, we showed that auditory smiles have a specific spectral signature, linked to the displacement and amplification of the voice's first formants (Study 1, 2).

2. We then built a computational model, able to simulate the sound of smile in running speech, leaving unchanged other aspects such as its prosody, speech rate or content (Study 3, 4, 5).

3. Using sounds generated with this computational model, we probed the cognitive processing of auditory smiles. Using facial EMG, we showed that the processing of smile-specific features trigger congruent zygomatic and corrugator activity in listeners, and that these motor reactions are, to some extent, independent of participants' judgement (Study 6).

4. To investigate how visual and auditory smile information interact during perception, we then created congruent and incongruent audiovisual stimuli and showed that, while visual information seemed to prevail in explicit valence ratings (Study 7), incongruent audiovisual information redirected gaze toward the eye region, and triggered larger pupil dilation (Study 8).

5. We then investigated the cognitive origin of the processes involved in auditory smile perception. We hypothesized two possible origins. Either (1) motor reactions to auditory smiles are learned from visual associations, due to the usual co-occurrence of auditory and visual smiles,

or, (2) the recognition and motor reactions to auditory smiles can develop independently from visual experience, in parallel to the visual social cognition machinery. To experimentally distinguish between these two we recruited a group of congenital and non-congenital blind participants and showed significant congruent zygomatic activity to auditory smiles, even though they lacked any form of visual experience of what a smile may look like (Study 9)

Taken together, these results show that empathic responses, such as facial mimicry, previously thought to be specific to visual social cognition, can take place and develop in full in the auditory modality, even in the absence of visual experience. These results highlight the striking human abilities to extract facial/motor information from vocal inputs and use them as communicative tools, both in- and outside of conscious awareness.

7.2 Four (relatively simple) experimental perspectives

Beyond the specific questions governing the experiments reported on here, the audiovisual effects built during this thesis can be used in a wide variety of additional experimental situations, in which they offer a unique way to control stimuli.

First, the real-time audiovisual smile effect could be used in the context of social interactions, in order to investigate how smiling is used in emotion co-regulation contexts or in joint decision making tasks. This tool can allow researchers to study causal relationships in such interactions by letting two participants talk (e.g., by skype) and shifting their voices and faces in congruent or incongruent directions. This procedure can be used to study judgements of willingness to cooperate, or even the impact of emotional processes on group productivity.

Second, the auditory smile effect could be used to manipulate the spectral envelope of musical intruments. Theoretical models of musical emotions suggest that music and speech use similar cues to convey and communicate emotions (Juslin and Laukka, 2003). Our stimuli manipulation paradigm can be transposed to the musical domain to test whether the valence changes induced by the auditory smile effect in the vocal domain correlate to those in musical sounds, and whether similar mechanisms subhold cognitive similarities. Preliminary experiments made in the team in the context of the MSc. thesis of D. Bedoya suggest that it is indeed the case (Bedoya, Goupil, and Aucouturier, 2018).

Third, some results in this thesis can be considered, to some extent, as biomarkers of emotion cognitive processing. These markers (for instance pupil dilation, or facial EMG reactions), could be transposed to populations who are known to have impaired emotional abilities, like populations with Autism Spectrum Disorder or depression. Another intriguing possibility, would be to probe the motor reactions to auditory smiles in patients in altered states of consciousness like coma. Indeed, because we find that some motor reactions to the stimuli are to some extent unconscious, would unconscious patients still show motor reactions? (see also Fiacconi and Owen, 2016)

Finally, is the acoustic consequence of the smile its origin, as proposed by (Ohala, 1980)? A first experiment to investigate this question would be to see whether speakers transformed with the smile effect are perceived as smaller, and whether the changes in body size and the changes in smiling correlate for different algorithm intensities. This would first control that the acoustic changes associated to smiles and body size are both changed in the same way and in the same direction by the same acoustic features. This study could be done both with human and animal vocalisations. Then, using this stimuli, one could investigate human and animal reactions to these cues, to see to what extent they influence dominance perception across the animal kingdom.

7.3 Two (hard) leftover theoretical questions

7.3.1 Are the underlying processes emotional or motor/articulatory ?

The data presented here does not allow us to draw strong conclusions on whether the observed effects are mediated either by emotional, motor-articulatory processes or both. First, motor imitations during auditory smile perception may be due either to the appraisal of the emotional content of the smile (i.e. smiling back because of the emotional significance of that gestural response), or the perception of the articulatory/oro-facial features (smiling back because it is the specific oro-facial configuration that is heard in the sound). Second, during audiovisual perception with eye tracking, we observe that pupil dilation indexes smile congruence strongly when the gaze is inside the mouth area —this can index either the articulatory mismatch of the stimuli, but also its emotional mismatch. In general, the distinction between emotion or articulation is particularly difficult to make in explicit judgements as the correlation between smile ratings and valence ratings is very strong (e.g. correlation between ratings in block 1 and 2 from chapter 4: r=80%; p=10e-24). In short, although our smile audio effect is a computational model of how vocal articulation changes during smile production (and incorporates no element of emotional voice production that is not stricto-sensu implied by the smile gesture, e.g. pitch), there is no way to know whether emotional processes are involved or even required to elicit the effects that we report on here.

These two alternatives (emotional or articulatory processess) are confounds that span the whole facial mimicry and audiovisual emotion integration literature. While facial mimicry is often interpreted as a mirroring phenomenon, it is rarely the case that a simpler, emotional-contagion mechanism can be ruled out ("smiling when looking at pictures of cute kittens", as put by Hess and Fischer, 2013). The fact that both mimicry and emotional congruency effects have been reported using non-smile-related (or further, non muscle-specific) vocal cues such as positive affective prosody (Hawk, Fischer, and Van Kleef, 2012; Rigoulot and Pell, 2014), suggests that these effects can be mediated as much by higher-level emotional processes than by lower-level sensorimotor reactions.

While not conclusive on that aspect, the signal-detection-theory analyses of chapter 4 and 6 may provide a methodological way to advance on this issue. In our models, while corrugator activity was entirely explained away by participant judgements, zygomatic activity was explained both by judgement and the acoustic features and, as such, may be mediated by different mechanisms: both a low-level "sensorimotor" process able to track articulatory acoustic features and operating somehow outside of conscious awareness (so that zygomatics react to missed trials more than false-alarms), and a higher-level process linked to emotional recognition, which follow false alarms, and maybe the only one active in the case of corrugator responses. These mechanisms may happen in parallel and explain, each, some amount of the motor reactions observed. One interesting experimental possibility to further explore this dissociation in the future would be to create positive and negative stimuli either by generatingsmile/unsmile transformations, or prosodic modifications, which also signal positive affect (for instance with pitch manipulations; Aucouturier et al., 2016) —one prediction would be that non-smile related prosodic cues can trigger facial reactions, but only those triggered by smile-specific spectral cues would remain unconscious.

7.3.2 Is task-dependency theoretically important?

Perhaps the most recurring pattern of results in all effects reported on in this work is their critical dependence to the task. In chapter 4, motor reactions to auditory smiles were stronger for tasks in which participants directly evaluated smiles, rather than their implied emotional valence. In chapter 5, pupil dilation and gaze only indexed audiovisual smile incongruency when participants' attention was directed to the emotional valence of the stimuli, even in the absence of explicit response, and the same stimuli in a passive observation task did not trigger the same physiological reactions. Finally, in chapter 6, motor reactions were observed in blind participants in a rating-scale evaluation of smiles, but not in a detection task (same question; same stimuli; different rating scales). While similar discrepancies are reported in the literature (e.g., for facial mimicry, Cannon, Hayes, and Tipper, 2009; Murata et al., 2016), they are often cast aside as a contingency, if not (we suspect) buried-away in the file-drawer of unpublished pilot studies (Rosenthal, 1979).

These differences can be caused by several factors. First, it is possible that the more explicit the task set, the more attention and, generally, cognitive

resources are involved in the parsing and processing of specific vocal/facial features. For instance, in Study 9 on blind participants, we observe effects when participants rate the stimulus in a scale ranging from 0 to 20, but not in a 2AFC procedure. Rating on a continuous scale may subhold a more precise parsing of the acoustic features, which entails more important facial mimicry. Similarly, in study 6 on sighted participants, we observed stronger zygomatic activity when participants were explicitly asked to rate smiles, rather than rate stimulus valence. Although no appropriate statistical comparisons between the tasks were presented here, cluster permutation tests show a more important cluster in the overt smile task. Finally, in Study 8 with eye-tracking, again attracting attention to the emotion cues in audiovisual smiles was enough to make pupil size and gaze index the audiovisual congruence of the smiles. These three examples suggest that attracting attention to specific emotion cues is essential for these facial reactions to take place, and that the more participants' attention is focused on these emotional cues, the more important are the physiological reactions.

These reactions may be mediated by the mental representations activated during the task. Indeed, when changing the task set, experimenters can to some extent control the cognitive processes involved in the parsing of auditory cues. If the task is a smile recognition task, participants will need to activate smile-related mental representations, possibly linked to motor-articulatory representations of how a smile is produced, sensorimotor representations of what a smile feels like, or to visual representations of what a smile looks like. If the task is totally passive, participants may not need to summon these representations. This is mechanistically different from the effect of attentional focus in the previous paragraph because of the causality of perceptual events. If attention is the main trigger of physiological reactions, focusing participants' attention on these cues even without an explicit evaluation task should trigger facial reactions, conversely, if mental imagery is crucial, emotional task sets are essential for these reactions to happen.

Third, task differences could be explained by "mechanistic satiation". For instance, in the experiment with blind participants (chapter 6), we observe the effects in the first task, but not in the second. Importantly, the order between the tasks was not counterbalanced, in order to keep the first task comparable to the task in chapter 4. Maybe we do not observe the effects at the group level in the second task because of cognitive fatigue —this experiment was

particularly long because we wanted to have enough trials to perform individual statistics.

Fourth, these reactions may be mediated by demand effects. In most of the experiments reported on here, behavioral responses co-occurred with motor activity. One possibility is that the preparation to respond "smile" is enough to trigger motor reactions. Although this interpretation is not supported by analyses showing that facial reactions can be independent from judgement (GLMM analysis with continous ratings in Chapter 4 and 6), as well as by other studies on visual facial mimicry (Tamietto et al., 2009; Dimberg, Thunberg, and Elmehed, 2000), it cannot be excluded that demand effects mediate, to some extent, some of these motor reactions. If anything, the fact that these reactions are observed across visual and auditory modalities, using the same task across modalities, suggests that the processes involved in mere responding to such tasks are important in the elicitation of the motor reactions.

Experimentally disentangling between these four possibilities is not straightforward. In the literature, discrepancy between tasks is often theoretically grouped under the umbrella of interpersonal and intergroup variations, to suggest a general influence of social context on facial mimicry (Hess and Fischer, 2016). In our view, this can be a slippery slope. Although it is attractive to group task and group differences effects into a common theoretical cocoon, the mechanisms underlying such differences can be different, and vary from simple demand effects to actual functional adaptations.

Similar discussions are ongoing in the field of facial mimicry. At this point, the results presented here can only acknowledge that similar patterns of task dependence are seen in the auditory modality, report that these can develop independently from visual experience, generalize them to other autonomic responses, but cannot yet draw any firm experimental nor theoretical conclusions on this matter.

7.4 Working with parametric transformations: some methodological after-thoughts

Across the 9 studies reported on here, this thesis used a relatively novel experimental methodology, namely, that of building computational models to create audio and visual stimuli. In fact, we used three different transformation algorithms. First, in Study 2 (reverse correlation), we created random changes in the spectral energy of a phoneme to generate a large number of frequency variations in order to uncover participants' mental representations of what a smile sounds like. Second, in chapter 3, we developed a 'smile transformation' audio effect, able to simulate the spectral cues of smiling in speech, and used it to create stimuli for mimicry and audiovisual integration studies (Studies 6 , 7, 8 and 9). Finally, we built a visual smile effect which, together with the audio effect, was used to construct audiovisual stimuli for studies 7 and 8. Such use of computationally-generated stimuli is not a broadly used technique in the experimental sciences and it involves a certain methodological *savoir-faire*, which I learned by trial and error during this thesis. As a closing methodological note to this work, I give here some reflections on the pros and cons of this technique.

7.4.1 The case for transforming (rather than creating) stimuli

While typical research may use pre-recorded actor vocalisations, or attempt to synthesize whole stimuli, the different techniques used here involve transformations, i.e. taking an existing input sound, and modifying it into an output with desired acoustic features. This has several advantages.

First, digital transformations can reliably conserve the original stimulus characteristics. In a transformation framework, the experimenter needs to transform only a few features in the stimulus, whereas with synthesis techniques the stimulus usually has to be recreated from scratch. Thus, transformations can result in highly natural and artefact free stimulus.

Second, when working with transformations in an experimental design, researchers have to deal with a two-stage process: (1) Choose what stimuli to transform and (2) transform them. This means that researchers have at least two independent levers, one for each stage. For instance, if I want

to test whether auditory smiles interact with semantic information (showing e.g. ventral and dorsal stream interactions), a potential experimental design could be to record words with positive and negative meaning (first stage), then transform these words with the smile algorithm (second stage) and then test, for instance, whether participants recognise semantic information as positive or negative in a faster or slower fashion depending on the acoustic manipulation. Here the lever in the first stage can be replaced by a large amount of variables, depending on the research question, making this methodology highly modular.

Third, the algorithmic architecture of transformation algorithms is suitable for real-time applications. The computations in such algorithms are usually performed for incoming data chunks, in the form of buffers for audio streams, or in the form of frames for video streams. As such, these transformations can be used in real-world situations like social interactions. Moreover, because of the simplicity of the input and output routing (e.g. audio-in, audio-out; video-in, video-out), transformations can be chained together —as we did for visual and auditory transformations in chapter 5, or as is done in the CLEESE tool used in chapter 2 for reverse correlation purposes; Burred et al., 2018).

Finally, transformation algorithms can usually be shared easily between research groups to support replicability.

7.4.2 The case for parametric-control over the stimuli

Another important constraint of algorithmic transformations in experimental contexts is the ability to change the transformation in a parametric fashion (e.g. with increasing levels of intensity). This is an important feature for different reasons.

First, having parametric control over the stimuli features has the advantage of ensuring that between different levels of transformations, the stimuli only change in one physical dimension. This represents an important level of control compared to e.g. actor vocalisations. This allows researchers, for instance, to integrate the physical variations triggered by the transformation in the data analysis stage to better depict underlying mechanisms, without having the confound of multi-dimensional feature variation.

Second, if the transformation is thought to change one specific percept, the intensity of the algorithm can be correlated to a measure (accuracy, intensity, reaction time etc.) of that percept, (as done in Study 7).

Third, and most importantly, transformation levels of intensity can be used to make predictions and hypotheses. The ability to predict how a specific behavior or electrophysiological reaction will co-vary depending on the intensity of a specific physical characteristic of the stimuli, matches well with the hypothesis-driven ideal of experimental research and can lead to fruitful mechanistic insights at the analysis stage. For instance, in chapter 4, we show a functional dissociation between zygomatic and corrugator muscle activity, which was observed because the muscles reacted differently to the acoustic cues manipulated by the smile algorithm.

Fourth, a parametric transformation allows researchers to go deeper in the description of the stimuli (e.g. how many pixels change across conditions? Where?), in order to better understand what the underlying mechanisms mediating the observed effects are. For instance, we saw in chapter 5 how pupil dilation was stronger during incongruent trials when participants' gaze was inside the mouth area, exactly where the articulation incongruence was created in stimuli. Having control over these features of the stimuli allowed us to draw deeper conclusions on the nature of the cognitive processes responsible for these reactions, and overall, to give more precise interpretations.

Finally, parametric manipulations can be easily deployed across several stimuli like e.g. cross-linguistic databases, or even in comparative approaches (animal vocalisations, musical instruments, etc..), while maintaining comparable stimulus characteristics.

In sum, working with parametric transformations allows researchers to focus on a few set of finely controlled features, rather than on a wide mixture of co-varying physical signals. This method can be used to draw causal links between physical intensities in the stimuli and psychological measures.

7.4.3 A step-by-step guide to work with parametric transformations

Through this thesis, I developed a step-by-step methodology in order to work with algorithmic transformations such as the 'smile effect'.

Step 1: what model to develop?

Here are some characteristics of a parametric transformation which are important to consider when working on a specific scientific question. First, parametric transformations allow researchers to draw stronger causal inferences if they are bidirectional. When choosing a model, it is useful to choose a dimension which can be increased or reduced. For instance, with the auditory smile manipulation, we can either increase or reduce how the contraction of AU12 changes speech. This allows experimenters to test for the directionality in behavioral and electrophysiological changes. For instance, in study 3 smiliness ratings increase when using the smile effect and decrease when applying the unsmile effect, as compared to the non-modified sound.

Second, consider the algorithmic directional changes as predictions (e.g. the smile effect will increase zygomatic activity, the unsmile effect will reduce it). This can be applied to several domains. For instance, in research studying how individuals compensate pitch during pitch-shifted auditory feedback, it is useful to have both increases and decreases of pitch information in order to have both effect directions (see e.g. Behroozmand et al., 2012), which allows us to draw stronger conclusions and have overall, more precise mechanistic insights. If the manipulation intensity is indeed causally involved in the cognitive processing of the algorithmically-manipulated features, the direction of the effects should closely match the predictions. If directions are at odds, either the transformation model, the predictions (or both) are wrong.

Third, the feature space used for the manipulation is crucial. For instance, in the reverse correlation study (Study 2), we used random spectral content variations, as opposed to simply adding white noise as done in vision (Gosselin and Schyns, 2003). This allowed our manipulation to be cognitively integrated as part of the speaker production, rather than perceived as a separate source in the audio stimuli. In order to correctly choose the feature space, it is useful to be inspired by natural phenomena that have a ground truth independent of perception like e.g. pitch height, face articulation, vocal roughness, gender, body size. These dimensions have a physical reality and can as such be modelled independently of perception by studying only their physical characteristics. The feature variation patterns can model the observed changes in stimuli (e.g. male voices have greater formant dispersion and lower pitch than female voices). When the feature space and the parametric variation used by the transformation closely match those of the

natural phenomena, it is easier to study the underlying mechanisms used for their perception.

In sum, parametric transformations of stimuli can be used to draw causal inferences, just as the neurosciences use transcranial magnetic stimulation (TMS). If I increase feature X, I predict I will observe more of Y; but also, if I decrease feature X, I predict I will observe less of Y.

Step 2: validating transformations

Two things have to be validated in a model for it to be useful in an experimental situation. First, the algorithmic output of the model has to be accurate. For instance, If I create a pitch modification algorithm, I have to ensure that every time I perform a 50 cents modification, the algorithm is reliably performing this modification independently of the incoming sound. If the experimenter needs to compare between male/female/animal/instrument the algorithmic performance has to be constant across these sound sources. If the algorithmic performance between sound conditions is not constant, the model can not be used to compare between these conditions, and no conclusions can be drawn from potential differences.

Second, somehow circularly, researchers should validate the perception of the effect. To do this, the more implicit the measure the better, as validating using explicit tasks, which can hide several forms of demand effects (see chapter 3).

Errors at either of these stages will propagate through the experimental pipeline, creating artefacts, noise or interpretation difficulties. If these errors are not controlled, the conclusions of subsequent studies can be affected by a confusion between perceptive effects and algorithmic limitations. For instance, one mistake done during this thesis was during the recording of an audiovisual dataset. The audio recording was done with an old built-in camera pre-amplifier which added hiss and noise to the background of the recordings. When transforming these sounds, the transformation increased these artefacts because of the spectral modifications, which made the sentences sound un-natural and deteriorated the quality of the perceptual measures in the experiment.

Step 3: experimental design, data analysis and interpretations

The cues manipulated by the algorithm will likely interact with the cues originally present in the stimuli. For instance, in the context of auditory smiles, a sentence with happy prosody and content, will likely be judged as positive even in the absence of smile specific cues. Prosody, content, identity, all interact with smile perception from the voice. It is important to ensure that such interactions are indeed perceptual interactions, and not algorithmic performance issues, when using e.g. a specific voice. Another example on how the cues of the original stimuli influence the observed results can be seen in the reverse correlation experiment (Study 2). The mean filter found with the reverse correlation technique is aligned with the features of the original phoneme. If we had taken too many or too different phonemes, the conclusions would likely be different, and would reflect a less precise, smoother filter, which takes into account only the common aspects of smiling between the phonemes, but not the formant-specific shifts we were able to observe here.

Another advantage of using parametric transformations is their fit to usual statistical models. Indeed, the intensity parameters, for instance in the form of physical change in stimuli (e.g. amount of formant shift) can be correlated with the psychological/physiological measures; the transformations can be easily integrated as factors of GLMMS or ANOVA analyses to test for main effects and interactions.

7.4.4 Some limitations

This research paradigm also has some limitations. First, one of the usual psychophysical techniques to normalise sensitivity across tasks and participants, the method of Just Noticeable Differences (JNDs), is not easily adaptable to this framework. JND is usually measured with an adaptive task, which measures participant's performance until they converge to performing a task with an accuracy of e.g. 70%. This often is a requirement when researchers want to assure sensitivity is constant across tasks and participants (Macmillan and Creelman, 2004). In the context of high-level tasks (involving e.g. emotions, or even articulation), algorithmic transformations such as the smile effect cannot easily be used in the same way, as would e.g. a classic pitch-shift. For instance, for an emotional task, as each pre-transformed

stimuli already contains some intrinsic emotional information, the algorithm transformation will not affect all stimuli in the same way. Imagine a 2AFC task, where participants have to answer whether the speaker is happy or sad, but the content varies between speakers. Here, the same amount of e.g. smile transformation will not lead to similar ratings across stimuli. In other words, each stimuli would need its own adaptive procedure in order to have an accurate measure of sensitivity.

A similar situation arises when chaining two parametric transformations together. For instance, in Study 7 (audiovisual perception of smiles), we found that the visual transformations more strongly influenced participant ratings than auditory transformations. What can be concluded from these results? Methodologically, it can not be said that visual information is cognitively more important than auditory information in this task, as the effects might be simply explained by a difference in the calibration of the algorithms. Imagine if I reduce the intensity of the visual algorithm until the auditory information becomes stronger. What would the conclusion be then? Without a proper way to normalise percepts between transformations, it is difficult to draw conclusions on the relationship between manipulations.

Another related aspect of the auditory smile effect —which was not developed upon here —is that its perception does not seem constant across speakers. For instance, in explicit tasks such as Study 3, 5 or 6, the mean shift of smiliness induced by the smile effect (measured as the difference smile - unsmile) changes for different voices. What is driving this difference? Are these algorithmics imperfections (of the kind discussed in step 2 above) or genuine perceptual interactions between the original cues of the stimuli and the smile effect? The answer is very likely both, with a bigger weight for perceptual interactions, but these questions need further research.

7.4.5 A final note on interdisciplinarity

Although using signal processing techniques to manipulate audio and video stimuli in the cognitive sciences can be scientifically fruitful, its implementation usually involves interdisciplinary methods, which do not always fit well in the disciplinary context of academic institutions. In the course of this thesis, we published signal processing papers, explaining algorithm design and architectures (Arias et al., 2018), a technical patent, describing the engineering novelty of the smile effect (Arias, Jean-Julien, and Roebel, 2017), and

psychology articles (Arias, Belin, and Aucouturier, 2018; Ponsot, Arias, and Aucouturier, 2018) investigating cognitive mechanisms. The work published in these documents is all on the same subject (i.e. auditory smiles) but describing different scientific stages of the research, to different audiences, and possibly with different impact in distinct fields. Indeed, the impact and novelty of our technical contributions (Arias et al., 2018) is not as important as those supported by those same contributions in the cognitive sciences. The algorithms implemented are interesting for the signal processing community, but a significant amount of their value lies in their use in experimental situations.

Is such an interdisciplinary approach useful for the science and for the apprentice researcher? First, "sound" is intrinsically interdisciplinary. Sound being the manifestation of physical signals interpreted by the brain, knowing the stages of information transformation from the physical signals all the way to behavior is scientifically useful to depict cognitive mechanisms. Second, in a world where important scientific advances come from the intertalk between scientific fields (e.g. machine learning and neuroscience), the merging of disciplines brings new methods and perspectives, stimulate new ideas and create new paradigms.

However, interdisciplinarity is not always easy to process for academic institutions (e.g. laboratories; job recruiters; doctoral schools), who usually expect PhDs to be experts in one specific scientific field. A researcher doing interdisciplinary research is usually not recognised as an expert in any of its sub-fields. Moreover, the interdisciplinary researcher has to learn the techniques, the theories and the social codes of two independent fields in parallel, which can lead to professional and personal imbalance (e.g. impostor syndrome or mental health problems) in today's highly competitive academic culture (Levecque et al., 2017; Hubel, 2009).

The value of interdisciplinary work lies in the creation of new links between otherwise hermetic scientific domains, and not only in the advance of one specific sub-field. These links should be encouraged and valued at an institutional level.

Appendix

A Modeling visual smiles

A.1 Introduction

Here we introduce a visual transformation technique able to manipulate an incoming audiovisual stream in real-time to parametrically control the amount of "smiliness" seen on the face, while preserving other characteristics such as the user's identity or the interaction's timing and content. The transformation is based on a warping technique informed by the real-time detection of visual landmarks.

The algorithm tracks morphological features of the face, such as the eyes and lip corners, stretches its position using a predefined parametric model, and resynthesizes pixel grey-levels to map the modified shape of the face. This algorithm significantly extends what constitute, to our knowledge, the only other example to date of real-time smile transformation Yoshida et al. (2013), by making it adaptive to the position of the user (more precisely, to camera-user distance and head pose), allowing users to speak during the transformation (an important limitation of previous work, allowing the simultaneous manipulation of smiled speech), as well as adding the possibility to use specific smile warpings that can be learned from a given user. We describe this algorithm, and how it relates to the existing literature, in Section A.2.

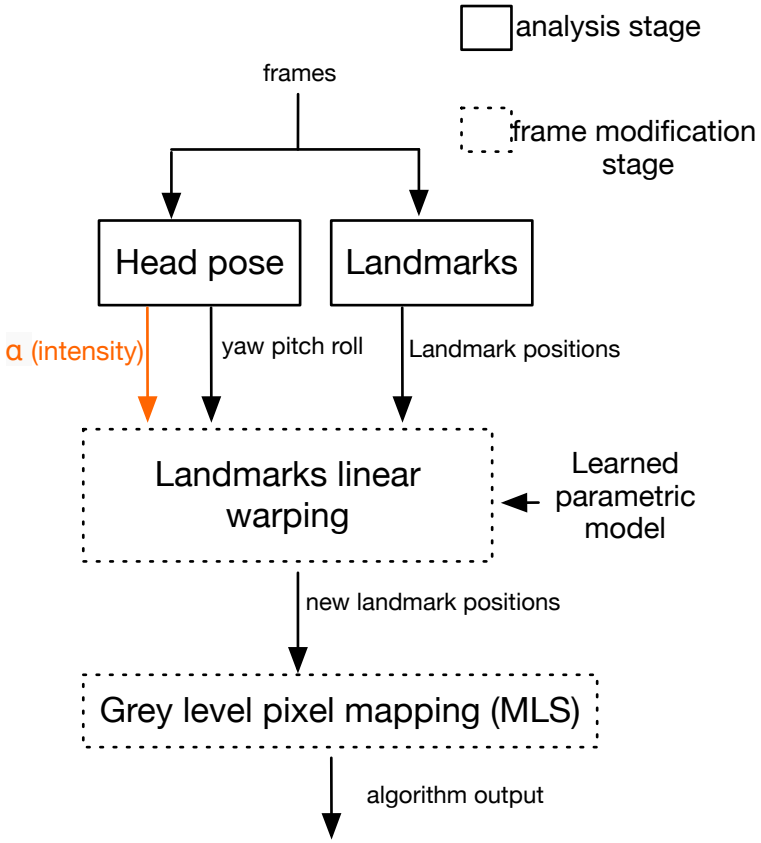
A.2 Algorithm architecture and design

A.2.1 Transformation algorithm

Smiling involves the activity of several muscles that raise the corners of the mouth and cheek, and lift the lower eyelids Ekman (2002). To recreate these distortions in real time on any face, we designed a two-stage image processing algorithm, which stretches morphological features of the face around the lips and the eyes using a pre-learned parametric model, and resynthesizes

pixel grey levels to correspond to the modified shape of the face. Figure A.1 illustrates the global process.

FIGURE A.1: Overview of the visual smile transformation. The first stage of the algorithm (solid line) extracts feature from the video frames: head pose and 84 landmarks, from which the system notably computes the distance between the subject's eyes. The second stage (dotted line) operate image manipulation: first, positions of 12 of the landmarks are modified using a learned linear model, then the grey-level pixel intensities of the image are changed using a Moving Least Square algorithm.



Landmarks linear warping

The algorithm works in real time and applies a pre-learned smile deformation on a frame by frame basis. For each frame, we first detect 84 landmarks on the face, as well as the head pose (roll, pitch, yaw) using a framework from a generic face tracking SDK provided by Dynamixyz Dynamixyz (2017) - see figure A.2-a.

Instead of heuristically designing a fixed warping function to simulate the expression of a smile, we made the choice to learn the pattern of landmark

distortion on one actor's expression, and then apply this pre-learned pattern to all subsequent input videos. The reasons for this choice are the following: first, while the smile expressions as defined e.g. in the Facial Action Coding System (FACS) Ekman (2002) make it possible to describe or detect such deformations, we did not find them sufficient to synthesize them with precision. Second, in an adaptive system, it appears interesting to learn the smile deformations that may be specific to a given person or attitude (e.g., genuine vs fake smiles Gunnery and Ruben (2016)).

Learning is based on two images of the same subject, a neutral face and a slightly smiling face (with mouth shut, no visible teeth). After aligning the two faces, we calculated a linear deformation model for 12 landmarks to model the changes in the Zygomaticus Major (AU 12) and the Orbicularis Oculi (AU 6) muscles involved in smiling Ekman (2002) (2 landmarks on the lower eyelid for each eye, 2 landmarks on the corners of the lips, three landmarks on the upper lip and three others on the lower lip - see figure A.2-b).

In more details, if i is one landmark ($i = 1..12$), X_i^n its 2D coordinates from the neutral face and X_i^s its 2D coordinates from the smiling face, the linear model can be described as

$$X_i^s = (X_i^n + Q_r * \Delta_{xy}) * scale * \alpha_{video} \tag{A.1}$$

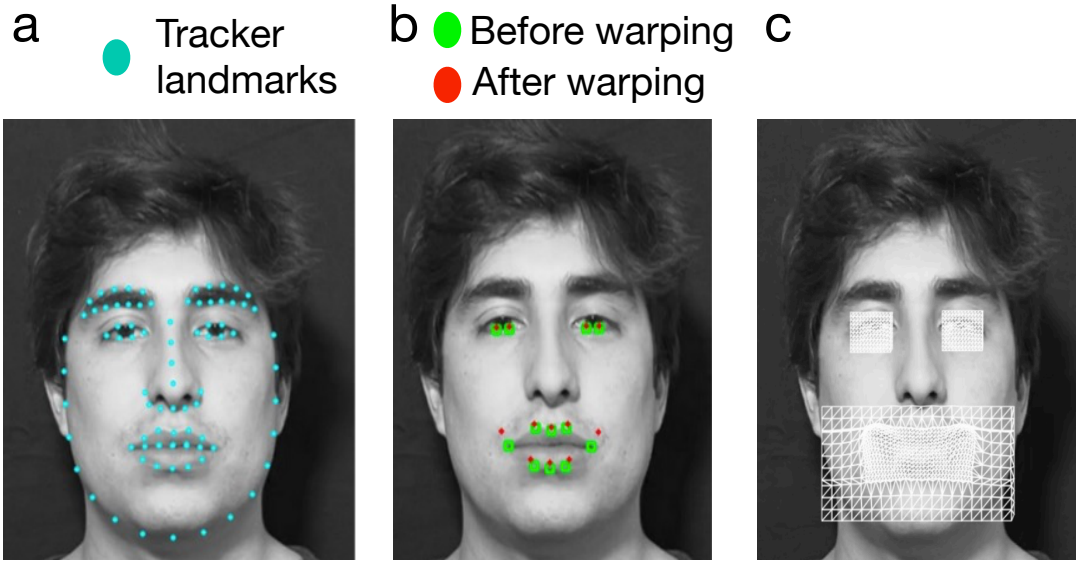
where Δ_{xy} is the learned parameters of the model and α_{video} is the intensity of the smile distortion. To adapt to face-camera distance and head pose, $scale$ is computed as the distance between the two eyes multiplied by cosine of the angle yaw, and Q_r is the rotation matrix corresponding to the roll. Figure A.2-b shows an example of original and modified landmarks for a frontal face with $\alpha_{video} = 2.5$.

Pixel grey levels mapping

The second step of the algorithm computes the impact of landmark warping on the pixel grey-level intensity. As in Yoshida et al. (2013), we use the rigid Moving Least Squares (MLS) method Schaefer, McPhail, and Warren (2006). MLS optimizes the deformation made on an image when the position of some landmarks is modified, while maintaining the spatial coherence of the overall shape.

We made two approximations to the standard MLS procedure in order to allow real-time performance. First, we apply MLS only to areas of the image around the mouth and the eyes. Second, we do not apply the algorithm to every pixel of these areas but first approximate the areas with grids (with smaller meshes close to the eyes and mouth) and apply the deformation function to each vertex in the grid. We then fill the resulting triangles using affine warping. Figure A.2-c shows an example of the grids after the MLS algorithm.

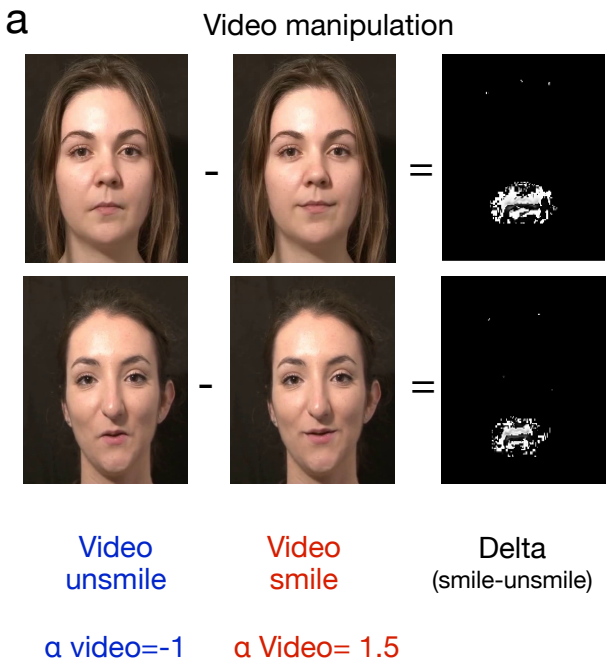
FIGURE A.2: Illustration of the tracking, warping and mapping steps in the visual smile transformation. (a) Tracking: 84 landmarks (turquoise dots) are automatically detected on the face. (b) Warping: the positions of 12 of the 84 landmarks (green dots) are transformed using a pre-learned linear model (red dots). (c) Mapping: we create a grid around the mouth and eyes, apply Moving Least Square deformation to each vertex of the grid and interpolate inside each resulting triangle using affine warping.



A.2.2 Video results

Figure A.3 present two examples of video frames transformed by the visual transformation. In the left/center are the unsmile/smile visual manipulations. The delta (difference smile-unsmile pixel wise) computed in the right highlights the area where the modification takes place, namely, the mouth, independently of the speaker (the pixels in black are identical between smile and unsmile conditions; more visual examples can be seen in Figure 5.1).

FIGURE A.3: Visual algorithmic performance (a) Example of the unsmile (left) smile (center) visual manipulations. (right) Pixel wise difference between smile and unsmile visual manipulations for a specific frame, which allows to see the differences between the two transformations. Black means there are no differences, white the opposite



A.2.3 Implementation discussion

One limitation of the warping model is that, while it is by construction adaptive to the frame-by-frame position of the mouth, it isn't to the qualitative nature of pronounced phonemes during speech. For instance, during real speech production, protruded/round vowels such as [y] may be incongruent with a large smile, whereas smiles on unround vowels such as [i] can be amplified without breaking the acoustic characteristics of the sound. A possible extension of the algorithm would be to learn a separate deformation pattern for different types of phonemes, and apply them adaptively, but this is beyond the scope of the current work.

One limitation of the MLS algorithm is that it cannot create textures that are not present in the original image, such as wrinkles. In particular, there are time-varying features (or "discontinuities") in the mouth and eyes areas (e.g., teeth which appear or disappear behind opening lip tissue, white sclera revealed by opening eyelids), which the algorithm cannot "add" to a frame if

not originally present. Finally, at large intensities, the MLS algorithm may stretch geometric shapes, resulting e.g. in unrealistically oval rather than round iris shape, although the effect is not observable at the intensities investigated here.

Finally, we measured the latency of the overall visual algorithm including the 3 processing stages. The mean time (over 1000 iterations) to process a frame depends on the processing power of the machine. Our tests resulted in a mean 61ms processing time for a single frame (45ms for landmark tracking, 7ms for warping and 9ms for MLS) which is suitable for real time applications, for instance at 15 fps. Anyway, the latency can be further diminished either by reducing the number of landmarks in the tracking stage—the most time consuming stage—, or by improving the machine processing power, specially, the CPU speed. You can find examples of the algorithm at https://archive.org/download/StimuliExample, where speaking and head poses/orientations variations are presented.

A.3 Conclusion

We created an visual smile transformation algorithm able to manipulate an incoming video stream in real-time to parametrically control the amount of smile seen on the users' face. To simulate visual smiles, we use face recognition and automatic landmark positioning, followed by a warping and a mapping stage. An experimental validation the parametric aspect of this effect can be found in (Arias et al., 2018).

B Publications by the author

Peer reviewed journal articles

Arias, Pablo, Belin, Pascal, and Jean-Julien Aucouturier. (2018). Auditory smiles trigger unconscious facial imitations. Current Biology, 28(14), R782-R783.

Arias, P., Soladie, C., Bouafif, O., Robel, A., Seguier, R., & Aucouturier, J. J. (2018). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. IEEE Transactions on Affective Computing.

Ponsot, E., Arias, P., & Aucouturier, J. J. (2018). Uncovering mental representations of smiled speech using reverse correlation. The Journal of the Acoustical Society of America, 143(1), EL19-EL24.

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., ... & Aucouturier, J. J. (2018). DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. Behavior research methods, 50(1), 323-343.

Abstracts and presentations at international conferences

Arias, P., Belin, P., & Aucouturier, J.-J., Hearing smiles and smiling back. Laughter workshop. Institut des Systèmes Intelligents et de Robotique (ISIR). Paris, France, September 2018.

Arias, P., Auditory smiles trigger unconscious facial reactions. Contextual: How the Social Context Shapes Brain and Behaviour. International conference of the European Society for Cognitive and Affective Neurosciences (ESCAN). Leiden, NL, July 2018.

Arias, P., Belin, P. , & Aucouturier, J-J., Unconsciously imitating smiles heard in speech —and hearing smiles in musical sounds. European Research Music Conference. Barcelona, Spain, June 2018.

Arias, P., Représentations mentales du sourire dans la voix parlé: une étude par corrélation inverse. Congrès Français d'Acoustique (CFA), Le Havre, France, Mars 2018.

Arias, P., Ziggy : the rise and fall of zygomatic muscles in speech. Conference of the Consortium of European Research on Emotions (CERE), Glasgow, United Kindgom, April 2018.

Arias, P., Unconscious physiological reactions to smiles in speech revealed by super-vp's frequency warping. Journée RIM. IRCAM, Paris, France, October 2017.

Arias, P., Spectral cues caused by smiling trigger unconscious facial imitation. Music Language and Cognition Summer School. Como, Italy, June 2017.

Arias, P., Emotional mimicry induced by manipulated speech. Workshop on Music cognition, emotion and audio technology in Tokyo. University of Tokyo, Tokyo, Japan, November 2016.

Arias, P., Emotional mimicry induced by manipulated speech. Journées Jeunes Chercheurs en Audition Acoustique musicale et Signal audio (JJCAAS). France, Paris, November 2016.

Arias, P., Time perception and neural oscillations modulated by speech rate. Festival IRCAM-Manifeste, Paris, France, October 2016.

Rachman, L., Liuni, M., Arias, P., & Aucouturier, J. J., Synthesizing speech-like emotional expression onto music and speech signals. In Fifth International Conference on Music and Emotions. ICME4, Geneva, CH, October 2015.

Patent

Arias, P., Aucouturier, J.-J., Roebel, A., (2018). Méthode et appareil de modification dynamique du timbre de la voix par décalage en fréquence des formants d'une enveloppe spectrale.

Teaching and Public dissemination

Arias, P., Perception of Smiles in the Voice., Voice Tech Podcast, Online, Paris, France, 2018.

Arias, P., Essential aspects of voice perception., IRCAM, Cursus- Program on Composition and Computer Music, IRCAM, Paris, 2018.

Arias, P., Trois outils de traitement de la voix émotionnelle et leurs effets physiologiques. École nationale d'ingénieurs de Tunis (ENIT), Journée d'études : TICs, Musique et émotion. Tunisie, Tunis, Janvier 2018.

Arias, P., Liuni M., Transformations émotionnelles de la voix parlée —conséquences comportementales et physiologiques. Journée voix - Studio 5 en direct, IRCAM, Paris, October 2017.

Conference organisation

Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio (JJCAAS), Paris, 23, 24, 25 November 2016

Master's Thesis

Arias, P., Françoise, J., Bevilacqua, F., & Schnell, N. (2014). Sound description and synthesis in the MaD (Mapping by demonstration) framework, Master thesis

Bibliography

Adolphs, Ralph, Daniel Tranel, Hanna Damasio, and Antonio Damasio (1994). "Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala". In: *Nature* 372.6507, p. 669.

Adolphs, Ralph, Frederic Gosselin, Tony W Buchanan, Daniel Tranel, Philippe Schyns, and Antonio R Damasio (2005). "A mechanism for impaired fear recognition after amygdala damage". In: *Nature* 433.7021, p. 68.

Ahumada Jr, Al and John Lovell (1971). "Stimulus features in signal detection". In: *The Journal of the Acoustical Society of America* 49.6B, pp. 1751–1756.

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Ansfield, Matthew E (2007). "Smiling when distressed: When a smile is a frown turned upside down". In: *Personality and Social Psychology Bulletin* 33.6, pp. 763–775.

Arias, Pablo, Pascal Belin, and Jean-Julien Aucouturier (2018). "Auditory smiles trigger unconscious facial imitations". In: *Current Biology*.

Arias, Pablo, Aucouturier Jean-Julien, and Axel Roebel (2017). "Méthode et appareil de modification dynamique du timbre de la voix par décalage en fréquence des formants d'une enveloppe spectrale". Patent CNRS 09652-01.

Arias, Pablo, Catherine Soladie, Oussema Bouafif, Axel Robel, Renaud Seguier, and Jean-Julien Aucouturier (2018). "Realistic transformation of facial and vocal smiles in real-time audiovisual streams". In: *IEEE Transactions on Affective Computing*.

Arnal, Luc H, Adeen Flinker, Andreas Kleinschmidt, Anne-Lise Giraud, and David Poeppel (2015). "Human screams occupy a privileged niche in the communication soundscape". In: *Current Biology* 25.15, pp. 2051–2056.

Arnott, Stephen R, Anthony Singhal, and Melvyn A Goodale (2009). "An investigation of auditory contagious yawning". In: *Cognitive, Affective, & Behavioral Neuroscience* 9.3, pp. 335–342.

Aucouturier, Jean-Julien, Petter Johansson, Lars Hall, Rodrigo Segnini, Lolita Mercadié, and Katsumi Watanabe (2016). "Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction". In: *Proceedings of the National Academy of Sciences* 113.4, pp. 948–953.

Baart, Martijn and Jean Vroomen (2018). "Recalibration of vocal affect by a dynamic face". In: *Experimental brain research*, pp. 1–8.

Bachorowski, Jo-Anne and Michael J Owren (1995). "Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context". In: *Psychological science* 6.4, pp. 219–224.

Banse, Rainer and Klaus R Scherer (1996). "Acoustic profiles in vocal emotion expression." In: *Journal of personality and social psychology* 70.3, p. 614.

Bargh, John A (1989). "Conditional automaticity: Varieties of automatic influence in social perception and cognition". In: *Unintended thought* 3, pp. 51–69.

Barrett, Lisa Feldman (2017). "The theory of constructed emotion: an active inference account of interoception and categorization". In: *Social cognitive and affective neuroscience* 12.1, pp. 1–23.

Barsalou, Lawrence W, Paula M Niedenthal, Aron K Barbey, and Jennifer A Ruppert (2003). "Social embodiment". In: *Psychology of learning and motivation* 43, pp. 43–92.

Barthel, Helen and Hugo Quené (2015). "Acoustic-phonetic properties of smiling revised–measurements on a natural video corpus". In: *Proceedings of the 18th International Congress of Phonetic Sciences.–Glasgow, UK: The University of Glasgow*.

Basso, Frédéric and Olivier Oullier (2010). ""Smile down the phone": Extending the effects of smiles to vocal social interactions". In: *Behavioral and Brain Sciences* 33.06, pp. 435–436.

Baudouin, Jean-Yves, Daniel Gilibert, Stephane Sansone, and Guy Tiberghien (2000). "When the smile is a cue to familiarity". In: *Memory* 8.5, pp. 285–292.

Beauchamp, Michael S, Audrey R Nath, and Siavash Pasalar (2010). "fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect". In: *Journal of Neuroscience* 30.7, pp. 2414–2417.

Bedoya, Daniel, Louise Goupil, and Jean-Julien Aucouturier (2018). "Les émotions sont-elles exprimées de la même façon en musique que dans la voix parlée ?" MA thesis. Sorbonne Université.

Behroozmand, Roozbeh, Oleg Korzyukov, Lindsey Sattler, and Charles R Larson (2012). "Opposing and following vocal responses to pitch-shifted

auditory feedback: evidence for different mechanisms of voice pitch control". In: *The Journal of the Acoustical Society of America* 132.4, pp. 2468–2477.

Belin, Pascal, Robert J Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike (2000). "Voice-selective areas in human auditory cortex". In: *Nature* 403.6767, p. 309.

Berg, Bruce G (1989). "Analysis of weights in multiple observation tasks". In: *The Journal of the Acoustical Society of America* 86.5, pp. 1743–1746.

Bernieri, Frank J, J Steven Reznick, and Robert Rosenthal (1988). "Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions." In: *Journal of personality and social psychology* 54.2, p. 243.

Bestelmeyer, Patricia EG, Pascal Belin, and Marie-Helene Grosbras (2011). "Right temporal TMS impairs voice detection". In: *Current Biology* 21.20, R838–R839.

Blairy, Sylvie, Pedro Herrera, and Ursula Hess (1999). "Mimicry and the judgment of emotional facial expressions". In: *Journal of Nonverbal behavior* 23.1, pp. 5–41.

Blasi, Anna, Evelyne Mercure, Sarah Lloyd-Fox, Alex Thomson, Michael Brammer, Disa Sauter, Quinton Deeley, Gareth J Barker, Ville Renvall, Sean Deoni, et al. (2011). "Early specialization for voice and emotion processing in the infant brain". In: *Current Biology* 21.14, pp. 1220–1224.

Bliss-Moreau, ELIZA and GILDA Moadab (2017). "The faces monkeys make". In: *The science of facial expression. New York, NY: Oxford*.

Blood, Anne J and Robert J Zatorre (2001). "Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion". In: *Proceedings of the National Academy of Sciences* 98.20, pp. 11818–11823.

Blumenthal, Terry D and Christopher T Goode (1991). "The startle eyeblink response to low intensity acoustic stimuli". In: *Psychophysiology* 28.3, pp. 296–306.

Boersma, Paul and David Weenink (2017). *Praat: doing phonetics by computer [Computer program]Version 6.0.24, retrieved 23 January 2017 from http://www.praat.org/*.

Boersma, Paulus Petrus Gerardus et al. (2002). "Praat, a system for doing phonetics by computer". In: *Glot international* 5.

Bourgeois, Patrick and Ursula Hess (2008). "The impact of social context on mimicry". In: *Biological psychology* 77.3, pp. 343–352.

Bowling, DL, M Garcia, JC Dunn, R Ruprecht, A Stewart, K-H Frommolt, and WT Fitch (2017). "Body size and vocalization in primates and carnivores". In: *Scientific reports* 7, p. 41070.

Boxtel, Jeroen JA van and Hongjing Lu (2015). "Joints and their relations as critical features in action discrimination: Evidence from a classification image method". In: *Journal of vision* 15.1, pp. 20–20.

Bradley, Margaret M, Laura Miccoli, Miguel A Escrig, and Peter J Lang (2008). "The pupil as a measure of emotional arousal and autonomic activation". In: *Psychophysiology* 45.4, pp. 602–607.

Breiter, Hans C, Nancy L Etcoff, Paul J Whalen, William A Kennedy, Scott L Rauch, Randy L Buckner, Monica M Strauss, Steven E Hyman, and Bruce R Rosen (1996). "Response and habituation of the human amygdala during visual processing of facial expression". In: *Neuron* 17.5, pp. 875–887.

Briefer, EF (2012). "Vocal expression of emotions in mammals: mechanisms of production and evidence". In: *Journal of Zoology* 288.1, pp. 1–20.

Bruder, Martin, Dina Dosmukhambetova, Josef Nerb, and Antony SR Manstead (2012). "Emotional signals in nonverbal interaction: Dyadic facilitation and convergence in expressions, appraisals, and feelings". In: *Cognition & emotion* 26.3, pp. 480–502.

Bryant, Gregory A, Daniel MT Fessler, Riccardo Fusaroli, Edward Clint, Lene Aarøe, Coren L Apicella, Michael Bang Petersen, Shaneikiah T Bickham, Alexander Bolyanatz, Brenda Chavez, et al. (2016). "Detecting affiliation in colaughter across 24 societies". In: *Proceedings of the National Academy of Sciences* 113.17, pp. 4682–4687.

Bryant, Gregory A, Daniel MT Fessler, Riccardo Fusaroli, Edward Clint, Dorsa Amir, Brenda Chávez, Kaleda K Denton, Cinthya Díaz, Lealaiauloto Togiaso Duran, Jana Fančovičová, et al. (2018). "The perception of spontaneous and volitional laughter across 21 societies". In: *Psychological science* 29.9, pp. 1515–1525.

Buchsbaum, Bradley R, Gregory Hickok, and Colin Humphries (2001). "Role of left posterior superior temporal gyrus in phonological processing for speech perception and production". In: *Cognitive Science* 25.5, pp. 663–678.

Burred, Juan José, Emmanuel Ponsot, Louise Goupil, Marco Liuni, and Jean-Julien Aucouturier (2018). "CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition". In: *bioRxiv*, p. 436477.

Camras, L., Serah Fatani, B. Fraumeni, and M. Shuster (2016). "The development of facial expressions: current perspectives on infant emotions". In:

Handbook of emotions, Fourth Edition. Ed. by L. F. Barrett, M. Lewis, and J. M. Haviland-Jones. NY: Guilford Press.

Cannizzaro, Michael, Brian Harel, Nicole Reilly, Phillip Chappell, and Peter J Snyder (2004). "Voice acoustical measurement of the severity of major depression". In: *Brain and cognition* 56.1, pp. 30–35.

Cannon, P., A. Hayes, and S Tipper (2009). "An electromyographic investigation of the impact of task relevance on facial mimicry". In: *Cognition & Emotion* 23(5), 918–929.

Chartrand, Tanya L and John A Bargh (1999). "The chameleon effect: the perception–behavior link and social interaction." In: *Journal of personality and social psychology* 76.6, p. 893.

Chartrand, Tanya L and Amy N Dalton (2009). "Mimicry: Its ubiquity, importance, and functionality". In: *Oxford handbook of human action*, pp. 458–483.

Cheng, Clara Michelle and Tanya L Chartrand (2003). "Self-monitoring without awareness: using mimicry as a nonconscious affiliation strategy." In: *Journal of personality and social psychology* 85.6, p. 1170.

Chevalier-Skolnikoff, Suzanne (1974). "The primate play face: A possible key to the determinants and evolution of play". In: *Rice Institute Pamphlet-Rice University Studies* 60.3.

Chong, Chee Seng, Jeesun Kim, and Chris Davis (2018). "Disgust expressive speech: The acoustic consequences of the facial expression of emotion". In: *Speech Communication* 98, pp. 68–72.

Collignon, Olivier, Simon Girard, Frederic Gosselin, Sylvain Roy, Dave Saint-Amour, Maryse Lassonde, and Franco Lepore (2008). "Audio-visual integration of emotion expression". In: *Brain research* 1242, pp. 126–135.

Cooper, Franklin S, Pierre C Delattre, Alvin M Liberman, John M Borst, and Louis J Gerstman (1952). "Some experiments on the perception of synthetic speech sounds". In: *The Journal of the Acoustical Society of America* 24.6, pp. 597–606.

Cosmides, Leda and John Tooby (2000). "Evolutionary psychology and the emotions". In: *Handbook of emotions* 2.2, pp. 91–115.

Cromheeke, Sofie and Sven C Mueller (2016). "The power of a smile: stronger working memory effects for happy faces in adolescents compared to adults". In: *Cognition and Emotion* 30.2, pp. 288–301.

Damasio, Antonio R, Thomas J Grabowski, Antoine Bechara, Hanna Damasio, Laura LB Ponto, Josef Parvizi, and Richard D Hichwa (2000). "Subcortical and cortical brain activity during the feeling of self-generated emotions". In: *Nature neuroscience* 3.10, p. 1049.

Darwin, Charles (1872). "The expression of the emotions in man and animals". In: *The American Journal of the Medical Sciences* 232.4, p. 477.

D'Ausilio, Alessandro, Friedemann Pulvermüller, Paola Salmas, Ilaria Bufalari, Chiara Begliomini, and Luciano Fadiga (2009). "The motor somatotopy of speech perception". In: *Current Biology* 19.5, pp. 381–385.

De Boer, Bart and Patricia K Kuhl (2003). "Investigating the role of infant-directed speech with a computer model". In: *Acoustics Research Letters Online* 4.4, pp. 129–134.

De Gelder, Beatrice and Jean Vroomen (2000). "The perception of emotions by ear and by eye". In: *Cognition & Emotion* 14.3, pp. 289–311.

De Vignemont, F. and T. Singer (2006a). "The empathic brain: how, when and why?" In: *Trends in cognitive sciences* 10(10), pp. 435–441.

De Vignemont, Frederique and Tania Singer (2006b). "The empathic brain: how, when and why?" In: *Trends in cognitive sciences* 10.10, pp. 435–441.

Di Pellegrino, Giuseppe, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti (1992). "Understanding motor events: a neurophysiological study". In: *Experimental brain research* 91.1, pp. 176–180.

Dimberg, Ulf and Monika Thunberg (1998). "Rapid facial reactions to emotional facial expressions". In: *Scandinavian journal of psychology* 39.1, pp. 39–45.

Dimberg, Ulf, Monika Thunberg, and Kurt Elmehed (2000). "Unconscious facial reactions to emotional facial expressions". In: *Psychological science* 11.1, pp. 86–89.

Dimberg, Ulf, Monika Thunberg, and Sara Grunedal (2002). "Facial reactions to emotional stimuli: Automatically controlled emotional responses". In: *Cognition & Emotion* 16.4, pp. 449–471.

Dolan, Raymond J, John S Morris, and Beatrice de Gelder (2001). "Crossmodal binding of fear in voice and face". In: *Proceedings of the National Academy of Sciences* 98.17, pp. 10006–10010.

Drahota, Amy, Alan Costall, and Vasudevi Reddy (2008). "The vocal communication of different kinds of smile". In: *Speech Communication* 50.4, pp. 278–287.

Driver, Jon (1996). "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading". In: *Nature* 381.6577, p. 66.

Dynamixyz (2017). "Dynamixyz Generic Face Tracking". In: URL: www.Dynamixyz.com.

Eibl-Eibesfeldt, Irenäus (1973). "The expressive behaviour of the deaf-andblind-born." In: *Social communication and movement*, pp. 163–194.

Ekman, Paul (2002). "Facial action coding system (FACS)". In: *A human face*.

Ekman, Paul and Wallace V Friesen (1978). *Manual for the facial action coding system*. Consulting Psychologists Press.

– (1982). "Felt, false, and miserable smiles". In: *Journal of nonverbal behavior* 6.4, pp. 238–252.

Ekman, Paul and Erika L Rosenberg (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Ekman, Paul, E Richard Sorenson, Wallace V Friesen, et al. (1969). "Pan-cultural elements in facial displays of emotion". In: *Science* 164.3875, pp. 86–88.

El Haddad, Kevin, Stéphane Dupont, Nicolas d'Alessandro, and Thierry Dutoit (2015a). "An HMM-based speech-smile synthesis system: An approach for amusement synthesis". In: *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. Vol. 5. IEEE, pp. 1–6.

El Haddad, Kevin, Hüseyin Cakmak, Stéphane Dupont, and Thierry Dutoit (2015b). "Towards a speech synthesis system with controllable amusement levels". In: *4th Interdisciplinary Workshop on Laughter and Other Non-Verbal Vocalisations in Speech, Enschede, Netherlands*, pp. 14–15.

– (2016). "Laughter and Smile Processing for Human-Computer Interactions". In: *Just talking-casual talk among humans and machines, Portoroz, Slovenia*, pp. 23–28.

El Haddad, Kevin, Ilaria Torre, Emer Gilmartin, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit, and Nick Campbell (2017). "Introducing AmuS: The Amused Speech Database". In: *International Conference on Statistical Language and Speech Processing*. Springer, pp. 229–240.

Ellsworth, Phoebe C (1994). "William James and emotion: is a century of fame worth a century of misunderstanding?" In: *Psychological Review* 101.2, p. 222.

Erro, Daniel, Eva Navas, and Inma Hernaez (2013). "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.3, pp. 556–566.

Ethofer, Thomas, Dimitri Van De Ville, Klaus Scherer, and Patrik Vuilleumier (2009). "Decoding of emotional information in voice-sensitive cortices". In: *Current Biology* 19.12, pp. 1028–1033.

Evans, Sarah, Nick Neave, and Delia Wakelin (2006). "Relationships between vocal characteristics and body size and shape in human males: an evolutionary explanation for a deep male voice". In: *Biological psychology* 72.2, pp. 160–163.

Fagel, Sascha (2010). "Effects of smiling on articulation: Lips, larynx and acoustics". In: *Development of multimodal interfaces: active listening and synchrony*. Springer, pp. 294–303.

Fecteau, Shirley, Pascal Belin, Yves Joanette, and Jorge L Armony (2007). "Amygdala responses to nonlinguistic emotional vocalizations". In: *Neuroimage* 36.2, pp. 480–487.

Fiacconi, Chris M and Adrian M Owen (2016). "Using facial electromyography to detect preserved emotional processing in disorders of consciousness: A proof-of-principle study". In: *Clinical Neurophysiology* 127.9, pp. 3000–3006.

Fischer, Agneta H, Antony SR Manstead, et al. (2008). "Social functions of emotion". In: *Handbook of emotions* 3, pp. 456–468.

Fitch, W Tecumseh (1997). "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques". In: *The Journal of the Acoustical Society of America* 102.2, pp. 1213–1222.

– (2000). "The evolution of speech: a comparative review". In: *Trends in cognitive sciences* 4.7, pp. 258–267.

Fitch, W Tecumseh, Jürgen Neubauer, and Hanspeter Herzel (2002). "Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production". In: *Animal behaviour* 63.3, pp. 407–418.

Föcker, Julia, Matthias Gondan, and Brigitte Röder (2011). "Preattentive processing of audio-visual emotional signals". In: *Acta psychologica* 137.1, pp. 36–47.

Frühholz, Sascha, Wiebke Trost, and Sonja A Kotz (2016). "The sound of emotions—Towards a unifying neural network perspective of affective sound processing". In: *Neuroscience & Biobehavioral Reviews* 68, pp. 96–110.

Galantucci, Bruno, Carol A Fowler, and Michael T Turvey (2006). "The motor theory of speech perception reviewed". In: *Psychonomic bulletin & review* 13.3, pp. 361–377.

Gallese, Vittorio, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti (1996). "Action recognition in the premotor cortex". In: *Brain* 119.2, pp. 593–609.

Gazzola, Valeria and Christian Keysers (2008). "The observation and execution of actions share motor and somatosensory voxels in all tested subjects: single-subject analyses of unsmoothed fMRI data". In: *Cerebral Cortex* 19.6, pp. 1239–1255.

Gelman, Andrew and Jennifer Hill (2007). *Data analysis using regression and multilevelhierarchical models*. Vol. 1. Cambridge University Press New York, NY, USA.

Gerdes, Antje, Matthias J Wieser, and Georg W Alpers (2014). "Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains". In: *Frontiers in Psychology* 5, p. 1351.

Giles, Howard (1973). "Accent mobility: A model and some data". In: *Anthropological linguistics*, pp. 87–105.

Glanz, Olga, Johanna Derix, Rajbir Kaur, Andreas Schulze-Bonhage, Peter Auer, Ad Aertsen, and Tonio Ball (2018). "Real-life speech production and perception have a shared premotor-cortical substrate". In: *Scientific reports* 8.1, p. 8898.

Gold, Jason M, Allison B Sekuler, and Partrick J Bennett (2004). "Characterizing perceptual learning with external noise". In: *Cognitive Science* 28.2, pp. 167–207.

Golle, Jessika, Fred W Mast, and Janek S Lobmaier (2014). "Something to smile about: The interrelationship between attractiveness and emotional expression". In: *Cognition & emotion* 28.2, pp. 298–310.

Goodale, Melvyn A and A David Milner (1992). "Separate visual pathways for perception and action". In: *Trends in neurosciences* 15.1, pp. 20–25.

Gosselin, Frédéric and Philippe G Schyns (2003). "Superstitious perceptions reveal properties of internal representations". In: *Psychological Science* 14.5, pp. 505–509.

Gosselin, Nathalie, Isabelle Peretz, Erica Johnsen, and Ralph Adolphs (2007). "Amygdala damage impairs emotion recognition from music". In: *Neuropsychologia* 45.2, pp. 236–244.

Gougoux, Frédéric, Franco Lepore, Maryse Lassonde, Patrice Voss, Robert J Zatorre, and Pascal Belin (2004). "Pitch discrimination in the early blind: People blinded in infancy have sharper listening skills than those who lost their sight later." In: *Nature*.

Gramfort, Alexandre, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen (2014). "MNE software for processing MEG and EEG data". In: *Neuroimage* 86, pp. 446–460.

Grant, Ken W and Philip-Franz Seitz (2000). "The use of visible speech cues for improving auditory detection of spoken sentences". In: *The Journal of the Acoustical Society of America* 108.3, pp. 1197–1208.

Gross, James J (1998). "Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology." In: *Journal of personality and social psychology* 74.1, p. 224.

Gunnery, Sarah D and Mollie A Ruben (2016). "Perceptions of Duchenne and non-Duchenne smiles: A meta-analysis". In: *Cognition and Emotion* 30.3, pp. 501–515.

Hasan, Bashar, Mitchell Valdes-Sosa, Joachim Gross, and Pascal Belin (2016a). "Hearing faces and seeing voices: Amodal coding of person identity in the human brain". In: *Scientific Reports* 6 (37494).

Hasan, Bashar Awwad Shiekh, Mitchell Valdes-Sosa, Joachim Gross, and Pascal Belin (2016b). ""Hearing faces and seeing voices": Amodal coding of person identity in the human brain". In: *Scientific reports* 6, p. 37494.

Hatfield, Elaine, John T Cacioppo, and Richard L Rapson (1993). "Emotional contagion". In: *Current directions in psychological science* 2.3, pp. 96–100.

Hawk, Skyler T, Agneta H Fischer, and Gerben A Van Kleef (2012). "Face the noise: Embodied responses to nonverbal vocalizations of discrete emotions." In: *Journal of Personality and Social Psychology* 102.4, p. 796.

Hepach, Robert and Gert Westermann (2013). "Infants' sensitivity to the congruence of others' emotions and actions". In: *Journal of experimental child psychology* 115.1, pp. 16–29.

Hess, Ursula, Martin G Beaupré, Nicole Cheung, et al. (2002). "Who to whom and why–cultural differences and similarities in the function of smiles". In: *An empirical reflection on the smile* 4, p. 187.

Hess, Ursula and Agneta Fischer (2013). "Emotional mimicry as social regulation". In: *Personality and Social Psychology Review* 17.2, pp. 142–157.

Hess, Ursula and Agneta H Fischer (2016). *Emotional mimicry in social context*. Cambridge University Press.

Hess, Ursula, Pierre Philippot, and Sylvie Blairy (1998). "Facial reactions to emotional facial expressions: Affect or cognition?" In: *Cognition & Emotion* 12.4, pp. 509–531.

Heyes, Cecilia M and Chris D Frith (2014). "The cultural evolution of mind reading". In: *Science* 344.6190, p. 1243091.

Hickok, Gregory (2014). *The myth of mirror neurons: The real neuroscience of communication and cognition*. WW Norton & Company.

Hickok, Gregory, John Houde, and Feng Rong (2011). "Sensorimotor integration in speech processing: computational basis and neural organization". In: *Neuron* 69.3, pp. 407–422.

Hickok, Gregory and David Poeppel (2007). "The cortical organization of speech processing". In: *Nature Reviews Neuroscience* 8.5, p. 393.

Hickok, Gregory, Maddalena Costanzo, Rita Capasso, and Gabriele Miceli (2011). "The role of Broca's area in speech perception: Evidence from aphasia revisited". In: *Brain and language* 119.3, pp. 214–220.

Hietanen, Jari K, Veikko Surakka, and Ilkka Linnankoski (1998). "Facial electromyographic responses to vocal affect expressions". In: *Psychophysiology* 35.05, pp. 530–536.

Hoehl, Stefanie, Kahl Hellmer, Maria Johansson, and Gustaf Gredebäck (2017). "Itsy bitsy spider...: infants react with increased arousal to spiders and snakes". In: *Frontiers in psychology* 8, p. 1710.

Hommel, Bernhard (2007). "Consciousness and control: not identical twins". In: *Journal of Consciousness Studies* 14.1-2, pp. 155–176.

Hubel, David H (2009). "The way biomedical research is organized has dramatically changed over the past half-century: Are the changes for the better?" In: *Neuron* 64.2, pp. 161–163.

Im, Hee Yeon and Justin Halberda (2013). "The effects of sampling and internal noise on the representation of ensemble average size". In: *Attention, Perception, & Psychophysics* 75.2, pp. 278–286.

Izard, Carroll E (1994). "Innate and universal facial expressions: evidence from developmental and cross-cultural research." In:

– (2007). "Basic emotions, natural kinds, emotion schemas, and a new paradigm". In: *Perspectives on psychological science* 2.3, pp. 260–280.

– (2009). "Emotion theory and research: Highlights, unanswered questions, and emerging issues". In: *Annual review of psychology* 60, pp. 1–25.

Jabbi, Mbemba, Marte Swart, and Christian Keysers (2007). "Empathy for positive and negative emotions in the gustatory cortex". In: *Neuroimage* 34.4, pp. 1744–1753.

Jack, Rachael E, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns (2012). "Facial expressions of emotion are not culturally universal". In: *Proceedings of the National Academy of Sciences* 109.19, pp. 7241–7244.

Jack, Rachael E, Wei Sun, Ioannis Delis, Oliver GB Garrod, and Philippe G Schyns (2016). "Four not six: Revealing culturally common facial expressions of emotion." In: *Journal of Experimental Psychology: General* 145.6, p. 708.

James, William (1884). "What is an emotion?" In: *Mind* 9.34, pp. 188–205.

Jessen, Sarah, Nicole Altvater-Mackensen, and Tobias Grossmann (2016). "Pupillary responses reveal infants' discrimination of facial emotions independent of conscious perception". In: *Cognition* 150, pp. 163–169.

Juslin, Patrik N and Petri Laukka (2003). "Communication of emotions in vocal expression and music performance: Different channels, same code?" In: *Psychological bulletin* 129.5, p. 770.

Kaczmarek, Lukasz D, Maciej Behnke, Todd B Kashdan, Aleksandra Kusiak, Katarzyna Marzec, Martyna Mistrzak, and Magdalena Włodarczyk (2018). "Smile intensity in social networking profile photographs is related to greater scientific achievements". In: *The Journal of Positive Psychology* 13.5, pp. 435–439.

Kaganovich, Natalya, Jennifer Schumaker, and Courtney Rowland (2016). "Matching heard and seen speech: an ERP study of audiovisual word recognition". In: *Brain and language* 157, pp. 14–24.

Kawakami, Kiyobumi, Kiyoko Takai-Kawakami, Masaki Tomonaga, Juri Suzuki, Tomiyo Kusaka, and Takashi Okai (2006). "Origins of smile and laughter: A preliminary study". In: *Early Human Development* 82.1, pp. 61–66.

Keough, Megan, Avery Ozburn, Elise Kedersha McClay, Michael David Schwan, Murray Schellenberg, Samuel Akinbo, and Bryan Gick (2015). "Acoustic and articulatory qualities of smiled speech". In: *Canadian Acoustics* 43.3.

Keysers, Christian and Valeria Gazzola (2010). "Social neuroscience: mirror neurons recorded in humans". In: *Current biology* 20.8, R353–R354.

– (2018). "Neural Correlates of Empathy in Humans, and the Need for Animal Models". In: *Neuronal Correlates of Empathy*. Elsevier, pp. 37–52.

Keysers, Christian, Evelyne Kohler, M Alessandra Umiltà, Luca Nanetti, Leonardo Fogassi, and Vittorio Gallese (2003). "Audiovisual mirror neurons and action recognition". In: *Experimental brain research* 153.4, pp. 628–636.

Kim, Yanghee, Jeffrey Thayne, and Quan Wei (2016). "An embodied agent helps anxious students in mathematics learning". In: *Educational Technology Research and Development*, pp. 1–17.

Koelsch, Stefan, Thomas Fritz, D Yves v. Cramon, Karsten Müller, and Angela D Friederici (2006). "Investigating emotion with music: an fMRI study". In: *Human brain mapping* 27.3, pp. 239–250.

Kohler, Evelyne, Christian Keysers, M Alessandra Umilta, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti (2002). "Hearing sounds, understanding actions: action representation in mirror neurons". In: *Science* 297.5582, pp. 846–848.

Kontsevich, Leonid L and Christopher W Tyler (2004). "What makes Mona Lisa smile?" In: *Vision research* 44.13, pp. 1493–1498.

KräMer, Nicole, Stefan Kopp, Christian Becker-Asano, and Nicole Sommer (2013). "Smile and the world will smile with you—The effects of a virtual agent's smile on users' evaluation and behavior". In: *International Journal of Human-Computer Studies* 71.3, pp. 335–349.

Kreifelts, Benjamin, Thomas Ethofer, Wolfgang Grodd, Michael Erb, and Dirk Wildgruber (2007). "Audiovisual integration of emotional signals in voice and face: an event-related fMRI study". In: *Neuroimage* 37.4, pp. 1445–1456.

Krumhuber, Eva, Antony SR Manstead, Darren Cosker, Dave Marshall, Paul L Rosin, and Arvid Kappas (2007). "Facial dynamics as indicators of trustworthiness and cooperative behavior." In: *Emotion* 7.4, p. 730.

Ku, Jeonghun, Hee Jeong Jang, Kwang Uk Kim, Jae Hun Kim, Sung Hyouk Park, Jang Han Lee, Jae Jin Kim, In Y Kim, and Sun I Kim (2005). "Experimental results of affective valence and arousal to avatar's facial expressions". In: *CyberPsychology & Behavior* 8.5, pp. 493–503.

Kuhl, Patricia K and Andrew N Meltzoff (1982). "The bimodal perception of speech in infancy". In: *Science* 218.4577, pp. 1138–1141.

– (1996). "Infant vocalizations in response to speech: Vocal imitation and developmental change". In: *The journal of the Acoustical Society of America* 100.4, pp. 2425–2438.

Künecke, Janina, Andrea Hildebrandt, Guillermo Recio, Werner Sommer, and Oliver Wilhelm (2014). "Facial EMG responses to emotional expressions are related to emotion perception ability". In: *PloS one* 9.1, e84053.

Kunz, Miriam, Kenneth Prkachin, and Stefan Lautenbacher (2009). "The smile of pain". In: *Pain* 145.3, pp. 273–275.

Lasarcyk, Eva and Jürgen Trouvain (2008). "Spread lips+ raised larynx+ higher f0= Smiled Speech?-An articulatory synthesis approach". In: *Proceedings of ISSP*, pp. 43–48.

Laurent, Raphaël, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière, and Julien Diard (2017). "The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception." In: *Psychological review* 124.5, p. 572.

LeDoux, Joseph E and Richard Brown (2017). "A higher-order theory of emotional consciousness". In: *Proceedings of the National Academy of Sciences*, p. 201619316.

Lee, DH and AK Anderson (2016). "Form and function in facial expressive behavior". In: *Handbook of emotions*, pp. 495–509.

Lee, Li and Richard C Rose (1996). "Speaker normalization using efficient frequency warping procedures". In: *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. Vol. 1. IEEE, pp. 353–356.

Lehmann, Christoph, Thomas Mueller, Andrea Federspiel, Daniela Hubl, Gerhard Schroth, Oswald Huber, Werner Strik, and Thomas Dierks (2004). "Dissociation between overt and unconscious face processing in fusiform face area". In: *Neuroimage* 21.1, pp. 75–83.

Lemaitre, Guillaume, John A Pyles, Andrea R Halpern, Nicole Navolio, Matthew Lehet, and Laurie M Heller (2017). "Who's that knocking at my door? Neural bases of sound source identification". In: *Cerebral cortex* bhw397, pp. 1–14.

Leopold, David A and Gillian Rhodes (2010). "A comparative view of face perception." In: *Journal of Comparative Psychology* 124.3, p. 233.

Lessard, Nadia, Michael Paré, Franco Lepore, and Maryse Lassonde (1998). "Early-blind human subjects localize sound sources better than sighted subjects". In: *Nature* 395.6699, p. 278.

Levecque, Katia, Frederik Anseel, Alain De Beuckelaer, Johan Van der Heyden, and Lydia Gisle (2017). "Work organization and mental health problems in PhD students". In: *Research Policy* 46.4, pp. 868–879.

Levelt, Willem JM and Stephanie Kelter (1982). "Surface form and memory in question answering". In:

Liberman, Alvin M and Ignatius G Mattingly (1985). "The motor theory of speech perception revised". In: *Cognition* 21.1, pp. 1–36.

Liberman, Alvin M, Franklin S Cooper, Donald P Shankweiler, and Michael Studdert-Kennedy (1967). "Perception of the speech code." In: *Psychological review* 74.6, p. 431.

Likowski, Katja U, Andreas Mühlberger, Antje BM Gerdes, Matthias J Wieser, Paul Pauli, and Peter Weyers (2012). "Facial mimicry and the mirror neuron system: simultaneous acquisition of facial electromyography and functional magnetic resonance imaging". In: *Frontiers in Human Neuroscience*.

Lima, César F, Saloni Krishnan, and Sophie K Scott (2016). "Roles of supplementary motor areas in auditory processing and auditory imagery". In: *Trends in neurosciences* 39.8, pp. 527–542.

Lipps, Theodor (1935). "Empathy, inner imitation, and sense-feelings". In: *A modern book of aesthetics, New York: Holt and Company.(Original work published 1903)*.

Liu, Huei-Mei, Patricia K Kuhl, and Feng-Ming Tsao (2003). "An association between mothers' speech clarity and infants' speech discrimination skills". In: *Developmental Science* 6.3, F1–F10.

Liuni, Marco and Roebel Axel (2013). "Phase vocoder and beyond". In: *Musica, Tecnologia* 7, pp. 73–120.

Lucey, P., J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews (2010). "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In: *Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis, San Francisco, USA*.

Macmillan, Neil A and C Douglas Creelman (2004). *Detection theory: A user's guide*. Psychology press.

Magariños, Carmen, Paula Lopez-Otero, Laura Docio-Fernandez, Eduardo R Banga, Carmen Garcia-Mateo, and Daniel Erro (2016). "Piecewise linear definition of transformation functions for speaker de-identification". In: *Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016 First International Workshop on*. IEEE, pp. 1–5.

Magnée, Maurice JCM, Jeroen J Stekelenburg, Chantal Kemner, and Beatrice de Gelder (2007). "Similar facial electromyographic responses to faces, voices, and body expressions". In: *Neuroreport* 18.4, pp. 369–372.

Maldonado, Heidy, Jong-Eun Roselyn Lee, Scott Brave, Cliff Nass, Hiroshi Nakajima, Ryota Yamada, Kimihiko Iwamura, and Yasunori Morishima (2005). "We learn better together: enhancing elearning with emotional characters". In: *Proceedings of th 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!* International Society of the Learning Sciences, pp. 408–417.

Mangini, Michael C and Irving Biederman (2004). "Making the ineffable explicit: Estimating the information employed for face classifications". In: *Cognitive Science* 28.2, pp. 209–226.

Maris, Eric and Robert Oostenveld (2007). "Nonparametric statistical testing of EEG-and MEG-data". In: *Journal of neuroscience methods* 164.1, pp. 177–190.

Martin, Jared, Magdalena Rychlowska, Adrienne Wood, and Paula Niedenthal (2017). "Smiles as Multipurpose Social Signals". In: *Trends in cognitive sciences*.

Matsumoto, David and Paul Ekman (1989). "American-Japanese cultural differences in intensity ratings of facial expressions of emotion". In: *Motivation and Emotion* 13.2, pp. 143–157.

Matsumoto, David and Tsutomu Kudoh (1993). "American-Japanese cultural differences in attributions of personality based on smiles". In: *Journal of Nonverbal Behavior* 17.4, pp. 231–243.

Matsumoto, David and Bob Willingham (2009). "Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals." In: *Journal of personality and social psychology* 96.1, p. 1.

McGurk, Harry and John MacDonald (1976). "Hearing lips and seeing voices". In: *Nature* 264.5588, p. 746.

Meij, Hans, Jan Meij, and Ruth Harmsen (2015). "Animated pedagogical agents effects on enhancing student motivation and learning in a science inquiry learning environment". In: *Educational technology research and development* 63.3, pp. 381–403.

Meltzoff, Andrew N (1985). "Immediate and deferred imitation in fourteen- and twenty-four-month-old infants". In: *Child Development*, pp. 62–72.

– (1988). "The Human Infant as Homo Imitans". In: *Social Learning: Psychological and Biological Perspectives*, p. 319.

Meltzoff, Andrew N and M Keith Moore (1983). "Newborn infants imitate adult facial gestures". In: *Child development*, pp. 702–709.

– (1997). "Explaining facial imitation: A theoretical model". In: *Infant and child development* 6.3-4, pp. 179–192.

Meltzoff, Andrew N, Lynne Murray, Elizabeth Simpson, Mikael Heimann, Emese Nagy, Jacqueline Nadel, Eric J Pedersen, Rechele Brooks, Daniel S Messinger, Leonardo De Pascalis, et al. (2018). "Re-examination of Oostenbroek et al.(2016): evidence for neonatal imitation of tongue protrusion". In: *Developmental Science* 21.4, e12609.

Mesgarani, Nima, Connie Cheung, Keith Johnson, and Edward F Chang (2014). "Phonetic feature encoding in human superior temporal gyrus". In: *Science* 343.6174, pp. 1006–1010.

Messinger, Daniel, Marco Dondi, G Christina Nelson-Goens, Alessia Beghi, Alan Fogel, and Francesca Simion (2002). "How sleeping neonates smile". In: *Developmental Science* 5.1, pp. 48–54.

Micheletta, Jérôme, Jamie Whitehouse, Lisa A Parr, and Bridget M Waller (2015). "Facial expression recognition in crested macaques (Macaca nigra)". In: *Animal Cognition* 18.4, pp. 985–990.

Minagawa-Kawai, Yasuyo, Sunao Matsuoka, Ippeita Dan, Nozomi Naoi, Katsuki Nakamura, and Shozo Kojima (2008). "Prefrontal activation associated with social attachment: facial-emotion recognition in mothers and infants". In: *Cerebral Cortex* 19.2, pp. 284–292.

Moineau, Suzanne, Nina F Dronkers, and Elizabeth Bates (2005). "Exploring the processing continuum of single-word comprehension in aphasia". In: *Journal of Speech, Language, and Hearing Research* 48.4, pp. 884–896.

Möttönen, Riikka, Rebekah Dutton, and Kate E Watkins (2012). "Auditory-motor processing of speech sounds". In: *Cerebral Cortex* 23.5, pp. 1190–1197.

Mukamel, Roy, Arne D Ekstrom, Jonas Kaplan, Marco Iacoboni, and Itzhak Fried (2010). "Single-neuron responses in humans during execution and observation of actions". In: *Current biology* 20.8, pp. 750–756.

Munhall, Kevin G, P Gribble, L Sacco, and M Ward (1996). "Temporal constraints on the McGurk effect". In: *Perception & psychophysics* 58.3, pp. 351–362.

Murata, Aiko, Hisamichi Saito, Joanna Schug, Kenji Ogawa, and Tatsuya Kameda (2016). "Spontaneous facial mimicry is enhanced by the goal of inferring emotional states: evidence for moderation of "automatic" mimicry by higher cognitive processes". In: *PloS one* 11.4, e0153128.

Natale, Michael (1975). "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." In: *Journal of Personality and Social Psychology* 32.5, p. 790.

Nath, Audrey R and Michael S Beauchamp (2012). "A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion". In: *Neuroimage* 59.1, pp. 781–787.

Navarra, Jordi and Salvador Soto-Faraco (2007). "Hearing lips in a second language: visual articulatory information enables the perception of second language sounds". In: *Psychological research* 71.1, pp. 4–12.

Neal, David T and Tanya L Chartrand (2011). "Embodied emotion perception: amplifying and dampening facial feedback modulates emotion perception accuracy". In: *Social Psychological and Personality Science* 2.6, pp. 673–678.

Neri, Peter (2010). "How inherently noisy is human sensory processing?" In: *Psychonomic Bulletin & Review* 17.6, pp. 802–808.

Neumann, Roland and Fritz Strack (2000). ""Mood contagion": the automatic transfer of mood between persons." In: *Journal of personality and social psychology* 79.2, p. 211.

Niedenthal, Paula M, Markus Brauer, Jamin B Halberstadt, and Åse H Innes-Ker (2001). "When did her smile drop? Facial mimicry and the influences of emotional state on the detection of change in emotional expression". In: *Cognition & Emotion* 15.6, pp. 853–864.

Niedenthal, Paula M, Martial Mermillod, Marcus Maringer, and Ursula Hess (2010). "The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression". In: *Behavioral and brain sciences* 33.06, pp. 417–433.

Noah, Tom, Yaacov Schul, and Ruth Mayo (2018). "When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect." In: *Journal of personality and social psychology* 114.5, p. 657.

Oberman, Lindsay M, Piotr Winkielman, and Vilayanur S Ramachandran (2007). "Face to face: Blocking facial mimicry can selectively impair recognition of emotional expressions". In: *Social neuroscience* 2.3-4, pp. 167–178.

Ochs, Magalie, Catherine Pelachaud, and Gary Mckeown (2017). "A User Perception-Based Approach to Create Smiling Embodied Conversational Agents". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.1, p. 4.

Oh, SY, J Bailenson, N Krämer, and B Li (2016). "Let the Avatar Brighten Your Smile: Effects of Enhancing Facial Expressions in Virtual Environments". In: *PLoS ONE* 11.9, e0161794.

Ohala, John J (1980). "The acoustic origin of the smile". In: *The Journal of the Acoustical Society of America* 68.S1, S33–S33.

Okun, Michael S, Dawn Bowers, Utaka Springer, Nathan A Shapira, Donald Malone, Ali R Rezai, Bart Nuttin, Kenneth M Heilman, Robert J Morecraft, Steven A Rasmussen, et al. (2004). "What's in a "smile?" Intra-operative observations of contralateral smiles induced by deep brain stimulation". In: *Neurocase* 10.4, pp. 271–279.

Oliva, Manuel and Andrey Anikin (2018). "Pupil dilation reflects the time course of emotion recognition in human vocalizations". In: *Scientific reports* 8.1, p. 4871.

Olsen, Anneli (2012). "The Tobii I-VT fixation filter". In: *Tobii Technology*.

Oostenbroek, Janine, Thomas Suddendorf, Mark Nielsen, Jonathan Redshaw, Siobhan Kennedy-Costantini, Jacqueline Davis, Sally Clark, and Virginia Slaughter (2016). "Comprehensive longitudinal study challenges the existence of neonatal imitation in humans". In: *Current Biology* 26.10, pp. 1334–1338.

Oostenbroek, Janine, Jonathan Redshaw, Jacqueline Davis, Siobhan Kennedy-Costantini, Mark Nielsen, Virginia Slaughter, and Thomas Suddendorf (2018). "Re-evaluating the neonatal imitation hypothesis." In: *Developmental science*, e12720–e12720.

Owen Brimijoin, W, Michael A Akeroyd, Emily Tilbury, and Bernd Porr (2013). "The internal representation of vowel spectra investigated using behavioral response-triggered averaging". In: *The Journal of the Acoustical Society of America* 133.2, EL118–EL122.

Pampalk, Elias (2004). "A Matlab Toolbox to Compute Similarity from Audio". In: *Proceedings of the ISMIR International Conference on Music Information Retrieval, Barcelona, Spain*.

Panksepp, Jaak (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

– (2007). "Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist". In: *Perspectives on psychological science* 2.3, pp. 281–296.

Parr, Lisa A and Bridget M Waller (2006). "Understanding chimpanzee facial expression: insights into the evolution of communication". In: *Social Cognitive and Affective Neuroscience* 1.3, pp. 221–228.

Partala, Timo and Veikko Surakka (2004). "The effects of affective interventions in human–computer interaction". In: *Interacting with computers* 16.2, pp. 295–309.

Paulmann, Silke and Marc D Pell (2011). "Is there an advantage for recognizing multi-modal emotional stimuli?" In: *Motivation and Emotion* 35.2, pp. 192–201.

Paulmann, Silke, Debra Titone, and Marc D Pell (2012). "How emotional prosody guides your way: evidence from eye movements". In: *Speech Communication* 54.1, pp. 92–107.

Peirce, Jonathan W (2008). "Generating stimuli for neuroscience using PsychoPy". In: *Frontiers in neuroinformatics* 2.

Pelphrey, Kevin A, Noah J Sasson, J Steven Reznick, Gregory Paul, Barbara D Goldman, and Joseph Piven (2002). "Visual scanning of faces in autism". In: *Journal of autism and developmental disorders* 32.4, pp. 249–261.

Pessoa, Luiz, Shruti Japee, David Sturman, and Leslie G Ungerleider (2005). "Target visibility and visual awareness modulate amygdala responses to fearful faces". In: *Cerebral cortex* 16.3, pp. 366–375.

Petkov, Christopher I, Christoph Kayser, Thomas Steudel, Kevin Whittingstall, Mark Augath, and Nikos K Logothetis (2008). "A voice region in the monkey brain". In: *Nature neuroscience* 11.3, p. 367.

Phillips, Mary L, Andy W Young, Carl Senior, Michael Brammer, Chris Andrew, Andrew J Calder, Edward T Bullmore, David I Perrett, Duncan Rowland, Steven CR Williams, et al. (1997). "A specific neural substrate for perceiving facial expressions of disgust". In: *Nature* 389.6650, p. 495.

Podesva, Robert J, Patrick Callier, Rob Voigt, and Dan Jurafsky (2015). "The connection between smiling and GOAT fronting: Embodied affect in sociophonetic variation". In: *Proceedings of the International Congress of Phonetic Sciences*. Vol. 18.

Ponsot, Emmanuel, Pablo Arias, and Jean-Julien Aucouturier (2018). "Uncovering mental representations of smiled speech using reverse correlation". In: *The Journal of the Acoustical Society of America* 143.1, EL19–EL24.

Ponsot, Emmanuel, Juan José Burred, Pascal Belin, and Jean-Julien Aucouturier (2018). "Cracking the social code of speech prosody using reverse correlation". In: *Proceedings of the National Academy of Sciences* 115.15, pp. 3972–3977.

Prochazkova, Eliska, Luisa Prochazkova, Michael Rojek Giffin, H Steven Scholte, Carsten KW De Dreu, and Mariska E Kret (2018). "Pupil mimicry promotes trust through the theory-of-mind network". In: *Proceedings of the National Academy of Sciences*, p. 201803916.

Provine, Robert R (2004). "Laughing, tickling, and the evolution of speech and self". In: *Current Directions in Psychological Science* 13.6, pp. 215–218.

Pulvermüller, Friedemann, Martina Huss, Ferath Kherif, Fermin Moscoso del Prado Martin, Olaf Hauk, and Yury Shtyrov (2006). "Motor cortex maps articulatory features of speech sounds". In: *Proceedings of the National Academy of Sciences* 103.20, pp. 7865–7870.

Quené, Hugo, Gün R Semin, and Francesco Foroni (2012). "Audible smiles and frowns affect speech comprehension". In: *Speech Communication* 54.7, pp. 917–922.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rachman, Laura, M. Liuni, P. Arias, A. Lind, P. Johansson, L. Hall, D. Richardson, K. Watanabe, S. Dubal, and J.J. Aucouturier (2017). "DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech". In: *Behavior Research Methods (in press)*.

Reber, Stephan A, Judith Janisch, Kevin Torregrosa, Jim Darlington, Kent A Vliet, and W Tecumseh Fitch (2017). "Formants provide honest acoustic cues to body size in American alligators". In: *Scientific reports* 7.1, p. 1816.

Reis, Harry T, Ilona McDougal Wilson, Carla Monestere, Stuart Bernstein, Kelly Clark, Edward Seidl, Michelle Franco, Ezia Gioioso, Lori Freeman, and Kimberly Radoane (1990). "What is smiling is beautiful and good". In: *European Journal of Social Psychology* 20.3, pp. 259–267.

Renner, Lena F and Marcin Włodarczak (2017). "When a dog is a cat and how it changes your pupil size: Pupil dilation in response to information mismatch". In: *Proc. Interspeech 2017*, pp. 674–678.

Rigoulot, Simon and Marc D Pell (2012). "Seeing emotion with your ears: emotional prosody implicitly guides visual attention to faces". In: *PloS one* 7.1, e30740.

– (2014). "Emotion in the voice influences the way we scan emotional faces". In: *Speech Communication* 65, pp. 36–49.

Rives Bogart, Kathleen and David Matsumoto (2010). "Facial mimicry is not necessary to recognize emotion: Facial expression recognition by people with Moebius syndrome". In: *Social Neuroscience* 5.2, pp. 241–251.

Röbel, Axel and Xavier Rodet (2005). "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation". In: *International Conference on Digital Audio Effects*, pp. 30–35.

Roebel, Axel (2010). "Shape-invariant speech transformation with the phase vocoder". In: *InterSpeech*, pp. 2146–2149.

Rosenblum, Lawrence D, Mark A Schmuckler, and Jennifer A Johnson (1997). "The McGurk effect in infants". In: *Perception & Psychophysics* 59.3, pp. 347–357.

Rosenthal, Robert (1979). "The file drawer problem and tolerance for null results." In: *Psychological bulletin* 86.3, p. 638.

Russ, Jeff B, Ruben C Gur, and Warren B Bilker (2008). "Validation of affective and neutral sentence content for prosodic testing". In: *Behavior research methods* 40.4, pp. 935–939.

Rychlowska, Magdalena, Yuri Miyamoto, David Matsumoto, Ursula Hess, Eva Gilboa-Schechtman, Shanmukh Kamble, Hamdi Muluk, Takahiko Masuda, and Paula Marie Niedenthal (2015). "Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles". In: *Proceedings of the National Academy of Sciences*, p. 201413661.

Rychlowska, Magdalena, Rachael E Jack, Oliver GB Garrod, Philippe G Schyns, Jared D Martin, and Paula M Niedenthal (2017). "Functional smiles: Tools for love, sympathy, and war". In: *Psychological science* 28.9, pp. 1259–1270.

Sauter, Disa A, Frank Eisner, Paul Ekman, and Sophie K Scott (2010). "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations". In: *Proceedings of the National Academy of Sciences* 107.6, pp. 2408–2412.

Schaefer, Scott, Travis McPhail, and Joe Warren (2006). "Image deformation using moving least squares". In: *ACM transactions on graphics (TOG)*. Vol. 25. 3. ACM, pp. 533–540.

Scheider, Linda, Bridget M Waller, Leonardo Oña, Anne M Burrows, and Katja Liebal (2016). "Social use of facial expressions in hylobatids". In: *PloS one* 11.3, e0151733.

Schwartz, Jean-Luc, Frédéric Berthommier, and Christophe Savariaux (2004). "Seeing to hear better: evidence for early audio-visual interactions in speech identification". In: *Cognition* 93.2, B69–B78.

Scott, Sophie K (2016). "Perception and production of speech: Connected, but how?" In: *Speech Perception and Spoken Word Recognition*. Psychology Press, pp. 33–46.

Sekuler, Robert (1997). "Sound alters visual motion perception". In: *Nature* 385.6614, p. 308.

Shams, Ladan, Yukiyasu Kamitani, and Shinsuke Shimojo (2000). "Illusions: What you see is what you hear". In: *Nature* 408.6814, p. 788.

Simner, Marvin L (1971). "Newborn's response to the cry of another infant." In: *Developmental psychology* 5.1, p. 136.

Singer, Tania, Ben Seymour, John O'doherty, Holger Kaube, Raymond J Dolan, and Chris D Frith (2004). "Empathy for pain involves the affective but not sensory components of pain". In: *Science* 303.5661, pp. 1157–1162.

Smith, PL and DR Little (2018). "Small is beautiful: In defense of the small-N design." In: *Psychonomic bulletin & review*.

Srinivasan, Ramprakash, Julie D Golomb, and Aleix M Martinez (2016). "A neural basis of facial action recognition in humans". In: *Journal of Neuroscience* 36.16, pp. 4434–4442.

Stasenko, Alena, Frank E Garcea, and Bradford Z Mahon (2013). "What happens to the motor theory of perception when the motor system is damaged?" In: *Language and cognition* 5.2-3, pp. 225–238.

Stel, Marille and Ad van Knippenberg (2008). "The role of facial mimicry in the recognition of affect". In: *Psychological Science* 19.10, p. 984.

Stevens, Kenneth N (2000). *Acoustic phonetics*. Vol. 30. MIT press.

Story, Brad H (2015). "Mechanisms of voice production". In: *The handbook of speech production*, pp. 34–58.

Story, Brad H and Kate Bunton (2017). "Vowel space density as an indicator of speech performance". In: *The Journal of the Acoustical Society of America* 141.5, EL458–EL464.

Strack, Fritz, Leonard L Martin, and Sabine Stepper (1988). "Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis." In: *Journal of personality and social psychology* 54.5, p. 768.

Street Jr, Richard L (1984). "Speech convergence and speech evaluation in fact-finding interviews". In: *Human Communication Research* 11.2, pp. 139–169.

Sumby, William H and Irwin Pollack (1954). "Visual contribution to speech intelligibility in noise". In: *The journal of the acoustical society of america* 26.2, pp. 212–215.

Susskind, Joshua M, Daniel H Lee, Andrée Cusi, Roman Feiman, Wojtek Grabski, and Adam K Anderson (2008). "Expressing fear enhances sensory acquisition". In: *Nature neuroscience* 11.7, p. 843.

Tamietto, Marco and Beatrice De Gelder (2010). "Neural bases of the nonconscious perception of emotional signals". In: *Nature Reviews Neuroscience* 11.10, p. 697.

Tamietto, Marco, Lorys Castelli, Sergio Vighetti, Paola Perozzo, Giuliano Geminiani, Lawrence Weiskrantz, and Beatrice de Gelder (2009). "Unseen facial and bodily expressions trigger fast emotional reactions". In: *Proceedings of the National Academy of Sciences* 106.42, pp. 17661–17666.

Tartter, Vivien C (1980). "Happy talk: Perceptual and acoustic effects of smiling on speech". In: *Attention, Perception, & Psychophysics* 27.1, pp. 24–27.

Tartter, Vivien C and David Braun (1994). "Hearing smiles and frowns in normal and whisper registers". In: *The Journal of the Acoustical Society of America* 96.4, pp. 2101–2107.

Thoret, Etienne, Philippe Depalle, and Stephen McAdams (2016). "Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments". In: *The Journal of the Acoustical Society of America* 140.6, EL478–EL483.

Tian, Xiaohai, Zhizheng Wu, Siu Wa Lee, and Eng Siong Chng (2014). "Correlation-based frequency warping for voice conversion". In: *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, pp. 211–215.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai (2014). "Mediation: R Package for Causal Mediation Analysis". In: *Journal of Statistical Software* 59.5.

Tomkins, Silvan (1962a). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

Tomkins, S.S. (1962b). *Affect, imagery and consciousness: The positive affects*. New York:Springer-Verlag.

Valente, Dannyelle, Anne Theurel, and Edouard Gentaz (2017). "The role of visual experience in the production of emotional facial expressions by blind people: a review". In: *Psychonomic bulletin & review*, pp. 1–15.

Van Boxtel, Anton (2010). "Facial EMG as a tool for inferring affective states". In: *Proceedings of measuring behavior*. Noldus Information Technology Wageningen, pp. 104–108.

Van Hooff, Jaram (1972). "A comparative approach to the phylogeny of laughter and smile". In: *Nonverbal Communication*.

Varnet, Léo, Fanny Meunier, Gwendoline Trollé, and Michel Hoen (2016). "Direct viewing of dyslexics' compensatory strategies in speech in noise using auditory classification images". In: *PloS one* 11.4, e0153781.

Venezia, Jonathan H, Gregory Hickok, and Virginia M Richards (2016). "Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility". In: *The Journal of the Acoustical Society of America* 140.2, pp. 1072–1088.

Verona, Edelyn, Christopher J Patrick, John J Curtin, Margaret M Bradley, and Peter J Lang (2004). "Psychopathy and physiological response to emotionally evocative sounds." In: *Journal of abnormal psychology* 113.1, p. 99.

Villavicencio, Fernando, Axel Robel, and Xavier Rodet (2006). "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation". In: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. Vol. 1. IEEE, pp. I–I.

Vrana, Scott R, Ellen L Spence, and Peter J Lang (1988). "The startle probe response: a new measure of emotion?" In: *Journal of abnormal psychology* 97.4, p. 487.

Vroomen, Jean, Jon Driver, and Beatrice De Gelder (2001). "Is cross-modal integration of emotional expressions independent of attentional resources?" In: *Cognitive, Affective, & Behavioral Neuroscience* 1.4, pp. 382–387.

Wagenmakers, E-J, Titia Beek, Laura Dijkhoff, Quentin F Gronau, A Acosta, RB Adams Jr, DN Albohn, ES Allard, SD Benning, E-M Blouin-Hudon, et al. (2016). "Registered Replication Report: Strack, Martin, & Stepper (1988)". In: *Perspectives on Psychological Science* 11.6, pp. 917–928.

Wan, Catherine Y, Amanda G Wood, David C Reutens, and Sarah J Wilson (2010). "Early but not late-blindness leads to enhanced auditory perception". In: *Neuropsychologia* 48.1, pp. 344–348.

Wang, Shuo, Ming Jiang, Xavier Morin Duchesne, Elizabeth A Laugeson, Daniel P Kennedy, Ralph Adolphs, and Qi Zhao (2015). "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking". In: *Neuron* 88.3, pp. 604–616.

Warren, Jane E, Disa A Sauter, Frank Eisner, Jade Wiland, M Alexander Dresner, Richard JS Wise, Stuart Rosen, and Sophie K Scott (2006). "Positive emotions preferentially engage an auditory–motor "mirror" system". In: *Journal of Neuroscience* 26.50, pp. 13067–13075.

Watkins, Kate E, Antonio P Strafella, and Tomáš Paus (2003). "Seeing and hearing speech excites the motor system involved in speech production". In: *Neuropsychologia* 41.8, pp. 989–994.

Watson, Rebecca, Marianne Latinus, Takao Noguchi, Oliver George Baring Garrod, Frances Crabbe, and Pascal Belin (2013). "Dissociating task difficulty from incongruence in face-voice emotion integration". In: *Frontiers in human neuroscience* 7, p. 744.

Watson, Rebecca, Marianne Latinus, Takao Noguchi, Oliver Garrod, Frances Crabbe, and Pascal Belin (2014). "Crossmodal adaptation in right posterior superior temporal sulcus during face–voice emotional integration". In: *Journal of Neuroscience* 34.20, pp. 6813–6821.

Wel, Pauline van der and Henk van Steenbergen (2018). "Pupil dilation as an index of effort in cognitive control tasks: A review". In: *Psychonomic bulletin & review*, pp. 1–11.

Wicker, Bruno, Christian Keysers, Jane Plailly, Jean-Pierre Royet, Vittorio Gallese, and Giacomo Rizzolatti (2003). "Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust". In: *Neuron* 40.3, pp. 655–664.

Williams, Kipling D (2002). *Ostracism: The power of silence*. Guilford Press.

Wilson, Stephen M, Ayşe Pinar Saygin, Martin I Sereno, and Marco Iacoboni (2004). "Listening to speech activates motor areas involved in speech production". In: *Nature neuroscience* 7.7, p. 701.

Wörmann, Viktoriya, Manfred Holodynski, Joscha Kärtner, and Heidi Keller (2012). "A cross-cultural comparison of the development of the social smile: A longitudinal study of maternal and infant imitation in 6-and 12-week-old infants". In: *Infant Behavior and Development* 35.3, pp. 335–347.

Yee, Nick, Jeremy N Bailenson, and Kathryn Rickertsen (2007). "A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, pp. 1–10.

Yoon, Jennifer MD and Claudio Tennie (2010). "Contagious yawning: a reflection of empathy, mimicry, or contagion?" In: *Animal Behaviour* 79.5, e1–e3.

Yoshida, Shigeo, Tomohiro Tanikawa, Sho Sakurai, Michitaka Hirose, and Takuji Narumi (2013). "Manipulation of an emotional experience by real-time deformed facial feedback". In: *Proceedings of the 4th Augmented Human International Conference*. ACM, pp. 35–42.

Yu, Hui, Oliver GB Garrod, and Philippe G Schyns (2012). "Perception-driven facial expression synthesis". In: *Computers & Graphics* 36.3, pp. 152–162.

Zafeiriou, DI, A Ververi, and E Vargiami (2007). "Childhood autism and associated comorbidities." In: *Brain & development* 29.5, pp. 257–272.

Zajonc, Robert B, Pamela K Adelmann, Sheila T Murphy, and Paula M Niedenthal (1987). "Convergence in the physical appearance of spouses". In: *Motivation and emotion* 11.4, pp. 335–346.

Zaki, Jamil and Kevin Ochsner (2016). "Chapter 50 : Empathy". In: *Handbook of emotions, Fourth Edition*. Ed. by L. F. Barrett, M. Lewis, and J. M. Haviland-Jones. NY: Guilford Press.

Zaki, Jamil and Kevin N Ochsner (2012). "The neuroscience of empathy: progress, pitfalls and promise". In: *Nature neuroscience* 15.5, p. 675.

Zilbovicius, Monica, Ana Saitovitch, Traian Popa, Elza Rechtman, Lafina Diamandis, Nadia Chabane, and N Boddaert (2013). "Autism, social cognition and superior temporal sulcus". In: *Open Journal of Psychiatry* 3.02, p. 46.