The ”other-voice” effect : how speaker identity and language familiarity influence the way we process emotional speech

Laura Rachman

Ecole Doctorale Cerveau Cognition Comportement
IRCAM - STMS - CNRS UMR 9912 - Perception et Design Sonore
ICM - INSERM U1127 - CNRS UMR S1127 - Social and Affective Neuroscience

Thèse de doctorat

The “other-voice” effect: how speaker identity and language familiarity influence the way we process emotional speech

Laura Rachman

Pour obtenir le grade de Docteur de Sorbonne Université

Sous la direction de Mme Stéphanie Dubal and M Jean-Julien Aucouturier

Soutenue le 7 décembre 2018 devant le jury composé de

Rapportrice	Mme Sonja Kotz
Rapporteur	M Pascal Belin
Examinatrice	Mme Disa Sauter
Examinatrice	Mme Marie Gomot
Examinateur	M Mohamed Chetouani
Directrice de thèse	Mme Stéphanie Dubal
Directeur de thèse	M Jean-Julien Aucouturier

Laura Rachman
The "other-voice" effect: how speaker identity and language familiarity influence the way we process emotional speech
Doctoral thesis, 7 December 2018
Supervised by : Stéphanie Dubal and Jean-Julien Aucouturier
This work was supported by an ERC Starting Grant awarded to Jean-Julien Aucouturier (ERC CREAM 335536)

Sorbonne Université
Perception et Design Sonore
STMS - IRCAM
1 place Stravinsky
75004 Paris

Abstract

The human voice is a powerful tool to convey emotions. Humans hear voices on a daily basis and are able to rapidly extract relevant information to successfully interact with others. The theoretical aim of this thesis is to investigate the role of familiarity on emotional voice processing. Chapters 2 and 3 present behavioral and electrophysiological studies investigating how self- versus non self-produced voices influence the processing of emotional speech utterances. By contrasting self and other, familiarity is here assessed at a personal level. The results of Chapter 2 show a dissociation of explicit and implicit processing of the self-voice. While explicit discrimination of an emotional self-voice and other-voice was somewhat impaired, implicit self-processing prompted a self-advantage in emotion recognition and speaker discrimination. Chapter 3 reports a prioritization for the non-self voice in the processing of emotional and low-level acoustic changes, reflected in faster electrophysiological (EEG) and behavioral responses. In Chapter 4, the effect of voice familiarity on is assessed at a larger sociocultural scale by comparing speech utterances in the native and a foreign language. Taken together, this thesis highlights some ways in which the 'otherness' of a voice - whether a non-self speaker or a foreign language speaker - is processed with a higher priority on the one hand, but with less acoustic precision on the other hand.

Résumé

L’objectif théorique de cette thèse est d’étudier le rôle de la familiarité vocale sur le traitement de la voix émotionnelle. Les chapitres 2 et 3 présentent des études comportementales et électrophysiologiques portant sur les contributions spécifiques de la voix du self et la voix de l’autre sur le traitement de la parole émotionnelle. En comparant le self et l’autre, la familiarité est évaluée ici à un niveau personnel. Les résultats du chapitre 2 montrent une dissociation chez les participants des traitements explicites et implicites de leur propre voix. Alors que la discrimination explicite de leur propre voix émotionnelle est réduite, le traitement implicite de soi entraîne un avantage pour la reconnaissance des émotions et la discrimination du locuteur. Le chapitre 3 montre que les voix inconnues sont priorisées par rapport à la voix du self dans le traitement des changements émotionnels et acoustiques de bas niveau, par des réponses électrophysiologiques (EEG) et comportementales plus rapides. Au chapitre 4, l’effet de la familiarité sur la perception des émotions vocales est évalué au niveau socioculturel en comparant la langue maternelle et étrangère. Au travers de ces études, cette thèse met en évidence les différentes manières par lesquelles «l’étrangeté» d’une voix - qu’il s’agisse d’un locuteur autre que le soi ou d’une langue étrangère - est traitée avec une priorité plus élevée, mais une précision acoustique diminuée.

Acknowledgements

First and foremost I would like to sincerely thank my thesis directors Stéphanie Dubal and Jean-Julien Aucouturier for their support and guidance throughout the past few years. Stéphanie, I am so happy to have had the opportunity to do my final master's project under your supervision and to continue working with you during this doctorate. Your trust in me and your patience, insights, and kindness have meant a lot to me throughout this journey.

Jean-Julien, your motivation and enthusiasm in building the CREAM project have been an absolute inspiration to me. I am so impressed by the energy you bring to the work and your unwavering involvement in everyone's projects, both from near and farther away. Working with you opened my (academic) life to a whole new world of sound technologies and it has been a pleasure to witness your creativity in building bridges between different disciplines.

Pablo and Marco, we have started out in the CREAM project at the same time and as a newbie in the audio world I am thankful to have had you by my side from the beginning. Pablo, I don't know how to express my gratitude for the many many (many!) times you have helped me out with Python problems, audio set-ups and the trials and tribulations with the EEG equipment. No matter the issue, you were always willing to get to the bottom of the problem and try to solve it. Marco, I can't believe we shared a desk for the first year and a half. I have since taken over the whole desk, but it is by no means a sign that I didn't like working with you side by side. Thank you for your invaluable presence in the project and for bringing DAVID to life, this thesis wouldn't be what it is without you. Thanks also to Louise, Emmanuel and most recently Daniel, whom I've shared the office with as well throughout my PhD studies, for contributing to a great atmosphere in the lab.

Vasso, thank you for your assistance and for taking care of the practical things that supported this work. Of almost equal value are all the delicious treats you brought us from Greece, you definitely provided a sweet side to this thesis.

Various shorter-term members of the CREAM project have helped me with the data collection for this thesis. I am thankful to Hugo, Maël, Louise (Loulou), and Gabriël for helping me out with running various experiments. Thanks also to Sarah who,

despite having her own internship project, helped me out on several occasions with installing and preparing participants for the EEG recording sessions. Tomoya, thank you for your work on the cross-cultural studies in this thesis and for welcoming us in Japan, this visit has been a personal highlight in this project.

Lou, you were the first intern that I supervised and it has been a pleasure working with you. Thank you for enthusiasm to partake in an exciting EEG study with me. Unfortunately the study did not make it into the thesis, we bit off a bit more than we could chew. Thank you nevertheless for your contribution during this work, tackling those EEG analyses is the first thing on the list after submitting this thesis.

I would also like to thank all the other members of the Perception and Sound Design team at IRCAM and the Social and Affective Neuroscience team at the ICM for welcoming me among them. I would especially like to express my gratitude to Nathalie George who was always willing to spend a moment sharing her invaluable knowledge of MEG and EEG data analyses with me.

Almost all the experiments presented in this thesis have been conducted at the INSEAD-Sorbonne Université Behavioural Lab and I would like to thank Liselott and Huong for running this place so efficiently, you have made the data collection during this project so much easier for me. Not any less important is the welcoming place you provided, it was always a joy to spend my days in your lab. Thanks also to Jean-Yves, Germain, and Seb for helping create this nice environment and for their assistance in running each experiment smoothly.

Furthermore, I would like to thank the people who so kindly agreed to evaluate this work and to join my jury: Sonja Kotz, Pascal Belin, Disa Sauter and Mohamed Chetouani. A special thanks to Marie Gomot who, besides being a member of the jury, took part in previous evaluations of this project and has supported this work by her thoughtful feedback at various occasions.

Finally, I would like to thank my friends, my parents and my sister for their love, support and wonderful company, not only throughout the realization of this thesis, but also long before.

Contents

Introduction

1

> *It is a peculiar truth that I see far less of myself than other people do. I can see my fingers typing when I look down at them. I can examine my shoes, the details of a shirt cuff, or admire a pair of new tights on my legs while I am sitting down, but the mirror is the only place where I am whole to myself. Only then do I see my body as others see it. But does my mirror-self really represent my persona in the world?*
>
> — **Siri Hustvedt**
> (Living, Thinking, Looking)

Humans are social beings and most of us interact with others on a daily basis. During social interactions, we are constantly monitoring other people's mental and emotional states in order to respond appropriately. However, in these interactions, we do not often see ourselves in the way we see others. When talking to someone, we can judge their level of involvement in the conversation from whether their eyes are focused on us or darting around the surrounding environment, but we cannot see our own eyes from a bystander's perspective; we can see whether someone has trouble understanding something by the frown that forms between their eyebrows, but we cannot see our own forehead wrinkling, only how our muscles are activated to produce this frown. Many of the cues that indicate our emotional state are not perceived through the same channels by the actor and the observer.

However, this limitation of the visual self is not present when we consider the auditory self. When we talk to someone else, not only does the other person hear what we have to say, but we also hear our own voice at the same time. When we communicate with our voice, we also perceive ourselves through the auditory channel and as such, we have become the actor and the observer in the way that others might observe us. The aim of this thesis is to investigate how the cognitive and neural processing of emotional voices depends on whether the voice is the self-voice or the voice of someone else, and more generally, whether emotional expressions are processed differently in familiar (e.g. one's native language) and unfamiliar (e.g. a foreign language) speech sounds.

Throughout this thesis, I will present the results of an electroencephalography (EEG) study as well as a series of behavioral experiments. Two common threads run through these experiments, one theoretical, and the other, methodological. From a theoretical point of view there are two subjects of interest in this thesis: speaker identity and language familiarity. These two subjects can be seen as different types of familiarity as people have generally been exposed to their own voice as well as their native language throughout their whole lives. Speech perception is a multidimensional process in which different types of information are processed in independent but potentially interacting subsystems that specialize e.g. in speech context, language, speaker identity and emotional expressions. The **theoretical contribution** of this thesis is to understand how familiarity with certain speech sounds - our own voice or our own language - influence emotion perception, thereby casting new light on the general organization of the cognition of human language.

From a methodological point of view, all the stimuli in this thesis have been created in the same manner, using a novel emotional transformation tool called DAVID. DAVID allows to take in real voice recordings (e.g. from participants who came in the lab to record isolated words or short sentences), and transform their acoustics properties to change their emotional expression, e.g. in the direction of being more happy or sad. DAVID is based on technologies that existed prior to this work, but had only been applied to entertainment or interactive contexts such as computer music or video games. The **methodological contribution** of this thesis is to test the applicability of technology like DAVID for experimental research, and to the study of vocal emotional processing in particular. In the last chapter I will discuss what the use of such transformation tools can bring to this research community, as well as its challenges and limitations.

1.1 The human voice signals identity and physical traits

The human voice provides a powerful tool to convey information. Among many competing hypotheses, the evolution of language was plausibly driven by increasing group size, for which it provides an evolutionary advantage over gestures and grooming (Dunbar et al., 1998). Language, in effect, is an efficient form of social grooming: while grooming takes place between two individuals, language can reach multiple individuals at the same time and a person does not have to be seen by or in touching distance of someone else in order to communicate vocally. From the sound of a voice, one can accurately determine various traits and physical characteristics of the speaker, such as whether the speaker is male or female (Mullennix et al., 1995), old or young (Smith and Patterson, 2005), someone we know or a stranger (Latinus et al., 2013). In this section I will first review the relevant acoustics characteristics with which speaker physical traits are conveyed, and then how these identity cues are processed in the brain.

1.1.1 Production

Vocal sounds are produced by the acoustic interaction of the larynx, which produces the sound source, and the vocal tract, which filters that source. When air from the lungs passes through the larynx, the vocal folds oscillate and the rate of these oscillations determines the pitch of the produced sound, where faster oscillation rates result in higher sounds (Titze and Strong, 1975). The vocal tract then acts as a filter where band-pass filters (or formants) cause the blocking of certain frequencies (Fitch, 2000; Lieberman et al., 1992).

Because the rate at which the vocal folds oscillate depends on the size and tension of the larynx, voice pitch can be an indicator of speaker characteristics such as sex (Mullennix et al., 1995) and age (Smith and Patterson, 2005). Because adult females typically have smaller vocal folds than males, female voices generally have a higher fundamental frequencies than male voices (Titze and Strong, 1975). (Note that sex differences in pitch are not merely to differences in body size, but have also evolved though sexual selection via female mate choice or male dominance competition - Puts et al., 2006). While pitch is not actually closely related to body size within sexes (see e.g. Pisanski et al., 2014), people's judgments of dominance and attractiveness can nevertheless be affected by average pitch and pitch variability (Hodges-Simeon et al., 2010), and lower pitch facilitates also accurate size assessment (Pisanski et al., 2014).

Formant information, related to the shape and size of the vocal tract, also conveys important speaker characteristics (Fitch and Giedd, 1999). In particular, formant dispersion, the distance between the frequency of the first and second resonance in the voice's spectral envelope, scales negatively with vocal tract length and is therefore a reliable signal of body size and physical strength (Sell et al., 2010). By shaping vocal timbre, formants (and how they change dynamically during speech) are also an important signature of speaker identity (McDougall, 2004; Latinus et al., 2013). Finally, because the different voiced phonemes of a language are produced by dynamically switching vocal tract configuration (Fitch and Giedd, 1999), formants are indicated of how these phonemes are pronounced (e.g. accents, Ferragne and Pellegrino, 2010) and what these phonemes are (e.g. foreign language, Flege, 1987).

1.1.2 Perception

The vocal signal, like other auditory signals, is sent from the cochlea to the medial geniculate body of the thalamus and the primary auditory cortex in the temporal lobe for low-level auditory analysis. Voice perception is a multidimensional process, and Belin and colleagues (2004) propose that different characteristics of the voice are processed along independent pathways that potentially interact with each other (Figure 1.1). This model of voice perception is based upon a model of face perception by Bruce and Young (1986), which involves three stages: first, the structural encoding of facial features and facial layout, such as the configuration of the mouth or the distance between the eyes; second, the analysis of emotional expression in which a smiling mouth for instance is interpreted as a positive affective signal; third, the incoming information is compared to prior knowledge of familiar faces which can involve both physical characteristics such as eye colour, as well as contextual information such as the job this person has or memories of specific encounters. The model of Belin and colleagues (2004) breaks down the process of voice perception in an analogous way, and posits that different subsystems exists for the perception of the speech content, the speaker identity, and the emotional expression of the voice. These subsystems can influence each other: the familiarity of a speaker can help speech intelligibility in challenging listening conditions (Johnsrude et al., 2013; Levi et al., 2011) and different speakers are better distinguished from each other in one's native language compared to in a foreign language (Fleming et al., 2014; Perrachione et al., 2011).

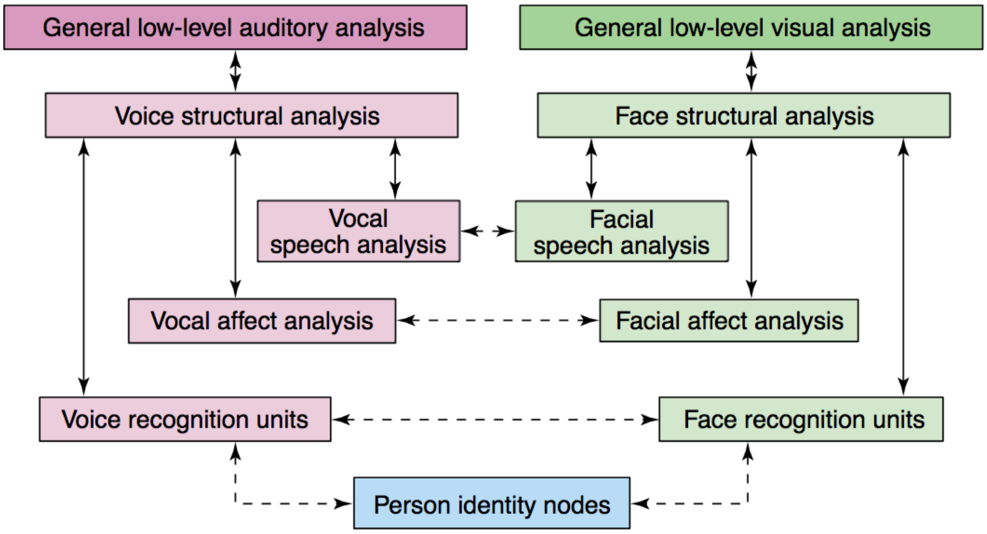


Fig. 1.1: Model of voice perception, adapted from Belin et al. (2004)

The middle and anterior part of the bilateral superior temporal gyrus (STG) and sulcus (STS) contain voice-selective regions called the temporal voice areas (TVA) (Figure 1.2). These areas are prefentially activated for voice sounds as opposed to non-voice sounds. While the strongest responses can be seen to speech signal, non-speech vocal signals have also been found to activate the TVAs. Electrophysiological results show that voices can be distinguished from other sound categories from around 200 ms after sound onset (Charest et al., 2009). This voice specific response at the latency of the auditory P2 has been reported in children as young as 4 years old (Rogier et al., 2010).

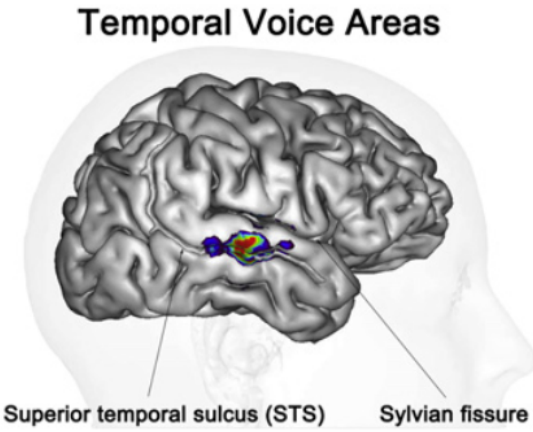
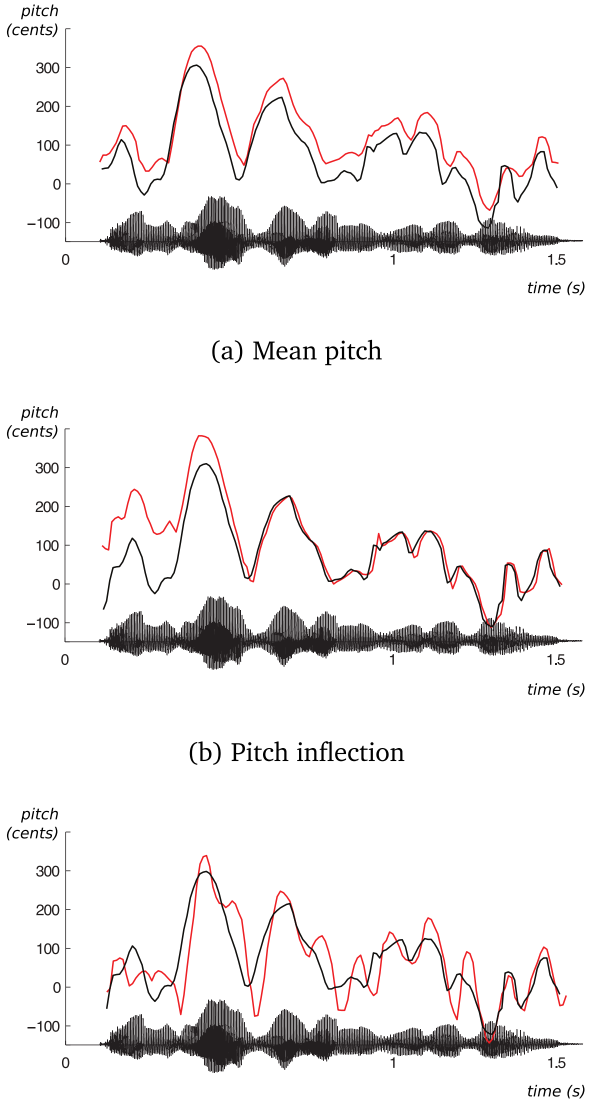


Fig. 1.2: Temporal voice areas (TVAs) along the superior temporal sulcus (STS) of the human brain, adapted from Belin and Grosbras (2010)

The TVAs are also involved in speaker identity analysis (von Kriegstein and Giraud, 2004; Roswandowitz et al., 2018; Warren et al., 2006). As described in section 1.1.1, the pitch of a voice is determined by the oscillation rate of the vocal folds. This characteristic is an important cue in speaker identity perception along with the spectral envelope of speech which is modulated by the vocal tract length. The prototype model of speaker identity processing describes that each speaker identity that is perceived is compared to a prototype voice (Latinus et al., 2013). In a series of experiments, Latinus and colleagues (2013) created a male and a female prototype voice by averaging a large number of individual voice samples of each gender. Looking at the fundamental frequency, formant dispersion, and the harmonic-to-noise ratio, they found that the perceived distinctiveness of an individual voice increased as these characteristics deviated more from the prototype voice. Moreover, in a subsequent functional magnetic resonance imaging (fMRI) study, speaker distinctiveness was found to correlate with neural activity in the bilateral TVAs.



(a) Mean pitch

(b) Pitch inflection

(c) Vibrato

Fig. 1.3: Three examples of auditory cues for emotional expression as implemented in DAVID (Rachman et al., 2018). The solid black line represents the time series of pitch values of a neutral voice recording and the red line represents the changes in pitch. (a) The pitch is shifted upwards by 40 cents, which is typically perceived as a more positive, happy voice. (b) Inflection kicks in at the start of the utterance, with an initial shift of +140 cents, and recedes after 500 ms, resulting in more pitch variation, typically perceived as higher in arousal. (c) Vibrato, an oscillatory variation of the mean pitch with a rate of 8.5 Hz and a depth of 40 cents, resulting in more pitch variation, typically perceived as a signal of negative arousal, anxiety or fear.

1.2 The human voice also signals emotions and psychological states

Voice is a tool that not only evolved to signal physical traits such as sex, age and identity, but also a speaker's psychological states such as their attitudes or emotions. There are many ways in which emotions can be expressed by the voice. First, non-verbal vocalizations such as screams, shrieks and laughter form an important part of the emotional repertoire (Anikin et al., 2018). Second, even in speech, the voice can carry affective information aside from the linguistic content of the speech signal. The focus of this thesis will be on speech stimuli specifically rather than voice perception in broader terms (i.e. non-speech vocalizations). In this section I will first review the relevant acoustic characteristics with which emotions are conveyed in voice, and then how these emotional cues are processed in the brain.

1.2.1 Production

Voice pitch (and its main physical correlate, the fundamental frequency or F0 of the vocal harmonic signal) is an important acoustic cue in the expression of emotions (Bachorowski and Owren, 1995; Banse and Scherer, 1996). Increased pitch often correlates with highly aroused states such as happiness, while decreased pitch correlates with low valence, such as sadness (Scherer, 2003; Juslin and Laukka, 2003; Patel and Scherer, 2013) (see Figure 1.3a). The dynamic changes of pitch over the course of an utterance also can lead to different levels of pitch variation. Increased variation in pitch is associated with high emotional intensity and positive valence (Laukka et al., 2005). For instance, Pell and Kotz, 2011 reported that expressions of happiness contain higher levels of pitch variation than expressions of fear, which in return contain more pitch variation than expressions of sadness (see Figures 1.3b and c). Sadness has also been related to a slower speech rate, whereas high arousal emotions such anger and fear are often expressed with a faster speech rate (Breitenstein et al., 2001).

Formants and voice spectrum are also manipulated to express emotions vocally, albeit perhaps less flexibly and prominently than pitch. Oro-facial gestures associated with emotions, such as stretching one's mouth corners in a smile (Arias et al., 2018a) or wrinkling one's nose in disgust (Chong et al., 2018) have audible consequences in vocal formants. Non-linearities in source spectrum, resulting in roughness and breathiness for instance, can also convey vocal arousal and anger or fear (Arnal et al., 2015).

By virtue of the source/filter model, vocal pitch and formants are largely independent characteristics of vocal sounds; in the study of vocal emotions in this thesis, the focus will be mainly on voice pitch, which is manipulated using the DAVID software described below.

1.2.2 Perception

The neural processing of vocal emotions is proposed to occur in three different stages (Schirmer and Kotz, 2006) (Figure 1.4). Expressive acoustic cues such as those described in section 1.2.1 need to be analyzed and integrated by the brain to evaluate the conveyed emotion. The initial cortical processing of voices occurs in the primary auditory cortex, which has a tonotopical configuration. The different frequency characteristics of a voice are then integrated in the peripheral areas of the auditory cortex. The auditory cortex is also able to encode intensity differences and the processing of these two acoustic cues takes place before 100 ms, forming the first stage of emotional prosody processing. During the second stage, different acoustic cues are integrated at around 200 ms in the STG and STS to form an 'emotional gestalt', for which pitch dynamics seem to be preferentially treated in the right hemisphere. Finally, in the third stage, the emotional information is subjected to higher-order cognitive processes in more frontal areas from around 400 ms. As the model suggests, all processing stages can be affected by contextual significance, such as attentional demands or stimulus salience, and individual significance, such as speaker identity. It is the topic of this thesis to investigate these interactions, and more specifically how the implicit and explicit information about a speaker's identity (self or other), physical (same or different sex) and cultural characteristics (same or different language) influence how emotional cues are processed.

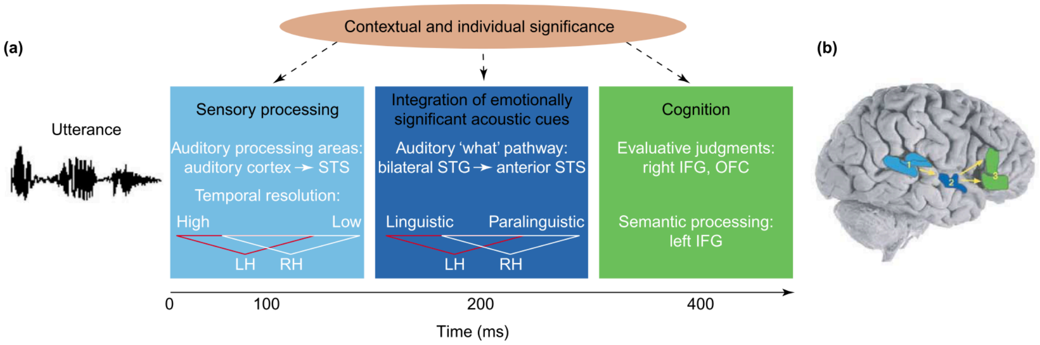

Contextual and individual significance
(a)
(b)
Utterance
Sensory processing
Auditory processing areas: auditory cortex → STS
Temporal resolution:
High
Low
LH
RH
Integration of emotionally significant acoustic cues
Auditory 'what' pathway: bilateral STG → anterior STS
Linguistic
Paralinguistic
LH
RH
Cognition
Evaluative judgments: right IFG, OFC
Semantic processing: left IFG
1
2
3
0
100
200
400
Time (ms)

Fig. 1.4: The three-stage model of emotional voice processing, adapted from Schirmer and Kotz, 2006

1.3 What is special about the self?

In this thesis, the influence of speaker identity on emotional voice processing is investigated by contrasting self-produced speech and speech produced by another person. Why would we be interested in the self and more specifically in the self-voice?

On the one hand, various lines of research have suggested that representations of self and others seem to be founded on shared neural representations. For example, the same motor areas are activated during action perception and the performance of that same action (Grèzes and Decety, 2001; Rizzolatti et al., 2001); experiencing pain, or empathy for someone else's pain, activate overlapping areas in the anterior insula and cingulate cortex (Lamm et al., 2011). On the other had, it is also necessary to be able to distinguish between self and other in order to understand another person's mental and emotional state (Steinbeis, 2016). We must be aware that another person has thoughts and feelings that differ from our own and therefore we cannot project our own mental or emotional state on someone else. For instance, in visual perspective-taking tasks where there is a conflict between two people's perspective, it is necessary to suppress automatic imitation tendencies of the other person and to recognize that the other has a different perspective in order to correctly perform the task (Santiesteban et al., 2012). The ability to distinguish between self and other and to overcome egocentricity bias has shown to be supported by the temporoparietal junction (Spengler et al., 2009) and the supramarginal gyrus (Silani et al., 2013) and is seen as the cornerstones of social cognition.

When it comes to studying emotions, the self is also of particular relevance. In outlining their higher-order theory of emotional consciousness, LeDoux and Brown (2017) underline the necessity of the self in the experience of emotion. For instance, the experience of fear is the experience that the self is threatened; there would be no feeling of fear without a representation of the self. Similarly, if an event or person affects someone else but also produces an emotion in you, it is because you are able to put yourself in someone else's place, which also requires a representation of the self (De Vignemont and Singer, 2006). *"Without the self there is no fear or love or joy"* (LeDoux and Brown, 2017, p.6).

The relevance of a sense of the self is supported by a range of studies that showed the existence of a self-bias using different experimental paradigms. This self-bias manifests itself as increased neural responses and faster behavioral responses for example. Words that are associated with the self as opposed to a stranger are better remembered (Conway and Pleydell-Pearce, 2000; Rogers et al., 1977) and meaningless stimuli such as geometric shapes that are associated with the self are more perceptually salient (Sui and Humphreys, 2013). Furthermore, in sequences of face or body representations, a visual mismatch response can be seen only to deviations of the self but not to deviations of a stranger (Sel et al., 2016). Other stimuli that are strongly associated with or relevant to the self such as someone's own name (Kotlewska and Nowicka, 2015; Tacikowski and Nowicka, 2010) or telephone ringtone (Roye et al., 2007) have also been found to evoke stronger neural responses related to redirecting of attentional resources. Differential neural responses to someone's own name versus another name have even been reported during sleep (Perrin et al., 1999) and in comatose patients (Fischer et al., 2008). Taken together, the self seems to hold a privileged position in various cognitive processes and Sui & Humphreys (2017) propose that the self is less affected by the external decision boundary and therefore serves as a stable point of reference during decision making. Additionaly, in line with LeDoux and Brown, 2017, they suggest that self-reference enhances the binding of different stimulus characteristics by increasing their salience.

The cognitive and neural processes involved in self-voice perception are especially relevant because, in social interactions, people have to process continuous changes not only in the vocal expressions of their interlocutors, but also in the feedback from their own vocal expressions. There is a long-ranging debate in the social-cognitive and meta-cognitive communities (James, 1884; Frith, 2012) about the mechanistic primacy of both types of inputs: on the one hand, the social-cognitive interpretation of other agents is believed to mobilize simulation mechanisms which supplement the processing of exteroceptive input (Gallese et al., 2004; Niedenthal, 2007). On the other hand, vocal (Aucouturier et al., 2016) and, to a lesser extent, facial feedback (Laird and Lacasse, 2014, but see also Wagenmakers et al., 2016) paradigms suggest that metacognitive evaluations of e.g. one's own emotional state are influenced by proprioceptive inputs (the sound of our voice, the motor pattern of our face) that are processed *as if* they were external stimuli.

In that perspective, recent studies on self-voice perception have not shown a clear pattern of self-bias. In an active detection task during EEG recordings, Conde et al., 2015 found an increased P3 amplitude when participants' heard their own voice compared to a stranger's voice, indexing the allocation of higher-order attentional resources. On the other hand, Graux and colleagues found that voices of strangers or familiar others evoked a larger P3a response compared to participants' own voice when using a passive auditory perception paradigm (Graux et al., 2013; Graux et al., 2015). Together, these results show that the self-voice, when task-relevant, seems to demand more attentional resources but, when task-irrelevant, seems to demand fewer attentional resources than other people's voices.

The question therefore remains whether there are fundamental mechanistic differences between e.g. hearing one's own voice suddenly change its pitch to sound brighter and happier, and processing the exact same cues on the voice of a conversation partner.

1.4 Familiarity with language

The other main case of contextual voice characteristics that will be discussed in this thesis is one's native language. Just as the intimate relationship with the self results in a high-level of familiarity with this particular vocal identity, so does one's native language.

A large body of psychological evidence shows that cultural familiarity crucially affects the way we process social signals such as face and voice characteristics. People recognize faces of their own race more accurately than faces of other races (other-race effect; Shapiro and Penrod, 1986; Meissner and Brigham, 2001), and identify speakers of their native language better than speakers of other languages (language-familiarity effect or LFE; (Perrachione et al., 2011; Fleming et al., 2014)). Within smaller social groups, the familiarity of a speaker's voice may facilitate the encoding of their spoken words in long-term memory (Pisoni, 1993), or the semantic processing of ambiguous prosodic cues (Chen et al., 2016). These effects are thought to result primarily from perceptual learning: an observer's perceptual space is modulated by exposure to informative distinctions, resulting in increased perceptual resolution for frequently-encountered items, with the result that rarely-encountered items are encoded in a less efficient manner (Valentine, 1991; Furl et al., 2002).

The effect of language familiarity on auditory processing, both at the level of phoneme discrimination and at the level of speaker identification, therefore suggests that the processing of emotional expressions may also be affected by language familiarity. However, this question so far has mainly been addressed from the point of view of cross-cultural perception (does one recognize joy or sadness differently in the English, French or Japanese language), which often confounds both how emotions are expressed and perceived (see e.g. Scherer et al., 2011, and Chapter 4). By using computer voice transformations with DAVID, it has become possible to systematically match vocal expressive cues between signals of different identities and different languages, thus allowing to examine both self-relevance and language familiarity in a common methodological and theoretical framework, namely the existence of an 'other-voice' effect in the way we process emotions in speech.

1.5 Methodological contribution

Throughout this thesis, the overarching aim is to investigate how certain characteristics of a voice, namely a speaker's identity or language, affects the processing of another vocal characteristic, emotional expression. When investigating just one dimension, such as affective information, it is usually acceptable to record tokens of different speakers in order to cancel out speaker-specific characteristics. However, when investigating how one dimension affects the perception of another, it is necessary to control one of these dimensions to properly assess this interaction.

In previous work, control over acoustical characteristics is usually achieved in a posthoc, descriptive manner, by comparing acoustic characteristics across contexts. Several signal processing toolboxes are available for researchers to analyse acoustic characteristics such as pitch, formants and intensity - one may cite PRAAT by P. Boersma for speech signals (`http://www.praat.org/`, Boersma and Weenink, 2006), MIR Toolbox by O. Lartillot for musical data (`https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox`) or Sound Analysis Pro by the team of O. Tchernichovski for non-verbal signals and bioacoustics `http://soundanalysispro.com`. However, since many acoustic properties are highly correlated, it is difficult to know which ones are responsible for the observed effects: all claims made with these tools are of a correlational, but not causal, nature.

An attractive methodological alternative is therefore to use interventional research design – i.e., to modify the sound parametrically, and observe the effect on listeners in the different contexts of interest. Unfortunately, in contrast with analysis tools such as PRAAT, only a handful of software tools is available to synthesize or transform voice/speech stimuli. PRAAT allows a small selection of manipulations of e.g. pitch, formants and intensity, but do so by resynthesis and often at the cost of stimulus realism (see e.g. Quené et al., 2012 for an example of formant resynthesis). Alternatively, the STRAIGHT Matlab toolbox (Kawahara and Matsui, 2003) can be used to morph different sounds over several voice characteristics. This technique has for example been used in the study by Latinus and colleagues (2013) described in Section 1.1.2. More recently, soundgen (Anikin, 2018) is a synthesis tool for nonverbal vocalizations and allows for parametric control over syllable length and intonation contour. While these tools have proven to be invaluable in the domain of speech research, they have not been specifically developed for the study of emotion, nor are they entirely appropriate to manipulate identical emotional cues in signals of different e.g. speakers or languages.

The work described in this thesis introduces a novel tool for the acoustic manipulation of vocal characteristics, called DAVID (short for 'Da amazing voice inflection device').

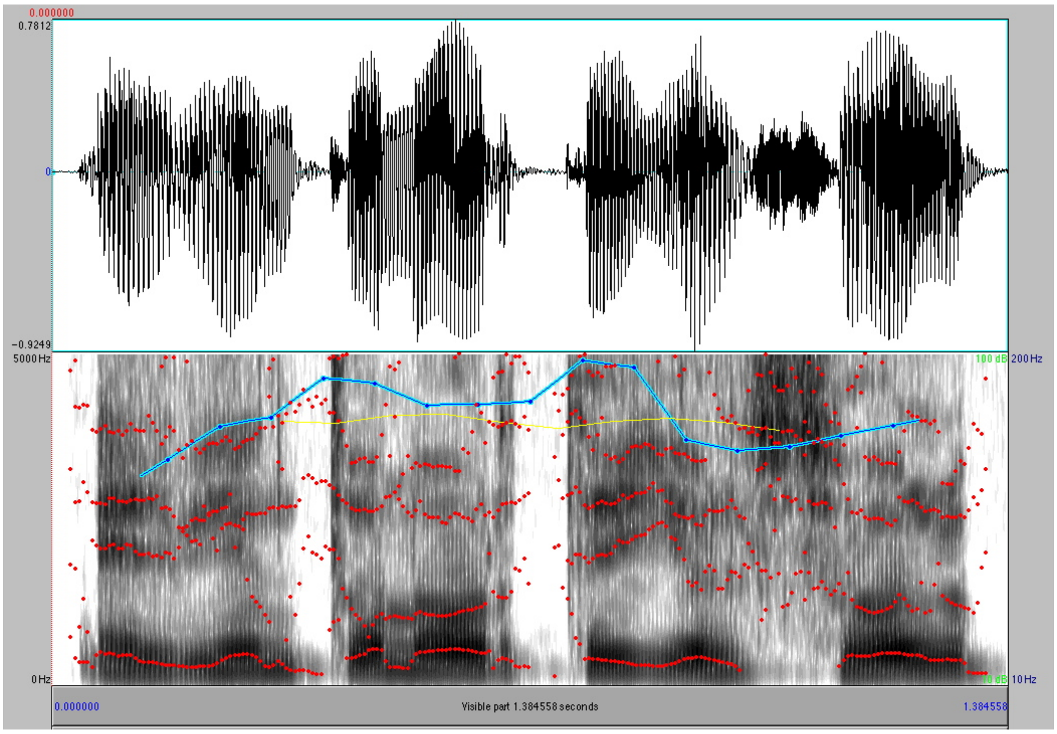
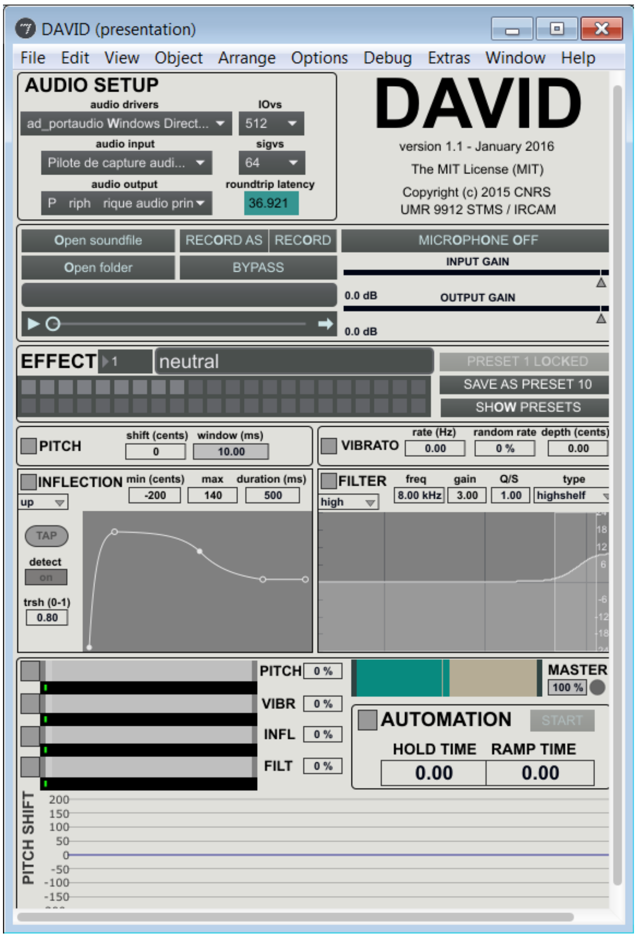


Fig. 1.5: Screenshot of the PRAAT software used to analyze the fundamental and harmonic frequencies (red) as well as amplitude profile (blue) of a voice sample. PRAAT, as other software of the same kind, offers only limited capacity to control for identical emotional cues in signals of different e.g. speakers or languages, and often at the expense of stimulus realism. Picture courtesy of J. Llisterri (UAB, Spain).

DAVID is based on technologies developed in IRCAM, which pre-existed before this work but had only been applied in mainly music production and live vocal interactions. In more details, DAVID is implemented as an open-source patch in the Max environment (Cycling74, `https://cycling74.com/`), a programming software developed for music and multimedia. It provides four types of audio effects, or building blocks, that can be combined in different configurations to create several emotions: pitch-shifting (the multiplication of the pitch of the original voice signal by a constant factor, see Figure 1.3a), inflection (a rapid modification of the pitch at the start of each utterance, which overshoots its target by several semitones but quickly decays to the normal value, see Figure 1.3b), vibrato (a periodic modulation of the pitch of the voice, occurring with a given rate and depth, see Figure 1.3c) and filtering (emphasizing or attenuating the energy contributions of certain areas of the frequency spectrum). A view of DAVID's user interface, showing the various parametric controls available for each of these transformations, is given in Figure 3.3.

The methodological contribution of this thesis is to test the applicability of DAVID for experimental research, and to the study of vocal emotional processing in particular. In the following chapters, we use DAVID to manipulate cues of happiness and sadness (Chapter 2, 3, 4), individual pitch (Chapter 3, 4) and cues of fear/anxiety (Chapter



DAVID (presentation)

AUDIO SETUP

audio drivers: ad_portaudio Windows Direct...
IOvs: 512
audio input: Pilote de capture audi...
sigvs: 64
audio output: P riph rique audio prin
roundtrip latency: 36.921

DAVID
version 1.1 - January 2016
The MIT License (MIT)
Copyright (c) 2015 CNRS
UMR 9912 STMS / IRCAM

INPUT GAIN: 0.0 dB
OUTPUT GAIN: 0.0 dB

EFFECT 1 neutral

PITCH shift (cents): 0 window (ms): 10.00
VIBRATO rate (Hz): 0.00 random rate: 0 % depth (cents): 0.00
INFLECTION min (cents): -200 max: 140 duration (ms): 500 up
FILTER freq: 8.00 kHz gain: 3.00 Q/S: 1.00 type: highshelf high
detect: on
trsh (0-1): 0.80

PITCH 0 %
VIBR 0 %
INFL 0 %
FILT 0 %
MASTER 100 %

AUTOMATION
HOLD TIME: 0.00
RAMP TIME: 0.00

Fig. 1.6: Screenshot of the DAVID software used in this work to manipulate acoustic cues of emotional expressions in vocal stimuli.

4). In Chapter 5, I will conclude on what the use of such transformation tools can bring to this research community, as well as its challenges and limitations.

1.6 Thesis outline

The aim of this thesis is to examine how the implicit and explicit processing of "other-voice" characteristics (e.g. cues that a vocal stimulus is mine or someone else's, or in my native or foreign language) influence how we process emotional cues in speech. At a *meta*, methodological level, this thesis further studies the applicability of the voice transformation tool DAVID in experimental research.

I will study two main types of other-voice characteristics. In the first two chapters, I will discuss the role of speaker identity and, in particular, the highly-familiar identity that is the self. Chapter 2 reports on a number of behavioural finding showing an implicit auditory self-advantage to categorizing emotions and recognizing speaker identities in the self-voice, which we argue plausibly results from a sensory familiarity effect. Conversely, Chapter 3 reports on behavioural and electroencephalography (EEG) studies showing an other-voice advantage for the latency of processing both emotional cues and low-level variations of pitch in the other-voice, which we argue plausibly results from the increased social relevance of processing signals from conspecifics, rather than the self.

The second part of the thesis addresses the influence of familiarity with language. In two experiments, Chapter 4 shows that French and Japanese speakers have an advantage when processing stimuli in their own native language, compared to a foreign language. This advantage is seen in the recognition of emotional expressions, but also in the detection of low-level variations of pitch. Furthermore, this advantage transfers to speakers of the same or opposite sex as the listener. Beyond their theoretical significance, these "other-voice" effects, which span the distinction between self and other, male and female, and native and foreign, have important consequences on how we interact vocally with others, within and across cultures.

The final chapter of this thesis offers an encompassing look on the common methodological thread running through this work, namely the use of emotional vocal transformations and the DAVID tool. I will discuss what the use of such transformation tools can bring to experimental research in the psychology and neurosciences of voice, as well as their challenges and limitations.

2

An implicit auditory processing advantage to the self-voice

This chapter presents the results of a series of behavioral experiments investigating the interaction between the processing of speaker identity and emotional voices. The first study aims at investigating the accuracy of emotion recognition in self- and other-voices, whereas the second study examines how emotional transformations impact speaker identity perception.

While the ability to recognize oneself by the face has been extensively studied (for a review, see Bruce, 1988), a comparatively smaller number of studies have investigated the processes involved in self-voice recognition (Nakamura et al., 2001; Rosa et al., 2008; Uttl and Morin, 2010; Graux et al., 2013; Xu et al., 2013; Graux et al., 2015; Conde et al., 2015; Pinheiro et al., 2016). Despite the common feeling that recordings of the self-voice sound 'foreign' and 'uncanny' (see Box 2.1), several behavioral studies suggest that people are able to recognize recordings of their own voice and, in some accordance with studies of the self-face (Keenan et al., 1999), that recognition is facilitated by left-hand responses, suggesting a right hemispheric lateralization (Rosa et al., 2008; Hughes and Nicholson, 2010). A more direct assessment of the brain regions involved in self-recognition was done in an fMRI study showing that hearing one's own voice leads to increased activation of the right inferior frontal gyrus (IFG) compared to hearing the voice of a friend (Kaplan et al., 2008). As this area was also activated during self-face recognition, it has been proposed that the right IFG is involved in the representation of an abstract and modality-independent concept of the self (Keenan et al., 2001).

What remains unknown, however, is how the implicit or explicit recognition of a voice being our own or someone else's impacts how we process attributes of this voice, such as its emotional expression. Limited evidence from other modalities, such as the face (Sugiura et al., 2000) or in emotional words (Pinheiro et al., 2016), have suggested that self-related stimuli may induce elevated emotional responses. In addition, information associated with the self is better remembered than when associated with someone else or with word semantics (Symons and Johnson, 1997). In more general terms of facilitating sensory processing however, behavioral evidence for a self-advantage is remarkably sparse: Knoblich and Flach (2001) let participants watch video clips of either themselves or somebody else throwing a dart at a target, and found that participants' predictions for the dart's landing position were more accurate when participants watched themselves performing the action (see Figure 2.1). Similarly, Tye-Murray et al. (2015) found that visual speech recognition (i.e. lip-reading) was facilitated when participants watched themselves talk rather than someone else. In a rare study of the self-voice, Xu et al. (2013) found that self-voice recognition was relatively preserved compared to other-voice recognition in severely degraded speech conditions (e.g. retaining only frequencies higher than the third formant), suggesting that self-voice auditory representations are perhaps especially robust. To directly test the hypothesis of facilitated auditory processing in self-voices, Study 1 examines recognition accuracy for two emotional manipulations, applied identically to self- and other- voices.

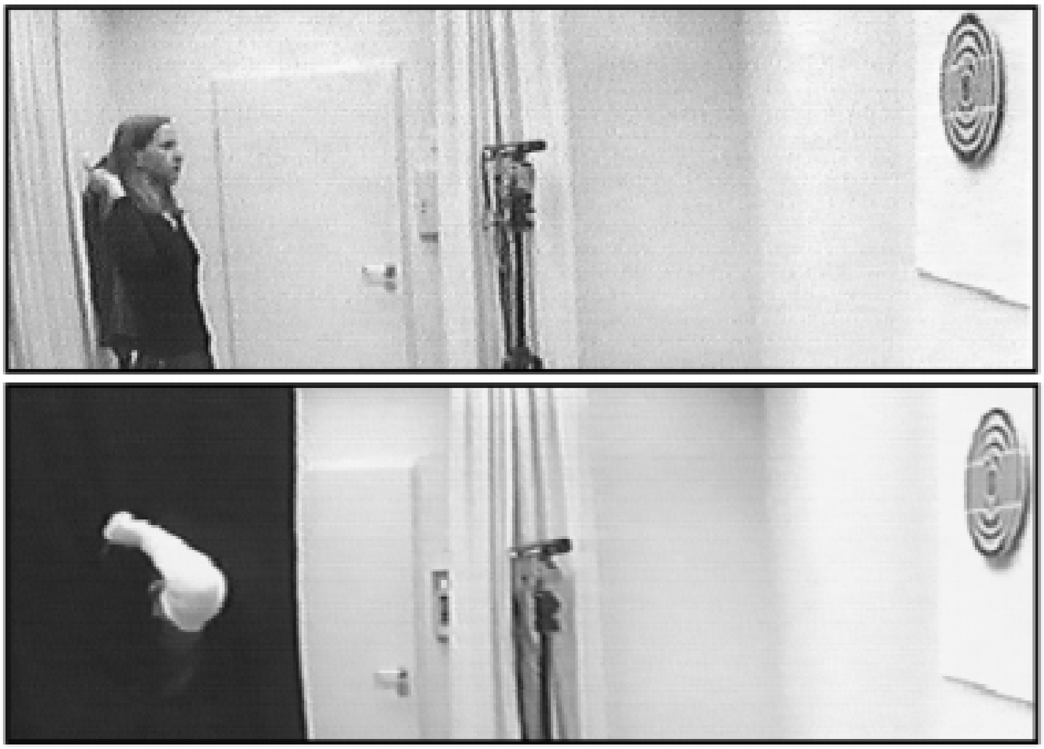


Fig. 2.1: Behavioural evidence for a sensory processing advantage to the self-body: Knoblich and Flach (2001) let participants watch video clips of either themselves or somebody else throwing darts (top), and found they better predicted the outcome of actions when they were self-performed, even when stimuli were degraded to only include the arm (bottom). Pictures adapted from Knoblich and Flach (2001).

Box 2.1: A note on why recordings of our own voice feel uncanny

Almost everyone is familiar with the experience of 'foreigness' when hearing recordings of our own voice. When we speak, we hear our own voice both through air conduction and bone conduction (Shuster and Durrant, 2003; Tonndorf, 1976) (see Figure 2.2). The signal that is transferred through air conduction, either straight from the mouth to the ear or reflected by objects in the environment, is perceived by ourselves and by others in the same way. However, the signal transmitted through our cranial bones is only perceived by the talker. Recordings of the self-voice, which only register what others hear of ourselves, are therefore often perceived differently than our real-time speech.

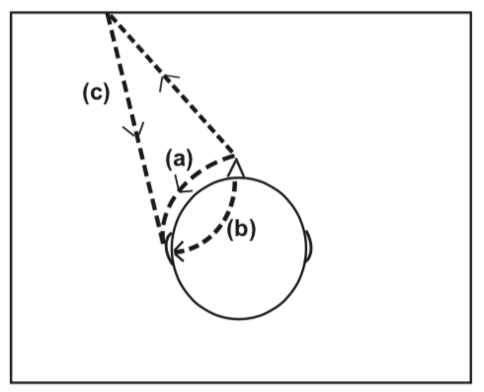


Fig. 2.2: Pathways for self-voice perception: direct (a) and indirect (c) sound transmission through air conduction, and through bone conduction (b). Adapted from Yadav et al. (2014)

Several studies have tried to correct for this difference in recordings either by shifting the pitch (Tian and Poeppel, 2014) or applying filters, such as a low-pass filter that amplified frequencies below 1000 Hz (Kaplan et al., 2008) or a band-pass filter amplifying frequencies between 300 Hz to 1200 Hz (Won et al., 2014). In a recent study, Kimura and Yotsumoto (2018) used our own DAVID tool to let participants adjust their own filter to match recordings of the self-voice and their own real-time speech, but found large individual differences suggesting that using the same strategy across participants might not improve self-recognition. At the same time, a number of studies have shown that participants typically recognize their own voice in self-recordings, perhaps since they are exposed to their own photos and voice recordings in numerous occasions in modern life, and that these are valid stimuli for investigating self-perception (Xu et al., 2013; Hughes and Nicholson, 2010). For these reasons, and in the absence of a better strategy, all self-stimuli used in this thesis were left uncorrected.

2.1 Study 1 - Emotion categorization is facilitated in the self-voice

2.1.1 Methods

Participants

Twenty female participants (mean age=21.4, SD=2.0 years) were recruited for this experiment. One participant was excluded due to technical problems during the experimental session.

Stimuli

Stimuli in Study 1 and later in Study 2 were created in the same manner. In each experimental session, participants were first asked to read a list of twenty bisyllabic neutral words and six pseudo-words (Table 2.1) with a neutral intonation. The recordings took place in a small room, using a headset microphone (DPA d:fine 4066), an external sound card (RME UCX Fireface) and the Garage-Band software (Apple Inc.) with a 44.1 kHz sampling rate and 16-bit resolution. All sounds were normalized at 70 dBA using a Matlab toolbox (Pampalk, 2004). All recorded words were transformed using DAVID software (Rachman et al., 2018) to create two variants of the neutral stimuli corresponding to happy and sad emotional expressions: happy variants of the stimuli had higher pitch and higher frequency content for positive valence, and some inflection for increased arousal; sad variants had lower pitch and lower frequency content (see Fig 2.3 for a schematic representation and Table 2.2 for parameter values). From these stimuli, five bisyllabic words and three pseudowords were used in Study 1, each word presented in the neutral, happy, and sad versions.

Procedure

Participants were presented with pairs of the same word produced by the same speaker (SV: self-voice, or OV: other-voice). The first stimulus was always an original, non-manipulated recording and the second stimulus was either a second neutral recording or a DAVID-manipulated recording in the happy or sad condition. Participants were asked to categorize the second stimulus in a 3-category emotion recognition task (neutral/happy/sad). In the second part, as control, participants performed an explicit self/other discrimination task. The stimuli in this task were

Tab. 2.1: Recorded words

French	English	Pseudowords
Cactus*	Cactus	ba-ba*
Poteau	Column	da-da
Moustache	Moustache	do-do*
Torchon	Tea towel	mo-mo
Cahier	Notebook	si-si
Compas	Compass	mi-mi*
Volcan	Volcano	
Bassine*	Tub	
Pendule*	Clockwork	
Briquet	Lighter	
Rideau	Curtain	
Ampoule	Light bulb	
Chaussette	Sock	
Chapeau	Hat	
Casquette	Cap	
Armoire*	Cupboard	
Commerce	Business	
Lunettes	Glasses	
Fourmi	Ant	
Bouteille*	Bottle	

Recorded French words, their English translations, and pseudowords. * Words used in study 1.

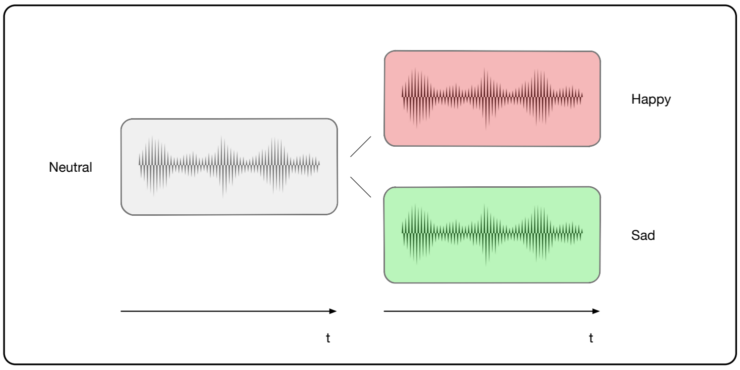


Fig. 2.3: Schematic presentation of the creation of the stimulus material for Study 1 using emotional voice transformations. Audio recordings of single words are transformed into two emotional variants (happy and sad), which alter the pitch and spectral content of the sounds in a fixed, parametric manner, while leaving their other characteristics, such as duration and content, unmodified.

the same as in the emotion categorization task, but no neutral reference sound was presented here. In each trial participants were asked to indicate whether they heard their own voice ("self") or the voice of someone else ("other"). During the experiment, participants were seated in individual booths and the stimuli were presented through closed headphones (Beyerdynamics, DT770, 250 ohm).

Tab. 2.2: Happy and sad parameter values

	Happy	Sad
Pitch		
shift, *cents*	+50	-70
Inflection		
duration, *ms*	500	–
min., *cents*	-200	–
max., *cents*	+140	–
Shelf filter		
cut-off, *Hz*	>8000	<8000
slope, *dB/octave*	+9.5	-12

Parameter values of the happy and sad transformations used in the experiments (refer to Rachman et al., 2018 for details).

Analysis

All statistical tests were two-tailed and used an alpha-level set at .05. Where appropriate, Greenhouse-Geisser corrections for non-sphericity were applied. Uncorrected degrees or freedom and corrected *p*-values are reported. Significant effects were followed up by paired *t*-tests using Bonferroni corrections for multiple comparisons.

2.1.2 Results

Emotion recognition

T-tests for each conditions showed that recognition accuracy was above chance level (33.3%) for all six conditions (*ps*$< .01$, Bonferroni-corrected, Figure 2.4, Left). A 2×3 (speaker type × emotion) rmANOVA showed a main effect of speaker [$F(1,18) = 14.9$, $p < .01$], with greater accuracy for the self-voice than the other-voice, a main effect of emotion [$F(2,36) = 20.5$, $p < .001$] and a significant speaker × emotion effect [$F(2,36) = 5.2$, $p < .05$]. Post-hoc *t*-tests between self and other for each of the three emotional conditions showed that, while there was no difference between accuracy on self and other voice in the neutral and sad conditions, the happy transformation was better categorized when applied on the self-voice than on a stranger's voice ($p < .01$).

Self/other discrimination

T-tests for each condition showed that discrimination accuracy was above chance level (50%) for all three OV conditions and the neutral and sad SV conditions

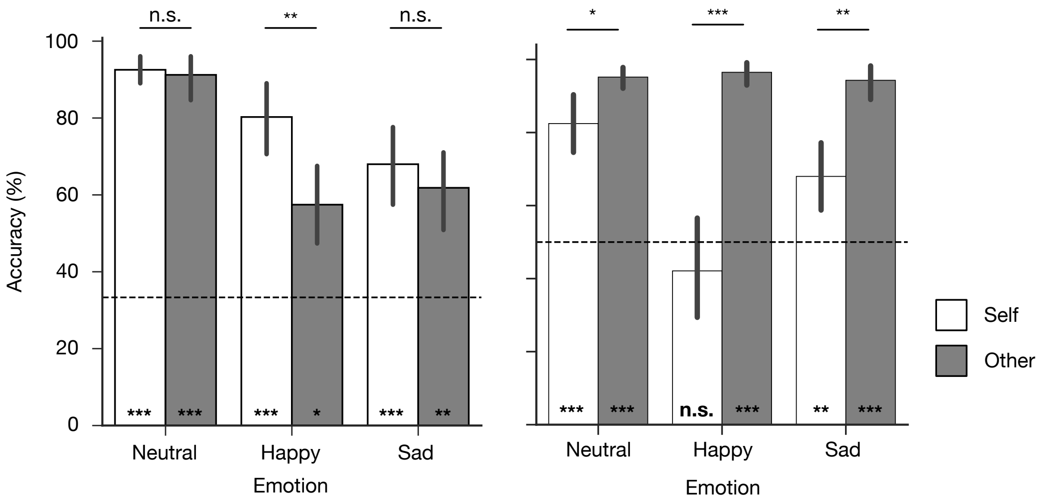


Fig. 2.4: Results of Study 1. **Left:** Emotion categorization in self and other voice. Happy (and to a lesser extent, sad) manipulations were better recognized when participants heard them on their own voice, rather than that of an unfamiliar stranger. **Right:** Self/other discrimination in the same three conditions. SV-advantage in recognizing happy emotions does not translate to an explicit advantage in recognizing these conditions as self-produced. Dotted line indicates chance level performance. Error bars represent 95% confidence intervals, $*p < .05$, $**p < .01$, $***p < .001$

($ps < .01$, Bonferroni-corrected). However, in contrast with the improved accuracy to recognizing happy expressions in SV stimuli, discrimination of the happy SV condition did not differ from chance level. A 2×3 (speaker type × emotion) rmANOVA showed a main effect of speaker [$F(1, 18) = 69.2$, $p < .001$] and a main effect of emotion [$F(2, 36) = 17.5$, $p < .001$]. There was a significant speaker × emotion effect [$F(2, 36) = 13.1$, $p < .001$]. Post-hoc t-tests between self and other for each of the three emotional conditions showed that discrimination accuracies were lower for the self voice compared to the other voice in all three emotional conditions ($ps < .05$) (see Figure 2.4, Right). Furthermore, for the self-voice, accuracies in the happy condition were lower than in the neutral and sad conditions and accuracies in the sad condition was also lower than in the neutral condition. The accuracies did not differ between any of the three emotional conditions for the OV.

2.1.3 Discussion

The results of this behavioral study are twofold. On the one hand, while manipulated emotions were recognized well above chance level for both self and other voices, happy (and to a lesser extent, sad) manipulations were better recognized when participants heard them on their own voice, rather than on unfamliar voices. To the best of our knowledge, a processing advantage in recognizing emotional expressions on the self-voice had never been reported in the literature and would only be manifest using voice transformations: first, if we had participants produce their own emotional expressions, memory of the self-produced stimuli may confound emotion recognition; second, if both self and other-voice emotional stimuli were produced by participants, there would be no way to control that these used the same acoustic cues, or that the production of these cues were equally salient.

The finding that participants more accurately recognized identical acoustic cues of emotional expressions in their own, rather than someone else's, voice is in line with a number of studies showing better recognition or prediction accuracy when one observes one's own actions than when one observes another person's actions (Knoblich and Flach, 2001; Tye-Murray et al., 2015), and is also consistent with the relative preservation of self-voice recognition in degraded audio conditions (Xu et al., 2013). One possible explanation to the self-voice processing advantage is that observers have greater auditory familiarity with their own voice compared to other people's voices. More robust, finely-grained auditory representations for the pitch and timbre of the self-voice would emphasize the small changes expressed in emotional variants, and help their recognition. This effect would be consistent with results documenting facilitating effects of language- or speaker-familiarity on phonological and semantic processing (Chen et al., 2016; Fleming et al., 2014). Another possible explanation for the self-advantage would be that auditory processing in the self-voice is supported by stronger associations with other representations of the self, such as the motor/articulatory representation (Hickok and Poeppel, 2007). Even if the emotional expressions tested here were computer-generated, an increased ability to simulate the phonological dynamics of the self-voice would likely help recognizing the nature of expressive variants. Related results would include for instance how inhibiting motor simulation with TMS degrades the discrimination of human versus machine sounds in singing voice stimuli (Lévêque et al., 2013), or how participants are able to discriminate the sound of their own hand-clapping from that of someone else (Repp, 1987).

On the other hand, the results of the explicit self-other discrimination task on the same stimuli show that improved self-voice emotion recognition is not mediated by explicit recognition of the stimuli as self-generated: while both neutral and sad versions of the self-voice are categorized correctly above chance level, the happy transformation on the self-voice reduces accuracy to chance-level performance. In other words, the emotions in self-voice stimuli were better recognized than in other voices, even though they were often mistaken for other-voice. This pattern of results is in line with previous findings in the visual modality, showing that explicit knowledge about the nature of stimuli is not a necessary condition to obtain a self-advantage: while participants in Knoblich and Flach's dart-throwing experiment were informed whether they were seeing themselves or another person, similar results were obtained when estimating trajectories of writing strokes even in the absence of such knowledge (Knoblich and Flach, 2001).

This discrepancy between explicit and implicit processing may be especially salient in voice processing. Candini and colleagues (2014) directly compared implicit and explicit self-voice perception in pairs of voices. The implicit task consisted of making same/different identity judgment for each pair of stimuli, for which the identity itself was not categorized. In contrast, in the explicit task, participants were asked to indicate whether or not they heard their own voice among each pair of stimuli. Their results showed that, while explicit self-recognition was less accurate than the identification of non-self voices, participants performed equally well in the same/different task for recordings of their own voice and recordings of another speaker. Similarly, in an effort to study the influence of speaker familiarity on speech intelligibility, Holmes and colleages (2018) also compared performances in an implicit and an explicit task. Their results showed that speaker familiarity, when task-irrelevant, seems to provide an advantage in speech comprehension during competing speech perception even when the familiar voices were not well recognized in an explicit task.

On the whole, it thus appears that learned speaker identities provide a speech processing or comprehension advantage, but that such auditory processing is facilitated only by implicit assessment of self-voices and does not necessarily "graduate" to explicit awareness (Kreitewolf et al., 2017). This interpretation would predict that, while impaired in explicit measures, self-recognition in emotional stimuli would be preserved or even enhanced compared to other-voice recognition if it is tested implicitly. This test will be the topic of Study 2 below.

On a final methodological note, results from Study 1 provide the incidental observation that, while typically recognized as emotional (Figure 2.4 Left), emotional voice transformations with DAVID do affect speaker identity, and in particular degrade the recognition of self-produced voice (Figure 2.4 Right). Preserving speaker identity, as well as ensuring that stimuli sound authentic and not overly artificial, is an important requirement for such transformations, and that aspect will be further discussed in Chapter 5. Suffice to observe here that the current results do not allow to distinguish the possibility that the transformations affect speaker identity in the self-voice specifically, or speaker identity in general, i.e. both in familiar and unfamiliar voices. In particular, the current task is necessarily asymmetric as it appears more likely to err by confusing the (unique) self-voice with one of a multiplicity of possible other-voices, than to mistakenly perceive an arbitrary unfamiliar voice as being the self. The implicit task of Study 2 will provide more data on that property of the transformations.

2.2 Study 2 - Implicit versus explicit identity perception

Because the results of Study 1 implied that distinct implicit and explicit processes are involved in self-voice perception, and how it may facilitate emotion recognition, we performed another behavioral study comparing these two processes more directly. To this end, we used an adapted version of the approach taken by Candini and colleagues (2014). Additionally, this experimental design also allowed us to address the question of whether the emotional voice transformation affect the perception of speaker identity in similar ways when it involves the self and unfamiliar speakers.

Where Study 1 involved judgments of emotions with an implicit knowledge of speaker identity, the current task involves judgments of speaker identity with an implicit knowledge of emotional expression. These judgments of speaker identity are collected in two tasks, involving either explicit judgments about the self-identity ("is one of these stimuli my own voice?") or implicit judgments ("are these two stimuli the same speaker?", only incidentally the self), allowing to compare performance in these two conditions.

2.2.1 Methods

Participants

Twenty right-handed participants (11 female, mean age=22.1, SD=2.4 years) were recruited for this experiment. Twelve of these participants had already participated in Experiment 1 and recordings of their voice were thus already available. The eight newly recruited participants came to the lab for a voice recording one week prior to the test session.

Stimuli

The stimulus material was created as described in Section 2.1.1. For each participant, two speaker identities of the same gender were randomly selected from the database recorded in a previous study to serve as the two "other" voices. Contrary to the previous study, all twenty bisyllabic words were included in the experiments described here.

Procedure

Each experimental session consisted of an implicit and an explicit task. In each trial, participants listened to two recorded voice samples after which they were asked to indicate whether the two words were produced by the same or different speakers in the implicit task, or whether at least one of the voices was their own voice in the explicit task. All pairs of stimuli consisted of two different words and we created twelve stimulus pairs for each of the experimental conditions by randomly selecting two words out of the twenty recordings (see Table 2.3 for all speaker combinations used in both tasks). In addition, to test how the emotional transformations affects performance, three emotional conditions were created: neutral, happy, and sad. In the neutral condition, both stimuli in each pair were the original voice recordings. In the emotional conditions (happy and sad), one stimulus in each pair was transformed as described in Study 1 while the other stimulus was presented in its original form. In self-other pairs, the transformation was always applied on the self-voice and the recording of the unknown speaker was always presented in its neutral form. The order of speakers and emotional transformations in each pair was counterbalanced for each condition throughout the whole experiment.

Tab. 2.3: Speaker combinations per condition

	Same speaker	**Different speakers**
Self	self - self	self - other A *or* self - other B
Other	other A - other A *or* other B - other B	other A - other B

All possible speaker pairs for each experimental condition. Note that the order of speakers in each pair was counterbalanced throughout the experiment.

2.2.2 Results

The conditions of pairs of the same speaker identity and pairs of different speaker identities were analyzed separately because the sources of variance differed between them; one speaker versus two speakers respectively. Therefore, two 2×3 (speaker×emotion) repeated measures ANOVAs were conducted for the "same speaker" and "different speaker" conditions separately.

Overall effect of task

To test the overall difference between implicit and explicit processing self-other disctinction, a 2×2 (task × speaker identity) rmANOVA was carried out to analyze differences in accuracy. Results confirmed a self-disadvantage to speaker recognition in the explicit task (as seen in Study 1) as well as, to some extent, a dissociation between implicit and explicit self-recognition : there was a main effect of task [$F(1,19) = 101.5$, $p < .001$] and a main effect of speaker identity [$F(1,19) = 16.9$, $p < .001$]. There was also an interaction effect of task and speaker identity [$F(1,19) = 59.1$, $p < .001$]. Post-hoc *t*-tests showed that accuracy for unknown voices, but not for the self-voice, differed between tasks. Furthermore, accuracy for self-voice and other-voice did not differ in the implicit task, but in the explicit task, accuracy was higher for other than for self (see Figure 2.5).

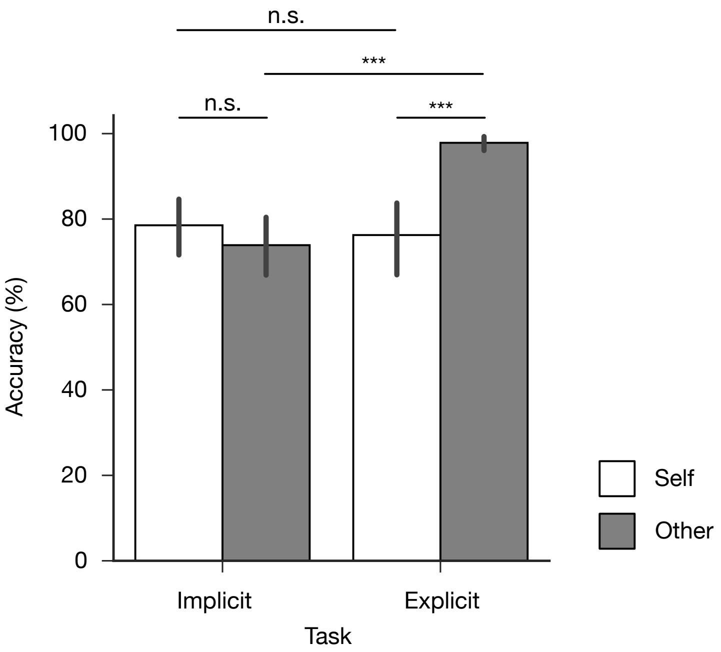


Fig. 2.5: Differences between the implicit and explicit task depending on speaker identity. Error bars represent 95% confidence intervals, $^{***}p < .001$

Implicit task

Main result: pairs of the same speaker Correct results in same-speaker pairs correspond to hits (i.e. correctly responding that these were voices of the same speaker). Results in implicit judgments of these pairs confirmed the prediction made in Study 1 that implicit self-recognition was in fact preserved, and even enhanced, in the emotional stimuli for which a self-advantage of emotional recognition was seen. There was a main effect of emotion [$F(2,38) = 479.3$, $p < .001$] and an emotion×speaker interaction [$F(2,38) = 3.4$, $p < .05$]. Post-hoc *t*-tests showed that, while accuracy for self and other did not differ in the neutral and sad conditions, accuracy was higher for pairs of the self than for pairs of unknown speaker in the happy condition (*p*$< .05$, see Figure 2.6, Left).

Results in Study 2 differed from the pattern seen in Study 1 regarding the sad manipulation: while the happy effect seemed to most hurt self recognition in Study 1, it was here the sad transformation which was almost systematically associated to changes to perceived identities. We conducted two separate one-way rmANOVAs for each speaker identity with Emotion as within-factor. There was a main effect of emotion in the self condition [$F(2,38) = 228.0$, *p*$< .001$], and follow-up paired comparisons indicated that, while the performance in the identity matching task did not differ between the neutral and happy conditions (*p*$> .05$), speakers were less often judged as the same identity in the sad condition when compared to the happy and neutral conditions (*ps*$< .001$). Similarly, the one-way rmANOVA for the other condition indicated an effect of emotion [$F(2,38) = 202.5$, *p*$< .001$], for which paired comparisons showed that the accuracy in the sad condition was lower than in the happy and neutral conditions and that the accuracy in the happy condition was also lower than in the neutral condition (all *ps*$< .01$).

Secondary result: pairs of different speakers Correct results in different-speaker pairs correspond to correct rejections (i.e. correctly responding that these were not voices of the same speaker). Implicit judgements in these pairs were only of secondary importance, because of the strong bias to confound the self with others more than others with the self. Accordingly, there were no main effects of emotion or speaker, nor was there an interaction between these two variables. Performance thus did not differ between pairs of two unknown speakers and pairs that included the self-voice (Figure 2.6, Right).

Explicit task

Secondary result: Pairs of the same speaker Correct responses in these pairs correspond to hits in the self condition (i.e. correctly detecting that one of two self-stimuli was the self) and to correct rejections in the other condition (i.e. correctly detecting that none of two other-stimuli were the self). Explicit judgements in such pairs were only of secondary importance, because they were based on twice the amount of stimulus evidence as different-speaker pairs, and could be based in each pair as much on the initial neutral stimulus as on the second emotional stimulus. There was a main effect of speaker [$F(1,19) = 4.8$, $p < .05$], a main effect of emotion [$F(2,38) = 8.4$, $p < .01$]. The interaction between speaker and emotion was marginally significant [$F(2,38) = 3.3$, $p = .068$].

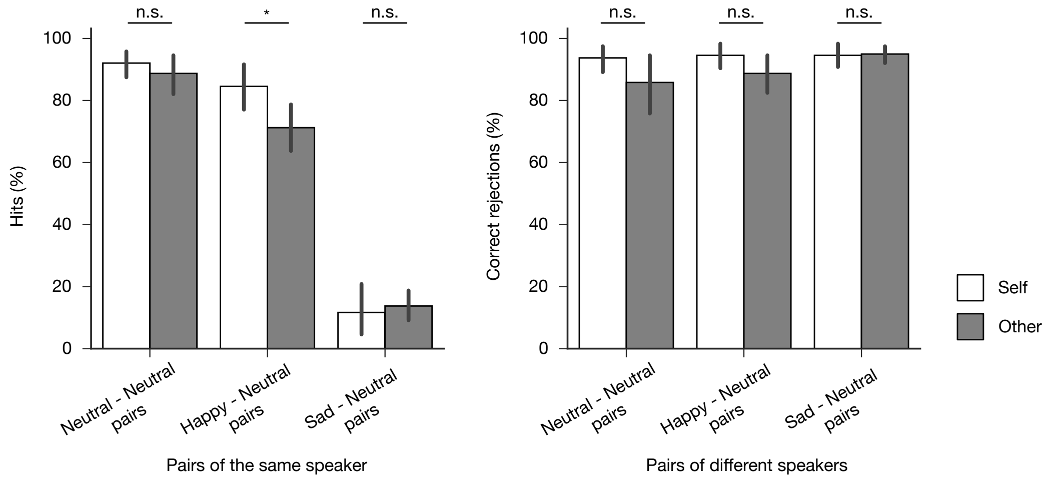


Fig. 2.6: Response accuracies in the implicit task. Left: pairs of the same speaker identity (self-self or other-other) in the neutral (neutral-neutral), happy (neutral-happy), and sad (neutral-sad) conditions. For the pairs in which one voice had been treated with the happy tranformation, speakers were more often judged to be the same person for the self-voice than for an unknown voice. Right: pairs of different speakers (self-other or two different other speakers). Error bars represent 95% confidence intervals, $*p< .05$.

Main result: Pairs of different speakers Correct responses in these pairs correspond to hits in the self condition (i.e. correctly detecting that the first stimulus of a self-other pair was the self) and to correct rejections in the other condition (i.e. correctly rejecting that neither stimulus of an other-other pair were the self). Explicit judgments in such pairs confirmed the results of Study 1, with a degradation of explicit self-recognition in the two emotional conditions. There was a main effect of speaker [$F(1,19) = 84.7$, $p < .001$], a main effect of emotion [$F(2,38) = 133.1$, $p < .001$], and a significant interaction between speaker and emotion [$F(2,38) = 83.4$, $p < .001$]. As in implicit judgments of same-speaker pairs above, it was this time the sad transformation that mostly hurt self-recognition, to the point that sad-transformed self-voices were systematically rejected as non-self.

2.2.3 Discussion

Results of Study 2 reinforce those of Study 1 on several important aspects. First, they confirm the intuition that, while impaired in explicit measures, self-recognition in emotional stimuli (and especially here happy stimuli) is preserved and even enhanced compared to other-voice recognition if it is tested implicitly. This provides support to the notion of a self-specific auditory advantage to recognizing emotions in the self-voice.

Second, results of Study 2 confirm a general dissociation between implicit and explicit processing of the self-voice: when doing implicit judgments of speaker identity,

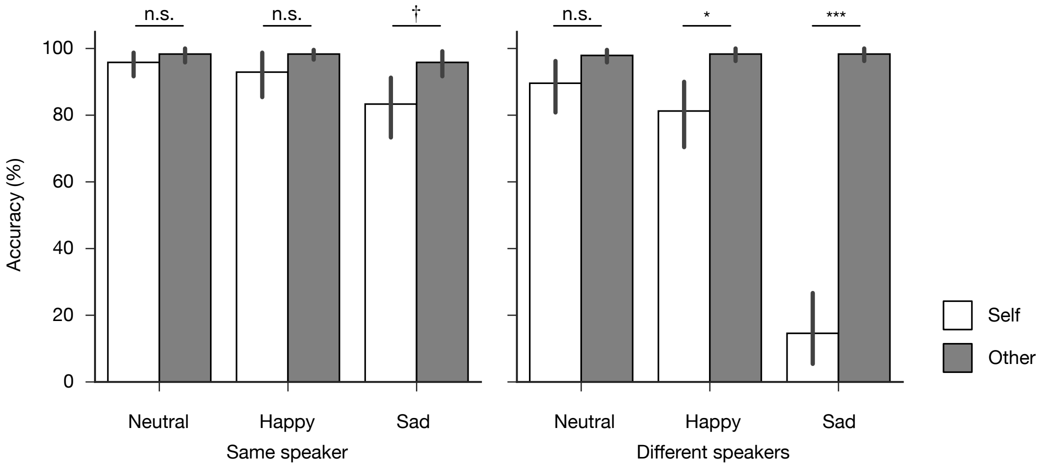


Fig. 2.7: Response accuracies in the explicit task. **Left**: pairs of the same speaker identity (self-self or other-other) in the neutral (neutral-neutral), happy (neutral-happy), and sad (neutral-sad) conditions. **Right**: pairs of different speakers (self-other or two different other speakers). Error bars represent 95% confidence intervals, $^{\dagger}p = .063$, $^{*}p < .05$, $^{***}p < .001$.

participants were more accurate to recognize two instances of their own voices, than two instances of someone else's voice (Figure 2.6, Left); on the contrary, when doing explicit judgements, participants were less accurate to recognize their own voice than to recognize two instances of someone else's (Figure 2.7, Right). These results are in line with those of Candini et al. (2014), which showed better self-voice recognition in implicit than in explicit tasks. They are also reminiscent of a number of other known dissociation between the implicit and explicit recognition of self-body parts (in which better performance for one's own body parts, in comparison with others', is not observed whenever participants are asked to explicitly recognize their own body - Frassinetti et al., 2011) or in blindsight (in which patients do not consciously perceive visual objects, but are still able to react to their emotional characteristics - Tamietto and De Gelder, 2010). All in all, they suggest that self-recognition is supported by richer auditory representations but that these representations contribute differently to our sense of self when implicitly or explicitly accessed.

Finally, results of both studies confirm that DAVID-generated vocal transformations are not transparent with respect to vocal identity. In Study 1, explicit self-recognition was mostly degraded by happy transformations. In Study 2, using the same stimuli but a different sample of participants and a different explicit task, it was the sad transformation which was almost systematically associated with mistaking the self for others. Importantly, the fact that this effect was seen in the implicit task for both self-self and other-other pairs (Figure 2.6, Left) indicates that the transformations do not only affect self-voice identity, but speaker identity in general, i.e. both in familiar and unfamiliar voices. Methodological implications will be discussed in Chapter 5.

2.3 Conclusion

Study 1 establishes a processing advantage in recognizing emotional expressions in one's own voice, compared to the voice of someone else. This processing advantage is not mediated by the explicit recognition of such voice as self-generated (Study 1 and 2), but co-exists with a self-advantage in the implicit recognition of speaker identity (Study 2). These results suggest that the auditory processing of emotions and speaker identity is supported by richer auditory representations for the self-voice, but that these representations contribute differently to our sense of self when implicitly or explicitly accessed.

This auditory processing advantage to the self-voice may be explained by a greater auditory familiarity with the sound of one's own voice and/or by stronger associations with other representations of the self, such as the motor/articulatory representations. The former explanation would predict that a similar effect of auditory familiarity can be observed to a lesser-extent for highly familiar other-voices, such as voices of the same sex or the same language - this will be the topic of Chapter 4. The latter explanation would predict a different involvement of the motor/premotor system in processing self- and other-voice stimuli. In the next chapter, we will seek evidence that this is indeed the case by turning to electrophysiological data and conducting an electroencephalography (EEG) study of the implicit processing of emotional cues in self and other-voice sequences.

3 A priority to change detection in the other-voice

The results of the previous chapter suggest that the self-voice perception benefits from an auditory processing advantage. This advantage may be driven by better-defined auditory and/or motor-articulatory representations in listeners, allowing for a finer processing of auditory cues (better emotion recognition: Study 1) while at the same time providing a more stable implicit representation of speaker identity allowing for better generalization of different utterances (better identity matching: Study 2). In order to get a better understanding of the neural processes underlying these effects, this chapter presents the results of an electroencephalography (EEG) study (Study 3) using an oddball paradigm and a behavioral study (Study 4) in which acoustic changes are applied on running speech.

The experimental paradigms in this chapter bring several advantages with respect to Chapter 2. First, the use of EEG methods allows us to investigate the neural mechanisms involved in emotional voice perception and to see how the behavioral findings of the previous chapter map onto these processes. EEG is a noninvasive method to record neural activity from the scalp. The signal that is recorded from multiple electrodes stems from the synchronous activity of large groups of neurons and can be measured at a millisecond temporal resolution. The use of EEG therefore allows us to precisely investigate the temporal dynamics of the processing of emotional expressions on different speaker identities.

Second, both experimental paradigms used in this chapter approach real-life situations of voice perception more closely. Compared to the isolated words of Chapter 2, the stimuli used in Study 3 are presented in long sequences, locally interspersed with acoustic changes, which is a step closer to mimicking realistic prosodic changes as they occur in daily social encounters. Similarly, Study 4 uses continuous pitch changes during recordings of complete sentences, as opposed to the bisyllabic words used in Chapter 2.

Third, the EEG paradigm used here is designed to elicit a mismatch negativity (MMN) response (Box 3.1). The MMN, which is measured pre-attentively and can be elicited even in the absence of explicit detection of deviants (Alho and Sinervo, 1997; Neuloh and Curio, 2004), is a known marker of implicit processes (van Zuijen et al., 2006). This paradigm is therefore particularly well-suited to investigate in more details the processes identified in the previous chapter, which we have found are not necessarily well-represented in explicit tasks.

Box 3.1: A note on mismatch negativity

The mismatch negativity (MMN) is a neural change detection response that is elicited by regularity violations in sequences of sensory stimuli, even without paying attention to them, and can be measured with EEG (Näätänen et al., 1978; Näätänen et al., 2007). Initial mismatch negativity studies involved the presentation of a series of pure tones at two different frequencies, for instance 1000 Hz and 1050 Hz. Regularity in an oddball paradigm is created by presenting one of the tones ("standard") often, while the other tone ("deviant") is presented rarely, thus violating the rule that is established by the standard stimulus. The neural signature of this change detection process is a negative deflection of the difference wave, measured by subtracting the averaged EEG response to standard stimuli from the averaged EEG response to the deviants.

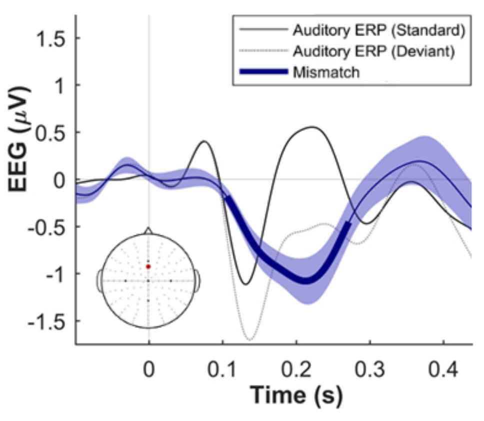


Fig. 3.1: The mismatch negativity is a frontal negative wave detected by subtracting the EEG response to standards from the response to deviants in a sequence of stimuli. Adapted from Iyer et al., 2017

The MMN is most clearly visible on frontocentral electrodes (Sams et al., 1985) and its amplitude is modulated by the extent to which the standard and deviant stimuli differ from each other and the lower predictability of the occurance of the deviant stimuli (Näätänen, 1990). The neural generators of the MMN are located in two different regions, the first one in bilateral supratemporal cortices (e.g. Giard et al., 1990; for a review see Näätänen et al., 2007) and the second one in the frontal cortex, predominantly right-lateralized (Escera et al., 1998; Giard et al., 1990).

A handful of studies have used MMN/oddball paradigms to study self-voice processing (Conde et al., 2015; Conde et al., 2016; Graux et al., 2013; Graux et al., 2015), with mixed conclusions. On the one hand, Conde and colleagues report that participants' own voice evoked larger N2 and P3 when compared to a stranger's voice in an active detection task (Conde et al., 2015). On the other hand, Graux and colleagues found that participants' own voice evoked a smaller P3a amplitude than the voice of a stranger or a familiar other when using passive oddball paradigms (Graux et al., 2013; Graux et al., 2015). Critically, these previous studies have used designs in which self- and other-stimuli are alternated. While such a contrast sheds light on the relative saliency of self-voice deviants in a context of other-voices standards, it does not address this thesis's question of whether we process expressive changes in our own voice (i.e. self-voice deviants in a sequence of self-voice standards) in the same way as in the voice of others (i.e. other-voice deviants in a sequence of other-voice standards). To this aim, Study 3 below will compare MMN responses to emotional deviants in two types of sequences composed either entirely of self-voice or of other-voice sounds.

Following results from Chapter 2, one can predict that processing of expressive changes in a sequence of self-voices may be facilitated compared to sequences of other-voices, because the self-voice is supported by richer auditory or motor/articulatory representations. Visual paradigms have consistently shown that deviants among familiar letters or shapes elicit faster mismatch responses (e.g. Sulykos et al., 2015), and similar results were found contrasting deviants in culturally familiar sounds (e.g. the Microsoft Windows chime) with deviants in sequences of the same sounds played backwards (Jacobsen et al., 2005). Specific to the current study, richer representations can be predicted to emphasize the contrast between standards and deviants (a bit like familiarity with a language increases the perceived difference between speakers of this language - Fleming et al., 2014), which would result in an increased MMN amplitude in sequences of self-voices.

An alternative possibility, however, is that the processing of expressive cues in a sequence of other-voices may be facilitated because the other-voice is less internally predictable, more socially relevant, and thus warrants more/faster reorientation of attention than self-stimuli. There are known effects of social relevance on mismatch responses in the visual and auditory modalities, notably when manipulating the communicative nature of the signals: in sequences of emotional face stimuli, Campanella and colleagues (2002) found earlier and larger mismatch responses to changes of expressions that led to a different emotional appraisal (e.g. a happy face in a sequence of sad faces) than to a different depiction of the same emotion (see also Bayer et al., 2017; Kovarski et al., 2017). In the auditory domain, affiliative signals such as laughter evoke larger MMN than a non-affiliative growl (e.g. Pinheiro et al., 2017), vowels expressing fear evoke both an earlier and larger MMN response than expressions of happiness and sadness (Carminati et al., 2018) and changes of the same intensity elicit larger MMNs on vocal than nonvocal stimuli (Schirmer et al., 2007), all of which can be interpreted as an effect of social relevance.

3.1 Study 3 - Earlier onset of MMN responses in sequences of other-voices

3.1.1 Methods

Participants

Twenty-five healthy, right-handed female participants took part in this study (27 came in for voice recordings, but two were not able to do the EEG session), two of which were excluded from analysis due to excessive EEG artifacts in the EEG, leaving 23 participants in the final analysis (mean age=21.2, SD=1.8 years).

For this experiment, we selected only female participants because the voice transformations we used worked more reliably for female than for deep, lower-pitch male voices (Rachman et al., 2018). The experimental protocol was approved by INSEAD's institutional review board and all participants gave written informed consent before the start of the study. Participants reported normal or corrected to normal vision, normal hearing, and an absence of neurological or psychiatric illness. They were financially compensated for their participation.

Stimuli

Participants came to the lab one week prior to the EEG experiment for a voice recording session. The recordings took place in a sound-attenuated booth, using a headset microphone (DPA d:fine 4066), an external sound card (RME UCX Fireface), and Garage-Band software (Apple Inc.) with a 44.1 kHz sampling rate and 16-bit resolution. Participants were asked to read a list of twenty disyllabic neutral words and six disyllabic pseudo-words with a neutral intonation, these were the same words as in Chapter 2 (Table 2.1). All sounds were normalized at 70 dBA using a Matlab toolbox (Pampalk, 2004). Because only the recordings of the pseudoword */ba-ba/* were used during the EEG session, these sound files were also normalized in time to have a duration of 550 ms using superVP/audiosculpt software. To ensure comparable amounts of vocal diversity in 'self' and 'other' stimuli, participants were grouped in pairs such that the 'self' voice (SV) of one participant served as the 'other' voice (OV) for the other participant, and vice-versa.

Finally, we processed all recordings with the DAVID software platform (Rachman et al., 2018) to generate happy and sad deviants from the standard utterance, by combining audio effects such as pitch shift (increasing the standard's pitch by 50 cents in the happy deviant, and decreasing by 70 cents in the sad deviant), inflection (increasing the beginning of the second syllable by an extra 70 cents in happy) and filtering (increasing high-frequency energy with a high-shelf filter in happy, and decreasing high-frequency energy with a low-shelf filter in sad; see Table 3.1 for parameter values).

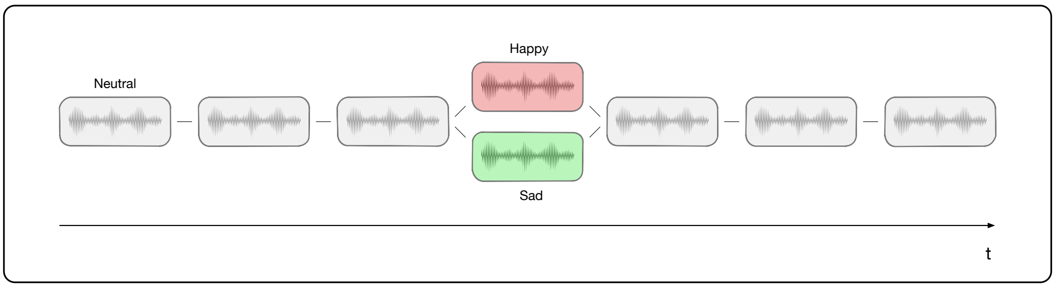


Fig. 3.2: Schematic representation of the application of DAVID in the mismatch negativity study. Here, a neutral utterance is transformed to create two other emotional stimuli that can be used in controlled oddball sequences in an MMN study.

One critical requirement to comparing mismatch responses to expressive deviants in self and other-voice sequences, is the need to control for similar changes to occur in both contexts. When relying on participant voices, it is always possible that one speaker expresses a given emotional change more clearly or loudly than another speaker (Jürgens et al., 2015), or with different cues (e.g. louder vs higher pitch), such that any difference observed in processing such changes cannot be unambiguously attributed to self/other processing differences, rather than individual production differences. By using programmable emotional transformations, we ensure that, in both the self and other sequences, deviants differed from the standards on *exactly* the same cues in *exactly* the same manner (e.g. a 50-cent pitch increase on the second syllable of the word), making EEG responses to such deviations comparable between sequences (Figure 3.3).

Oddball paradigm

We used an oddball paradigm with two different sequences: one 'self sequence' and one 'other sequence'. In the 'self sequence', the neutral recording of the SV served as the standard stimulus and the 'happy' and 'sad' transformation of the standard stimulus served as the two emotional deviants. Following the same logic, the 'other-sequence' used the neutral recording of the OV as the standard and its 'happy' and 'sad' transformations as deviants (see Figure 3.4). Additionally, both sequences also contained an identity deviant to try to replicate previous studies by Graux and

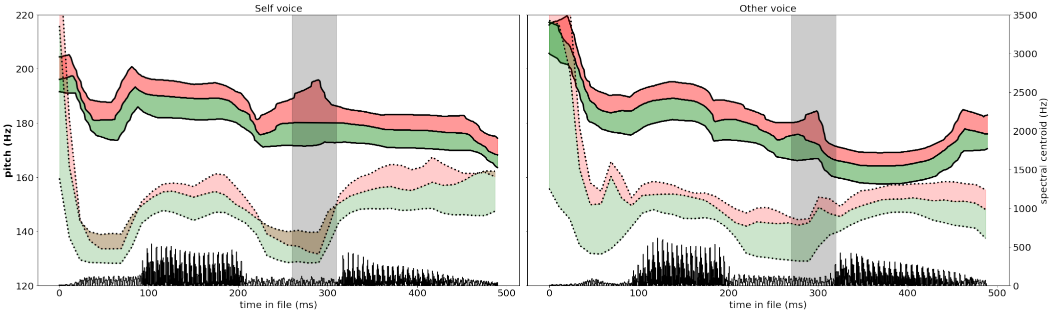


Fig. 3.3: Acoustic content of two representative stimuli used in Study 3. Solid line, black: pitch of the standard; red: increase of pitch in the happy deviant; green: decrease of pitch in the sad deviant. Shaded area indicates second-syllable inflection in the happy deviant. Dotted line, black: spectral centroid (centre of mass) of the standard; red: high-frequency energy added in the happy deviant; green: high-frequency energy removed in the sad deviant. Bottom, black: half-corrected waveforms of the standard. Left: participant's own voice (SELF). Right: another participant's voice (OTHER).

Tab. 3.1: Deviant parameter values

	Happy	Sad
Pitch		
shift, *cents*	+50	-70
Inflection		
duration, *ms*	500	–
min., *cents*	-200	–
max., *cents*	+140	–
Shelf filter		
cut-off, *Hz*	>8000	<8000
slope, *dB/octave*	+9.5	-12

Parameter values of the happy and sad transformations used in this study (refer to Rachman et al., 2018 for details).

colleagues (2013; 2015): the neutral SV was presented as the identity deviant in the 'other sequence' and vice versa (results for this condition are not presented in this thesis). We counterbalanced the order of the sequences across participants. Each sequence contained 1080 stimuli in total with the standard stimulus occurring 80% of the time and each of the three deviant stimuli ('happy', 'sad' and 'identity') occurring 6.7% of the time (72 stimuli). Each sequence started with 10 standard stimuli and 2-7 standards occurred between successive deviants. All stimuli lasted 550 ms and were presented with a stimulus onset asynchrony (SOA) of 1000 ms.

Discrimination and intensity rating tasks

To replicate some of the findings from Chapter 2 and control that this new sample of participants behaved similarly, participants performed a behavioral self-other

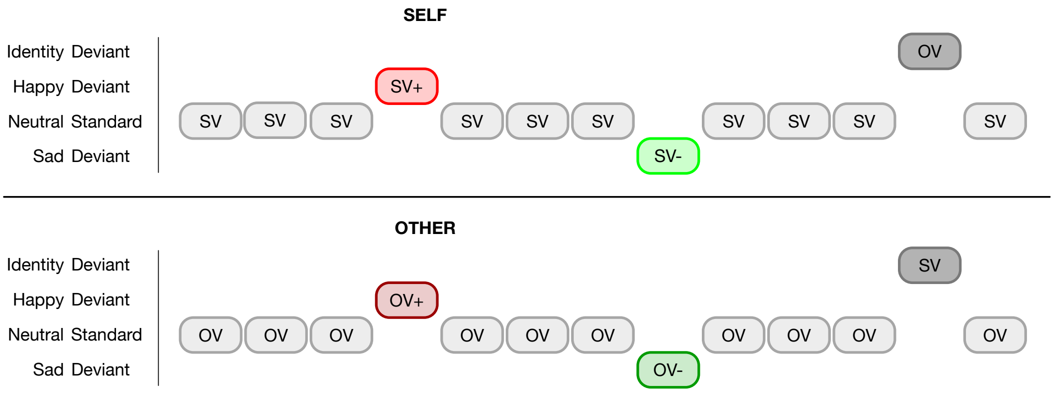


Fig. 3.4: Schematic representation of the oddball sequences for the self (above) and other (below) conditions. In the self-voice sequences, standards are neutral self-voice (SV) and deviants are happy (SV+) and sad (SV-) manipulations of the standard, as well as one other-voice recording of the same word (OV). In the other-voice sequences, standards are neutral other-voices (OV) and deviants are happy (OV+) and sad (OV-) manipulations as well as one self-voice recording of the same word (SV).

discrimination task after the EEG recordings. Five bisyllabic words and three pseudowords (Table 2.1), produced by the participant and another person (the same 'other' as was presented during the EEG recording) were presented in neutral, happy, and sad versions. Participants were asked to indicate for each stimulus if it was their own voice or the voice of someone else. Participants then listened again to the same stimuli and were asked to rate the emotional intensity of the voice on a continuous rating scale (0-100). All stimuli were separately randomized for both tasks.

Procedure

During the EEG recordings, subjects were seated in front of a computer screen (55×32 cm) on which they watched a silent subtitled movie. Participants were asked to pay attention to the movie and to ignore the sounds. Auditory stimulus presentation was controlled with PsychoPy (Peirce, 2007) and sounds were delivered through Sennheiser CX 300-II earphones at 70 dB SPL.

EEG data acquisition

Electroencephalographic (EEG) data were recorded from 63 scalp locations (actiCHamp, Brain Products GmbH, Germany) with a sampling rate of 500 Hz, relative to a nose tip reference, and filtered with a bandpass of 0.01-100 Hz (12 db/octave roll-off). Four electrodes were placed on the left and right temples (horizontal electrooculogram [EOG]) and above and below the left eye (vertical EOG) to monitor eye movements and blinks respectively.

Pre-processing and statistical analyses were performed in FieldTrip (Oostenveld et al., 2011). Offline, the continuous data were re-referenced to the average of the left and right mastoid electrodes (TP9 and TP10) and filtered with a 0.1 Hz high-pass filter (Butterworth, 12 dB/octave roll-off) and a 30 Hz low-pass filter (Butterworth, 48 dB/octave roll-off). The data were then visually inspected to remove epochs with artifacts, such as muscle activity and signal drifts. Next, eye blinks and movements were corrected using the fast independent component analysis (fastICA) method.

To get a better estimation of the MMN, we equated the number of deviants and standards by randomly selecting 69 standards (as many as the mean number of deviants after artifact rejection) that immediately preceded a deviant in the self and other sequences. Individual EEG epochs were averaged separately for each type of standard stimulus (self, other; Figure 3.5) and deviant stimulus (neutral self, neutral other, happy self, happy other, sad self, sad other), with a 200 ms pre-stimulus baseline and a 700 ms post-stimulus period. After artifact rejection, each subject had at least 75% trials remaining in each condition and the number of trials did not differ across conditions (Self standard: M=831.5, happy: M=69.7, sad: M=69.4; Other standard: M=831.3, happy: M=68.7, sad: M=69.0; $ps > .05$). Finally, four difference waves were calculated by subtracting the grand average waveform of the standard stimuli from each of the deviant grand averages within each sequence type (i.e., for each speaker separately), yielding the following conditions: 'Happy Self', 'Happy Other', 'Sad Self', 'Sad Other' (Figure 3.6).

Statistical analyses

Statistical analyses were conducted in Python 2.7. The alpha level was set at .05 and all statistical tests were two-tailed.

Behavioral data Accuracy scores and ratings were computed from the discrimination and intensity rating tasks respectively. We conducted one-sample *t*-tests on the accuracy scores to test whether SV and OV were discriminated above chance level (50%). For the intensity task, we conducted a two-way repeated-measures analysis of variance (rmANOVA) with the factors Emotion (neutral, happy, sad) and Identity (self, other). Significant main effects and interactions were followed up with Tukey HSD for post-hoc comparisons.

EEG data EEG data were analyzed using cluster-based statistics implemented in FieldTrip (Maris and Oostenveld, 2007). In total, four cluster-based permutation tests were performed: one on the standard grand-averages to test for an effect of speaker identity and three on the difference waves to investigate main effects of identity and emotion and their interaction. For the interaction we first calculated

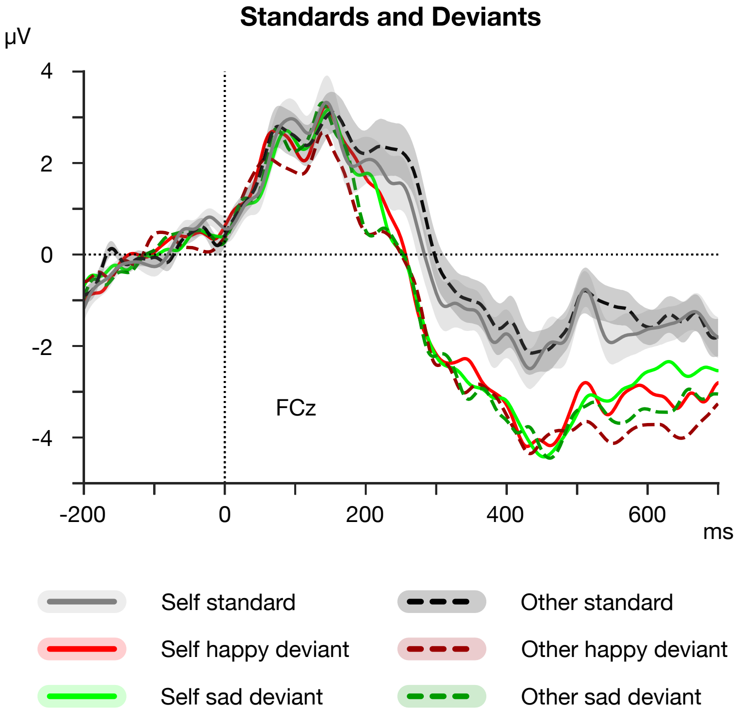


Fig. 3.5: Grand-average ERPs to the self (solid lines) and other (dashed lines) standard and deviant stimuli. Shaded area represents bootstrap standard error of the mean (SEM).

the difference between the happy and sad difference waves for each speaker identity separately before entering these data into the analysis. Based on prior hypotheses about the temporal location of the MMN component (e.g. Graux et al., 2013; Pinheiro et al., 2017; Beauchemin et al., 2006), analyses were carried out within a 50-300ms time window across all electrodes. For each cluster-based permutation test we first conducted pairwise *t*-tests between two conditions at each time-point and channel in the predefined time-window and region of interest. The cluster-level statistic was defined as the sum of all individual *t*-statistics that belong to the cluster. This procedure then controls the type I error rate by evaluating the cluster-level statistic under the randomization null distribution of the maximum cluster-level test statistic, obtained by 5000 permutations.

As an alternative parametric analysis strategy, we analyzed the mean MMN amplitude and MMN peak and onset latencies within a region of interest comprising electrodes F1, Fz, F2, FC1, FCz, FC2, C1, Cz, and C2. We extracted the mean amplitude over a 40 ms time window around the averaged MMN peak across conditions, participants, and electrodes (280 ± 20 ms), to avoid a possible bias introduced by differences in conditions. We extracted the MMN onset and peak latencies using a jackknife procedure and tested for differences in the four conditions (Identity × Emotion). The

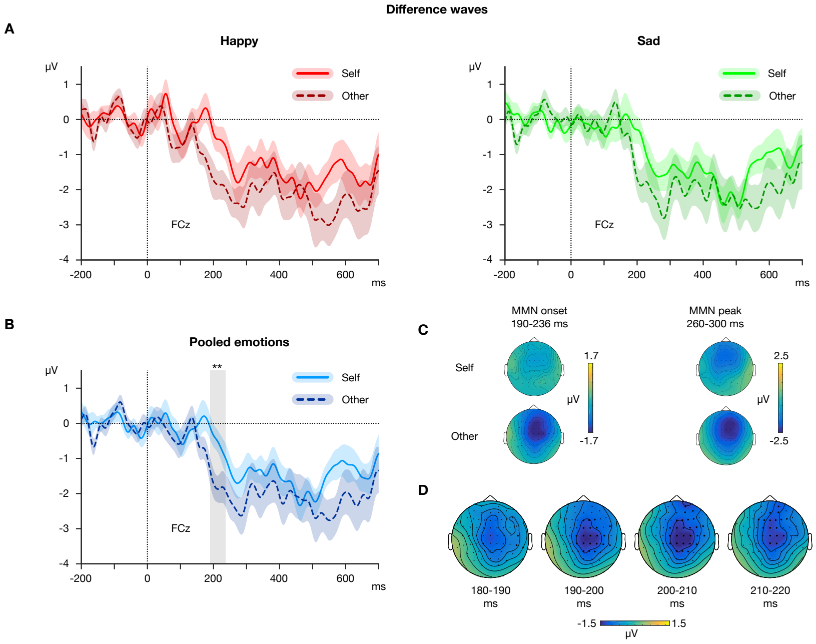


Fig. 3.6: Difference waves (A) of the happy (red) and sad (green) transformations on the self (light) and other (dark) voice. (B) Differences waves of the pooled happy and sad deviants of the self (light) and other (dark) voice. Shaded area represents bootstrap SEM, $^{**}p < .01$. (C) Topographies of the pooled happy and sad deviants of the self and other voice at MMN onset and peak. (D) Significant cluster of the contrast between the 'other' and 'self' difference waves represented in four 10 ms time windows between 180 and 220 ms. Highlighted channels belong to the cluster and were significant across the whole time 10 ms window.

jackknife procedure improves statistical power by taking the latencies of the grand average using a leave-one-out method (Kiesel et al., 2008; Ulrich and Miller, 2001): for N=23 participants, we calculated 23 grand averages, each leaving out one of the participants and including the other 22. We then determined the onset latency for each of these 23 grand averages as the time where the difference wave reached 50% of the MMN peak amplitude. In a similar way, we defined the MMN peak latency as the time at which the difference wave reached the most negative amplitude. These values were entered into two separate two-way rmANOVAs with identity (Self, Other) emotion (Happy, Sad), antero-posterior site (Frontal, Frontocentral, Central) and lateralization (1-line, z-line, 2-line) as within-subject factors. Finally, we divided the resulting F-value by $(N-1)^2$ to correct for the artificially low error variance introduced by the leave-one-out procedure (Ulrich and Miller, 2001). Furthermore, Greenhouse-Geisser correction for non-sphericity was applied when necessary. We report uncorrected degrees of freedom and corrected p-values.

Source localization Estimation of cortical current source density was performed with Brainstorm (Tadel et al., 2011). The cortical current source density mapping was obtained from a distributed source model of 15000 current dipoles. The dipoles were unconstrained to the cortical mantle of a generic brain model built from the standard MNI template brain provided in Brainstorm. EEG electrode positions were determined for each subject using a CapTrak system (Brain Products GmbH, Germany) and aligned to the standard MNI template brain. The forward model was computed with the OpenMEEG Boundary Element Method (Gramfort et al., 2010). A noise-covariance matrix was computed for each subject by taking the 200 ms baseline period of each trial and was taken into account in the inversion algorithm. The cortical current source density mapping was then obtained for each subject from the time series of each condition by means of the weighted minimum-norm estimate (wMNE). Z-scored cortical maps across all conditions were used to define the regions of interest that are activated irrespective of emotion and identity within the time window in which there was a significant difference between self and other conditions. Regions of interest (ROIs) contained at least 30 vertices with a z-score above 60% of the maximum z-score. To analyze the cortical sources of the difference waves we performed paired *t*-tests for each vertex within the defined ROIs, taking the mean values across the 190-230 ms window. This time window was chosen to span the interval between the average MMN onset latency in the OV condition (190 ms) and the average MMN onset latency in the SV condition (236 ms), in order to identify sources for the activity explaining the effect (see Results section). Activations within an ROI were considered significant whenever at least ten adjacent vertices reached statistical significance.

3.1.2 Results

Behavioral results

Results in the self-other discrimination task replicated those of Study 1. The accuracy of self-other discrimination was greater in OVs than in SVs (main effect of speaker identity: $F(1, 22) = 80.7$, $p < .001$), and there was also a main effect of emotional manipulation on discrimination accuracy ($F(2, 44) = 19.2$, $p < .001$) and an identity × emotion interaction ($F(2, 44) = 20.7$, $p < .001$), showing that manipulated self-voices were more easily confused for other identities than non-manipulated voices. Self-other discrimination was more accurate than chance for both neutral and emotional other- voices ($ts(22) > 19$, $ps < .001$), more accurate than chance in the neutral ($t(22) = 9.66$, $p < .001$) and sad self-voices ($t(22) = 2.60$, $p < .05$), but not in happy self-voices ($t(22) = -1.07$, $p > .05$; Figure 3.7, Left).

Auditory familiarity with the self-voice would predict, to some extent, that the perceived difference between standard and deviants would be emphasized in self-voice stimuli (see e.g. Fleming et al., 2014, for a related effect). Ratings of emotional intensity made after the EEG paradigm using similar stimuli did not show this effect. There was a main effect of emotion ($F(1, 22) = 21.63$, $p < .001$), in which sad voices were rated as less intense than happy ($p < .01$), but neither emotion differed from neutral (see Figure 3.7, Right). There was no effect or interaction with speaker identity, suggesting that emotional deviants were perceived as comparable in both types of sequences.

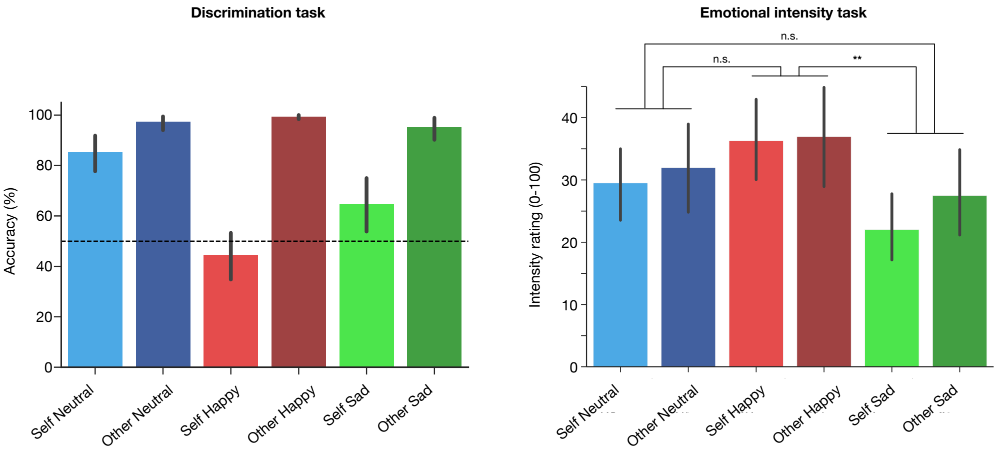


Fig. 3.7: Behavioral results of Study 3. **Left**: Discrimination accuracy (%) for neutral, happy, and sad versions of SV and OV. Dotted line indicates chance level performance (50%). **Right**: Intensity ratings (0-100) for neutral, happy, and sad versions of self and other voice. Error bars represent standard error of the mean. $^{**}p < .01$

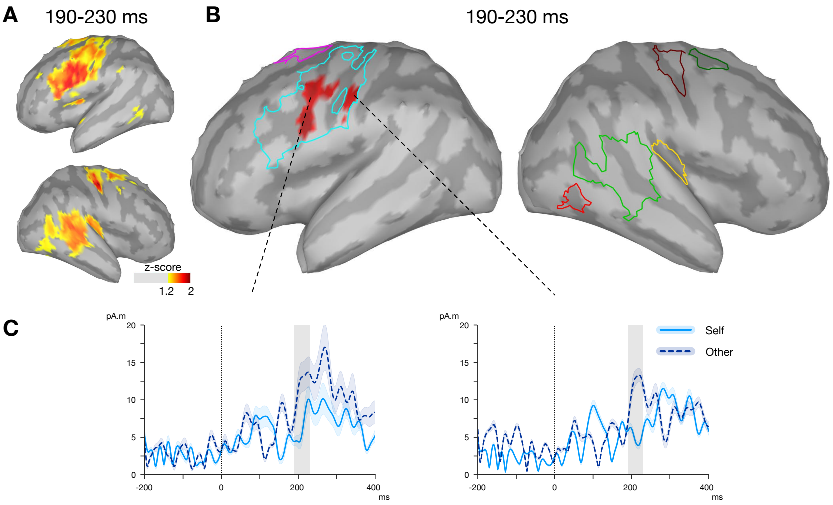


Fig. 3.8: EEG source analyses. (A) Source localizations across all conditions in the 190-230 ms showing maxima of activation (>60%, z-scores) to determine Regions of Interest (ROIs). (B) Modulations of cortical activity as a function of speaker identity in the 190-230 ms time window. Only clusters containing at least 10 contiguous vertices with $p < .05$ in this time window were considered statistically significant. The source activations are color-coded only for *t*-values corresponding to $p < .05$.(C) Time courses of the grand mean amplitude of the current sources in each activated region for self and other conditions. Shaded areas represent the standard deviation, the grey area represents the 190-230 ms time window in which the analyses took place.

Standards

The cluster-based permutation test and rmANOVAs did not reveal any differences between the self and other standard conditions ($ps > .05$).

Difference waves

Difference waves showed a relatively small (-2μV) fronto-central negativity peaking at 280 ± 20 ms, compatible with an MMN (Figure 3.6). The difference waves were also re-referenced to the nose reference to ensure the typical polarity inversion between Fz/Cz and the mastoid electrodes. However, because mastoid-referenced averages typically show a better signal-to-noise ratio than the nose-referenced averages, the former were used in all subsequent analyses (Kujala et al., 2007; Martínez-Montes et al., 2013).

We found a significant cluster when testing for a main effect of identity (Monte Carlo $p < .05$; Figure 3.6D) but none for a main effect of emotion or an interaction. Parametric analyses with the jackknife procedure revealed that this difference was driven by the onset of the MMN, rather than its peak. There was a main effect of identity on the MMN onset latency ($F_{\text{corrected}}$(1,22) = 10.14, $p < .01$), with the OV onset latency at 190 ms, compared to 236 ms in the SV condition, a considerable difference of 46 ms (see Figure 3.6A-C for the difference waves and topographies). There were no effects of emotion, electrode antero-posterior location or lateralization on onset latency, nor was there a significant interaction between any of the factors. In contrast, no main effects of identity or emotion were observed in the amplitude (-2 μV) or the peak latency (280±20 ms) of MMN peak and no interaction effects were observed on the MMN peak latency. The repeated-measures ANOVA on the mean MMN amplitude showed only an identity × lateralization interaction effect ($F(2, 44) = 7.04$, $p < .01$), but follow-up analyses at each antero-posterior site (Frontal, Frontocentral, Central) did not reveal an effect of identity (all $ps > .05$).

Sources

ROIs identified using source activation maps across all conditions in the 190-230 ms window (spanning the difference between other- and self- MMN onset latencies) included bilateral regions in the precentral gyri, large insulo-temporal regions in the right hemisphere and large fronto-parietal regions in the left hemisphere. Right-lateralized temporal activations are in line with previous MMN studies that reported right activations for pitch deviants in tones and voice (Jiang et al., 2014; Lappe et al., 2016). In addition, the right anterior insula is involved in processing vocal emotions Belin et al., 2004, and has also been associated to MMN responses to emotional syllable deviants (Chen et al., 2014).

Source activations for activity discriminative of OV versus SV deviants in these ROIs were stronger in the left precentral gyrus/sulcus (47 vertices), and the left postcentral gyrus (16 vertices) (Figure 3.8).

3.1.3 Discussion

The results presented in Chapter 2 showed an auditory processing advantage when recognizing emotional changes in self-voice stimuli, which did not appear to be mediated by the explicit recognition of the self speaker identity. Translated to the implicit MMN paradigm used here, this pattern of results would predict a larger MMN amplitude, or earlier peak MMN latency, for responses to emotional deviants in a sequence of self-voices. We observed no such difference in MMN amplitude and peak latency. Rather, we found an earlier MMN onset for emotional deviants on the *other-voice* compared to the self-voice. This MMN onset latency effect was seen in both emotional transformations, and amounted to a considerable difference of 46 ms.

Because we did not find any significant difference between the self and other conditions on the waveform of the standard stimuli, and because both self and other deviants were generated from the standards with identical algorithmic procedures, it is unlikely that such a large onset effect results from the differential processing of the standards, or differences in refractory states (Jacobsen and Schröger, 2001). Rather, the shorter MMN onset latency in the OV condition suggests that changes on a stranger's voice are highly prioritized in auditory processing. While this is in stark contrast with the accuracy advantage seen for self-stimuli in Chapter 2, it is possible that the auditory processing advantage to the self-voice seen in Chapter 2 and the prioritized change detection in the other-voice seen here are in fact two sides of the same coin. If the self-advantage is explained by the additional recruiting of articulatory/motor representations, it could be that richer self-voice processing is in fact delayed with respect to faster, but more shallow processing of other-voices. It is also possible that both effects (more accuracy for the self, faster/earlier responses to other-voices) result from distinct, independent processes - the former linked e.g. to auditory representations, and the latter to social relevance and attentional biases. In a recent study, effects of emotion were seen earlier in a communicative context when compared to a non-communicative context (Rohr and Abdel Rahman, 2015). It is therefore possible that our design of other-deviants in a sequence of other-standards is implicitly treated as a context akin to social communication ("other speaking to self"), more so than changes embedded in a sequence of self-sounds.

While their interpretation must remain conservative, source estimations for EEG activity discriminative of self and other mismatches during the MMN onset temporal window (190-230 ms) provide some elements to that question. Activity that occurred after the onset of MMN for OVs, but before its onset for SVs, did not occur within the typical supra-temporal or frontal MMN generators (Garrido et al., 2009), which suggests that processing other-voice stimuli was accompanied neither by any detectable enhancement of sensory processes, nor by any switch of attention. Neither did it occur within the predominantly right-lateralized regions previously associated with speaker identity tasks, such as the right temporoparietal junction (Schall et al., 2015) and right inferior frontal gyrus (Kaplan et al., 2008), or with MMN sources associated with emotional vocal stimuli such as the right anterior insula (Chen et al., 2014).

Instead, when contrasting responses to 'self' and 'other' deviants within the above ROIs, we found increased activations in the left precentral gyrus/sulcus and the left postcentral gyrus for deviants on the other voice. These regions suggest that emotional vocal deviants recruit a network of motor and somatosensory areas which are generally considered to be involved in mapping heard speech onto articulatory representations (Scott and Johnsrude, 2003; Evans and Davis, 2015; Skipper et al., 2017). The left somatomotor cortex in particular has been associated with phoneme discrimination tasks (Sato et al., 2009), and appears to be especially recruited for more effortful conditions involving noisy (Hervais-Adelman et al., 2012; D'Ausilio et al., 2012) or non-native speech (Wilson and Iacoboni, 2006), in which articulatory representations may provide a processing advantage. In the visual modality, left somatosensory areas have also been associated with unpredicted deviations from the self face (Sel et al., 2016), or facial emotion recognition in the other (Sel et al., 2014), both of which are also believed to involve processes of embodied simulation or predictions. Earlier activity in this network of regions for the other-voice deviants is therefore compatible either with a greater recruitment of resources for less-internally predictable signals such as speech produced by an unfamiliar stranger, for which listeners may lack an adequate internal template, or with the involvement of richer representations to process the self-voices, leading to the processing advantages described in Chapter 2.

In sum, the results of Study 3 suggest that expressive changes on a stranger's voice are highly prioritized in perceptual processing compared to identical changes on the self-voice. Other-voice deviants generate an earlier onset of the MMN without modulating the MMN amplitude. In the next study we aim to investigate whether this other-voice prioritization also manifests itself on a behavioral level. To this aim we conducted a pitch change perception task in which participants had to detect pitch changes in running speech. In Study 4, rather than using the emotional effects of the previous studies, we only manipulated pitch because it was the most important cue that was manipulated in the emotional transformations and because changing only one parameter made it easier to perform an adaptive procedure to equalize participants' task performance.

Two final methodological considerations are worth discussing. First, the MMN usually peaks at 150-250 ms from change onset, with this peak latency getting larger with the decreased magnitude, or increased processing difficulty, of stimulus change (Garrido et al., 2009). Here, across all conditions, the MMN peak latency was a relatively late 280 ms. It is possible that the late peak latency observed here reflects a late onset of observable stimulus change in our two-syllable words. In particular, spectral changes associated with happy or sad deviants may only become manifest on the vowel portion of the first syllable (onset ca. 100 ms, see Figure 3.3). In similar studies of two-syllable emotional words with a variety of changes (e.g. consonant duration, omission of second syllable, etc.), Pakarinen et al. (2014) report MMN peak latencies ranging between 126-355 ms post stimulus-onset, and Chen et al. (2014) a peak MMNm of 265 ms; in contrast, with single-vowel stimuli involving more immediate timbre changes and no initial consonant, Carminati et al. (2018) report MMN latencies around 200 ms. Future work should better document the temporal profile of physical information available in the signal to discriminate deviants from standards, in order to more precisely determine the chronometry of their auditory processing. Second, it should be emphasized that only female participants were included in this study. While both women and men typically show an MMN response to emotional deviants, previous work has showed that this preattentive response can be amplified in women, possibly because of a greater social relevance of emotional information for women (Schirmer et al., 2005). Importantly, this amplification seems to be specific to vocal sounds and has not been found in nonvocal sounds (Hung and Cheng, 2014). As such, it remains to be determined whether male participants show a similar difference in MMN onset latency as what we report here. Study 4 below will provide evidence that prioritized change detection in the other-voice also generalizes to male participants.

3.2 Study 4 - Faster detection of continuous pitch changes in other-voices

3.2.1 Methods

Participants

Twenty-four participants (12 female, 21 right-handed, mean age=22.3, SD=2.9 years) were recruited for this study. One female participant was excluded due to technical problems during the experimental session.

Stimuli

During the first part of the experimental session, each participant's voice was recorded in a sound-attenuated booth. Participants were asked to read twenty sentences out loud (see Table 3.2). Each sentence appeared on a computer screen, after three seconds the microphone was automatically turned on. Participants had four seconds to produce the sentence and the remaining time was indicated by a countdown on the screen. All voice samples had thus the same length and the experimenter verified whether the each recording contained the complete sentence. In case sentences were not recorded completely, participants were asked to do the recording procedure again. The voice recordings were made using a headset microphone (DPA d:fine 4066) and an external sound card (RME UCX Fireface) with a 44.1 kHz sampling rate and 16-bit resolution. All recordings were then normalized at 70 dBA using a Matlab toolbox (Pampalk, 2004). Recordings of two males and two females from a different study (not reported here) following the same procedure were used as stimuli for the "other-voice" condition.

Procedure

The experimental part of this study comprised two phases. In the first phase, participants performed a two-interval forced choice (2IFC) adaptive staircase procedure with a one-up, two-down progression rule converging to a 70.7% level on the psychometric function (Leek, 2001; Levitt, 1971). In each trial one of the twenty recorded sentences of the self-voice was presented twice. In one of the presentations the pitch was shifted up midway through the sentence and participants were asked to indicate whether they heard the change in pitch in the first or the second sentence. Participants could respond by pressing one of two buttons on a keyboard with their

Tab. 3.2: Recorded sentences

French	English
Les longs repas de famille au coin du feu me manquent	I miss the long family meals by the fireplace
Je suis toujours en retard au travail le lundi matin	I’m always late for work on Monday morning
J’ai déjà vu votre visage quelque part	I have already seen your face somewhere
J’aime beaucoup nos vacances de Noël en famille	I love our Christmas holidays with the family
Ils reprendront le travail après le déjeuner	They will resume work after lunch
J’aimerais beaucoup voyager mais je n’ai pas assez d’argent	I would love to travel but I don’t have enough money
Mon frère a déménagé de l’autre coté du pays	My brother moved to the other side of the country
Elle donne des cours de français pour arrondir ses fins de mois	She gives French lessons to make ends meet
Il aimerait devenir archéologue l’année prochaine	He would like to become an archaeologist next year
Je vous conseille ce site internet pour réserver un logement	I recommend this website to book accommodation
La première réunion va bientôt commencer	The first meeting will start soon
Je ne sais pas si je vais réussir mon examen	I do not know if I will pass my exam
Ils regardent un film assis dans leur canapé	They are watching a movie on their sofa
J’aime beaucoup le travail de ce compositeur	I really like the work of this composer
Je connais les chansons de cet artiste par coeur	I know this artist’s songs by heart
Elles n’ont pas assez dormi la nuit dernière	They did not sleep enough last night
Je n’arrive pas à fermer correctement ma valise	I can not properly close my suitcase
Tu as de la chance d’avoir trouvé un objet aussi rare	You are lucky to have found such a rare object
Ce musée expose de très belles oeuvres d’art	This museum exhibits beautiful works of art
Je n’ai jamais emprunté ce chemin auparavant	I never took this path before

French sentences and English translations used in Study 4 3.2.

right index and middle fingers after the second sentence had been presented. The initial pitch shift was set at 200 Hz and the initial step size fixed at 40 Hz and halved after each two reversals until a 5 Hz step size that was maintained until the end of the procedure. The pitch shifts were controlled by pyDAVID (see Figure 3.9 and Chapter 5 for further details). The timing of the pitch shift onset randomly varied between 1.0 and 2.0 seconds with 250 millisecond increments. The adaptive procedure stopped after twelve reversals and the pitch shift values of the last six reversals were averaged to estimate the pitch shift value targeting a 70.7% response accuracy. Only recordings of the participants' own voices were used in the adaptive staircase procedure.

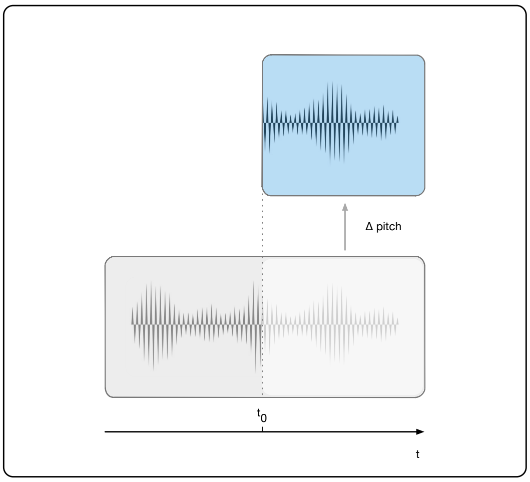


Fig. 3.9: Schematic representation of the use of DAVID to apply parametrically controlled changes of pitch on running speech. Audio recordings of neutral sentences are modified with a sudden positive shift of pitch (of a fixed, parametric magnitude) at a variable time onset, which participants are tasked to detect.

In the second phase, participants performed a pitch shift detection task following the same paradigm that was used in the adaptive staircase procedure. This time however, the size of the pitch change was fixed according to the outcome of the adaptive procedure. Trials included either recordings of the self-voice or recordings of the two other speaker identities of the same gender as the participant and participants performed twenty trials for each speaker.

Trial outliers with a response time exceeding M ± 2 SD were removed prior to analysis. All statistical tests were two-tailed and used an alpha-level set at .05.

3.2.2 Results

Across participants, pitch shifts of 45 ± 18 (M ± SD) were applied. A paired *t*-test of the accuracy for self and other did not show a significant difference between the two conditions [$t(22) = -0.8, p > .05$]. On the other hand, the paired *t*-test of the response time showed that responses were faster in the other condition compared to the self condition [$t(22) = 3.9$, $p < .001$; mean other = 0.96 s, mean self = 1.15 s].

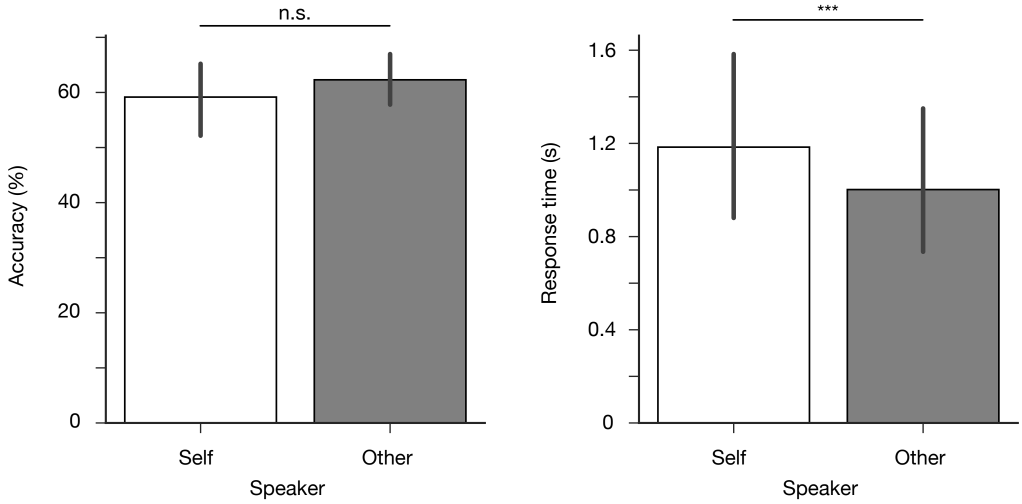


Fig. 3.10: Results of study 4. **Left:** accuracies of pitch shift detection on self and other. **Right:** response times for self and other. Error bars represent 95% confidence intervals, $^{***}p < .001$

3.2.3 Discussion

The results of this study show that participants' accuracy in pitch shift detection did not differ between the different speaker identities. Participants were as accurate to detect pitch changes on their own voice as they were on unknown voices. However, participants did respond faster for pitch changes on unknown voices. These results are in line with the results of the EEG study in which we found no differences in MMN amplitude, but rather an effect in timing that was reflected in an earlier onset of the MMN for changes on the voice of an unknown speaker. It extends the MMN results in two important directions: first, the effect is manifest in behaviour, resulting in responses that were 20ms faster for other-voices than self-voices; second, the effect is not only seen for composite happy/sad emotional changes, but for low-level acoustic changes in pitch.

3.3 Conclusion

Taken together, Chapters 2 and 3 seem to show two different effects. First, a self-advantage shown in Chapter 2 was reflected in better performance when the self-voice was evaluated explicitly for emotional content (Study 1), or implicitly for speaker identity (Study 2). In Chapter 3 we extended the investigation of implicit processes involved in emotional voice perception by conducting an MMN study and a behavioral study of change detection, this time showing that the detection of expressive changes in non-self voices is supported by earlier MMN onset (Study 3) and faster responses (Study 4). Source analyses in Study 3 show a different involvement of the motor/premotor system during the time-window corresponding to the latency difference betwen processing self- and other-voice stimuli. This suggests that both effects are in fact two sides of the same coin, with richer but slower auditory processing for the self-voice, and faster but shallower processing for other-voices.

This effect of increased familiarity with the auditory characteristics and/or the internal articulatory dynamics of the self-voice stimuli predicts that an other-voice effect would also manifest itself in an analogous manner for highly familiar other-voices, such as voices of the same sex or the same language. Using a cross-cultural design between France and Japan, the next chapter will examine whether people are more familiar with the sounds of their native language than sounds of a foreign language, and whether this familiarity translates to an "other-voice effect" on the recognition of cross-cultural emotions (Study 5), and the detection of low-level pitch changes (Study 6).

4

A language familiarity effect on the processing of pitch and emotional expressions

The results of the previous chapters show that, on the one hand, the self-voice benefits from an auditory processing advantage in both the recognition of emotion recognition and speaker identity, an effect we propose is based on greater auditory familiarity (Chapter 2); on the other hand, detection of pitch and emotional changes is prioritized on other-voices, both in electrophysiology and behaviour, plausibly because of greater social relevance (Chapter 3). Taken together, these results raise the question whether this 'other-voice' effect is specific to self versus other-voices, or might map onto other instances of vocal familiarity, such as processing the sounds of one's native - versus a foreign - language.

Box 4.1: Cross-cultural research in emotion science

The universality of emotions has been debated extensively in the domains of psychology and biology. Historically, this debate has focused on facial expressions, which were hypothesized to convey six basic emotions produced and recognized by people across all cultures (e.g. Ekman et al., 1969; Ekman and Friesen, 1971). Claims of universality are based on a tight link between the experience of emotion and the underlying physiological changes that share a common biological basis across the human species (Levenson, 2003). Recent work, however, has challenged the universality of facial expressions (e.g. Jack et al., 2012), as well as the notion of natural, basic emotions (e.g Barrett, 2006). In addition, cross-cultural research has expanded its fervor to vocal expressions of emotion in the past decade, most notably investigating nonverbal vocalizations such as growls or cheers (Laukka et al., 2013; Sauter et al., 2010; Sauter et al., 2015). While many studies report evidence that supports some level of cross-cultural validity to recognizing emotions in both the visual and auditory modality, an *in-group advantage* to recognizing emotions produced by member of one's own culture/language is often shown (Elfenbein, 2013), although the origin of in-group stimuli does not have to be explicitly recognized in order to provide this advantage (Sauter, 2013).

Moving from the personal sphere, where the self served as a stimulus, to the larger sociocultural sphere, in which the in-group language serves as a reference over the out-group, opens up connections with a number of new, broader research questions. Cross-cultural research constitutes a large part of modern emotion science, which is broadly concerned with the question of whether emotional expressions are universal (Box 4.1). Even beyond emotions, a large body of psychological evidence also shows that lack of familiarity with other cultures crucially affects how we process social signals such as their facial or vocal expressions. People recognize faces in their own culture more accurately than faces in other cultures (other-race effect; Shapiro and Penrod, 1986; Meissner and Brigham, 2001, and Figure 4.1), and identify speakers of their native language better than speakers of other languages (language-familiarity effect or LFE; Perrachione et al., 2011; Fleming et al., 2014). Even within a given cultural or language group, familiarity with a given speaker's voice may facilitate how his or her spoken words are remembered (Pisoni, 1993), or how ambiguous prosodic cues are processed semantically (Chen et al., 2016).

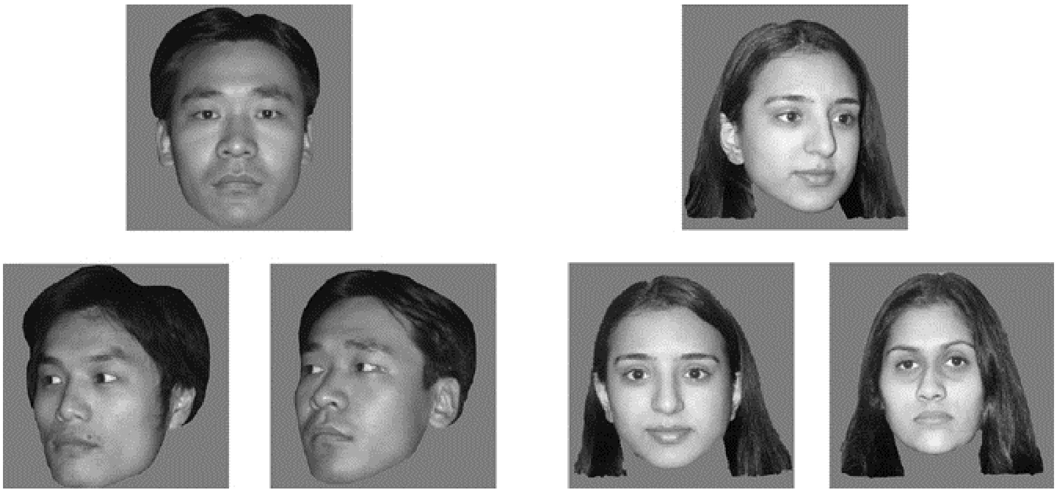


Fig. 4.1: The visual other-race effect: people recognize faces in their own culture more accurately than faces in other cultures. Left: Sample stimuli from a typical study showing a Chinese male face for habituation (top) and two test faces (novel and familiar) in the bottom. Right: A similar triad with Middle-Eastern female faces. Western observers typically perform more accurately on the latter than the former type of triad. Figure adapted from Kelly et al., 2007.

Interestingly, in the context of the self-other voice effects identified in this work, these effects of cultural familiarity are thought to result primarily from perceptual learning: differential exposure wraps an observer's perceptual space to better accommodate distinctions that are informative to discriminate between common in-group items, with the result that comparatively less-common out-group items are encoded in a less efficient manner (Valentine, 1991; Furl et al., 2002). Such perceptual warping is manifest for instance in the fact that listeners rate pairs of speakers of their own language as more dissimilar than pairs of speakers of the other language (Fleming et al., 2014), or that young infants can discriminate sounds from all languages equally well before 6 months of age, but develop a native-language advantage by about 6-12 months (Kuhl et al., 1992; Johnson et al., 2011). It would therefore appear plausible that, if a familiarity with the self-voice results in processing advantages to recognize emotions (Chapter 2), so would one observer's auditory familiarity with the sounds of a given language: judging whether a particular speech utterance is unusually high or low, a given phoneme bright or dark, whether a specific pitch inflection is expressive or phonological (Scherer and Oshinsky, 1977; Juslin and Laukka, 2003), all would appear to be advantaged, or conversely impaired, by acquired auditory representations optimized for the sounds of one language or another; conversely, if other people's voices are prioritized in auditory processing, possibly because of their social relevance, one could also predict differences in response times between in-group and out-group emotional expressions.

Most cross-cultural data indeed reveal an in-group advantage for identifying emotions displayed by members of the same rather than a different culture (Scherer et al., 2001; Thompson and Balkwill, 2006; Pell et al., 2009a; see Elfenbein et al., 2002 for review), showing that vocal emotion recognition is governed by both language-independent (universal) and language-specific processes. However, it has been difficult to conclusively determine to what extent such differences arise from language-familiarity effects in the sense of the above, or more generally from learned cultural differences in how these emotions are expressed (Elfenbein et al., 2002). For instance, LFEs would predict better cross-cultural recognition with increasing language similarity, but such evidence is mixed: Scherer et al. (2001) found Dutch listeners better at decoding German utterances than listeners of other, less similar European and Asian languages, but other studies found no differences in e.g. how Spanish listeners identified vocal emotions from the related English and German languages, and the unrelated Arabic (Pell et al., 2009a), or how English listeners processed utterances in the related German language and the unrelated Tagalog of the Philippines (Thompson and Balkwill, 2006). Most critically, because most of such cross-cultural investigations use stimuli recorded by actors of each culture, differences in the agreement levels across groups may simply arise because of cultural differences in emotional expressions. If I, a Dutch speaker, have difficulties processing emotional cues spoken by the people I meet while traveling through Japan, is it because my auditory representations of the Japanese phonetic inventory are poorly differentiated (see for instance Dupoux et al., 1999 for similar difficulties with French speakers), or because one simply does not use the same cues to express *joy* in Japanese and in Dutch (see Kitayama et al., 2006)? To progress on this debate, it would be necessary to employ emotional voice stimuli which, while being recognized as culturally-appropriate emotional expressions in different languages, utilize *exactly* the same prosodic cues in *exactly* the same manner (e.g. a 50-cent pitch increase on the second syllable), something which is at best impractical with vocal actors.

Once again, our software platform DAVID (Rachman et al., 2018) comes particularly handy in such an experimental context. In Study 5, we manipulate both Japanese (JP) and French (FR) voices with the same set of parametric transformations by DAVID, so as to display emotions of happiness, sadness and fear/anxiety, and let both native Japanese and native French speakers forced-choice categorize these emotions. This procedure excludes the possible effect of different emotion expressions and individual or cultural variability of recordings, because emotion categories of two languages are produced exactly in the same manner (i.e., with the same algorithmic parameters). In Study 6, we further test the effect of language familiarity with French of Japanese on the lower-level auditory processing of pitch changes in continuous French or Japanese sentences, using a similar experimental set-up as Study 4 which allowed to investigate both accuracy and reaction times.

4.1 Study 5 - Emotion recognition in native and foreign language

4.1.1 Methods

Participants

22 native Japanese (JP) speakers (9 female, M=19.7 years) and 22 French (FR) speakers (12 female, M=23.6 years) participated in the study. None of the JP speakers had ever learned French, and none of the FR speakers had ever learned Japanese. Experiments with the Japanese speakers were conducted in the University of Tokyo, Japan, while those with the French speakers were conducted at the INSEAD/Sorbonne Université Behavioral platform in Paris, France. Volunteers were recruited through local databases and mailing lists in the respective countries and were financially compensated for their participation. Two French participants were excluded because they did not satisfy language requirements (not native French speakers). Furthermore, one Japanese speaker was excluded because they claimed that they could hear the auditory stimuli only from one side of the headphone.

Stimuli

We prepared four normal JP sentences and four normal FR sentences, translated from the same four semantically-neutral English sentences (Russ et al., 2008). To test which linguistic component contributed to such LFE, we further prepared syntactically/semantically modified sentences, as well as time-reversed voices and sentences in a non-familiarized language for both participant groups (Swedish sentences). Jabberwocky variants (Hahne and Jescheniak, 2001) of the 8 sentences were created by replacing each content word in the normal sentences with a pseudo-word with the same number of syllables, without changing functional words (e.g. JP: *Uwagi wo wasureta* (I forgot my jacket) -> *Etaki* wo morushita**). Jabberwocky sentences did not have any meaning (or content), but still maintained the grammatical structures of their original languages. Shuffled Jabberwocky sentences were produced by changing the word order of the corresponding Jabberwocky sentences so that they did not maintain any grammatical structures of their original languages (e.g. JP: *Etaki wo morushita* -> *Wo* morushita* etaki**; see Table 4.1). The number of syllables was balanced across the two languages (7.25 ± 1.5 [mean ± SD] for the normal JP stimuli, and 7.5 ± 1.3 for the normal FR sentences), and the Jabberwocky and shuffled variants of each sentence had the same number of syllables as the original version, in both languages.

Tab. 4.1: Original sentences

JP sentences	FR sentences	English translations
Normal		
Uwagi wo wasureta	J'ai oublié mon pardessus	I forgot my jacket
Kaigi ni itta	Je suis allé à la réunion	I attended the meeting
Hikouki wa ippai	L'avion est complet	The airplane is full
Hyoumen wa nameraka	La surface est lisse	The surface is slick
Jabberwocky		
Etaki wo morushita	J'ai odrié mon tarpodu	
Goito ni etta	Je suis ijé à la boussion	
Komeubi wa ottai	L'ation est bondret	
Bouren wa soniyaka	La borpaque est nitte	
Shuffled		
Wo morushita etaki	Odrié j'ai tarpodu mon	
Etta ni goite	Boussion ijé la je à suis	
Komeubi ottai wa	L'est ation bondret	
Bouren soniyaka wa	La est nitte borpaque	

A male and female Japanese native speaker then recorded the four normal JP sentences, four jabberwocky JP sentences, and four shuffled JP sentences, and used the four utterances of the normal JP sentences to make four reverse JP recordings, by playing them backwards. Similarly, we recorded four normal FR sentences, four jabberwocky FR sentences, four shuffled FR sentences, and four reverse FR sentences in the same manner, with a male and female French native speaker. The recordings took place in a sound-attenuated booth, using Garage-Band software (Apple Inc.) with a headset microphone. In addition, we used four male and four female normal Swedish (SE) recordings from a previous study (Rachman et al., 2018). All stimuli were normalized in time to have a duration of 1.5 s, and normalized for the root mean square of the intensity.

Finally, all the above recordings (incl. reverse JP, reverse FR, and normal SE) were processed with DAVID and transformed into happy, sad, and afraid variants, resulting in 96 manipulated recordings for both JP and FR (3 emotions × 4 sentences × 4 conditions × 2 speakers), and 24 manipulated SE recordings (see Figure 4.2).

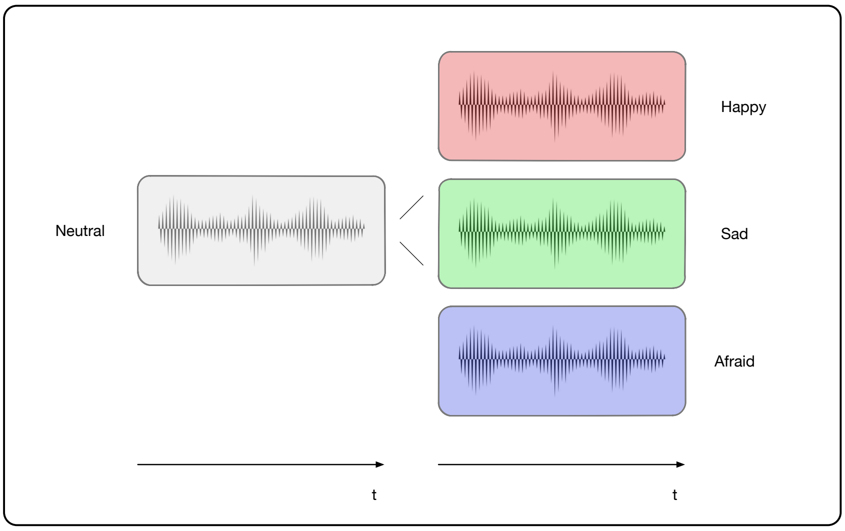


Fig. 4.2: Schematic representation of the use of DAVID to create the emotional transformations in Study 5. This set-up is similar to the one used in Study 1, only with complete sentences instead of single words, and one extra emotional transformation (afraid).

Table 4.2 describes the transformation parameter values used in this study to simulate the emotions of happy, sad, and afraid. These parameters, which were applied identically to both Japanese and French-language stimuli, are similar to those used for happy and sad in Studies 1, 2, and 3. In addition, we included an ‘afraid’ effect consisting of vibrato, a sinusoidal modulation of the voice’s pitch, with a rate (e.g. 8 Hz) and depth (e.g. + 40 cents) simulating a trembling, anxious voice. All three effects were validated to be both recognizeable and natural in both JP and FR (Rachman et al., 2018).

Tab. 4.2: Transformation parameter values

	Transformations		
	Happy	Sad	Afraid
Pitch			
shift, *cents*	$+50$	-50	–
Vibrato			
rate, *Hz*	–	–	8.5
depth, *cents*	–	–	30
Inflection			
duration, *ms*	500	–	150
min., *cents*	-200	–	-200
max., *cents*	$+140$	–	$+200$
Filter			
cut-off, *Hz*	> 8000	< 8000	–
slope, *dB/octave*	$+9.5$	-12	–

Parameter values of the happy, sad and afraid transformations used in Study 5 (refer to Rachman et al., 2018 for details).

Procedure

In each trial, participants listened to pairs of utterances of the same sentence by the same speaker (presented with an inter-stimuli interval= 0.7-1.3 s). The first utterance was always the neutral recording and the second utterance was either the same recording unprocessed (neutral condition) or processed with one of the emotional transformations (happy, sad, afraid). After hearing the two utterances, participants (all right-handed) were instructed to answer whether the second utterance sounded happy, sad, afraid, or neutral by pressing one of four keys ("S", "D", "F", and "G") with their left fourth, third, second, and first finger, respectively. The next trial started when participants pressed the "ENTER" key with their right first finger. All participants were presented with the 96 JP pairs, 96 FR pairs, and 24 SE pairs, randomized across participants in one single block. The correspondence of keys and response categories was randomized across trials. Visual stimuli were displayed on a laptop-computer screen, and the voice stimuli were presented through closed headphones. Stimulus presentation and data collection were controlled using PsychoPy toolbox (Peirce, 2007). Response categories in the recognition task for the French group in fact used the English terms (happy, sad, afraid) instead of the French equivalents, but were defined in the instructions using the equivalent French terms. Response categories used in Japanese groups were presented in Japanese terms.

Data analysis

We computed mean hit rate of the three emotion categories (happy, sad, and afraid) for each of the nine conditions (JP, FR: normal, jabberwocky, shuffled, reversed; SE: normal). The Swedish sentences are excluded from the analyses here. To take a possible response bias into account, we also calculated unbiased hit rates (Hu) for each participant, averaged over the three emotion categories (Wagner, 1993).

4.1.2 Results

Both participant groups recognized the three emotional transformations above chance level (25%) when applied to the normal speech utterances in their native language. For French participants, average hit rate over the three emotional categories was significantly larger than chance level in the normal French sentences, but not in the normal Japanese sentences ($\alpha_{Bonfer.} = .0125$). Average hit rates for the Japanese participants were higher than chance level in the normal French and Japanese sentences. In both participant groups, hit rates on native language were at comparable levels [FR: 37%; JP: 44%; $t(39) = 1.40$, $p = .85$, $d = 0.44$]. These results confirmed both that participants had good emotion recognition abilities in their native language, and that the emotional transformations used in this study are appropriate (and of comparable discriminability) in both languages.

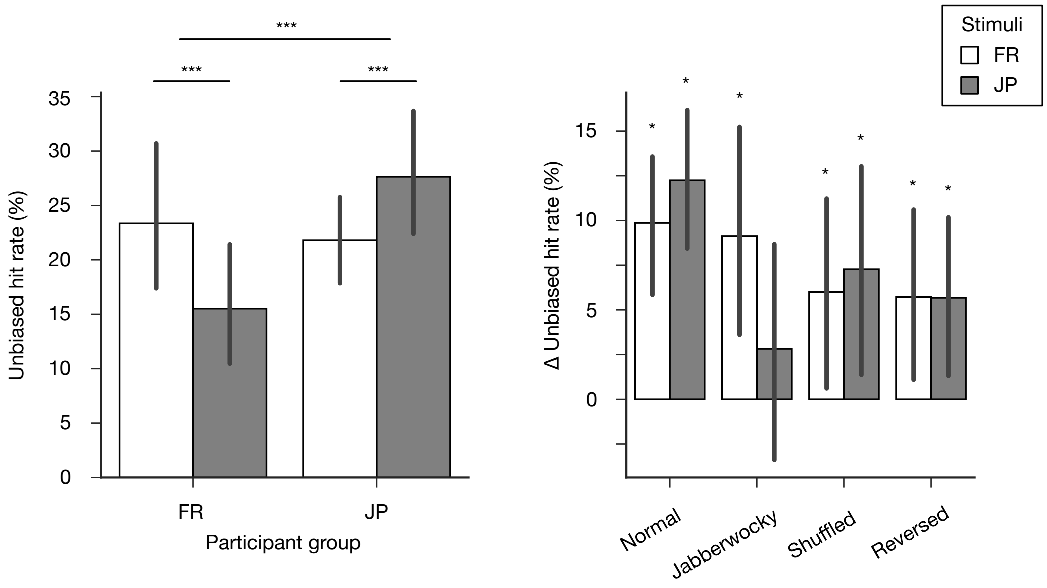


Fig. 4.3: Results of Study 5. **Left:** Unbiased hit rates averaged over the 3 non-neutral emotion categories, and the 4 sentence types. In each participant group, unbiased hit rates for French (white) and Japanese (grey) stimuli are shown. **Right:** The effect of language familiarity on the unbiased hit rates ($\mathrm{Hu}_{\mathrm{Native}}$ - $\mathrm{Hu}_{\mathrm{Foreign}}$) averaged over the 3 non-neutral emotion categories per sentence type (normal, jabberwocky, shuffled, and reversed). ***p < .001. Error bars represent 95% confidence intervals.

To test for language-familiarity effects, we then examined unbiased hit rates. A two-way repeated measures analysis of variance (rmANOVA) of Participant group × Stimulus language showed a significant interaction [$F(1{,}39) = 22.02$, $p < .001$, $\eta_p^2 = 0.36$], as well as a significant main effect of participant group [$F(1{,}39) = 8.3$, $p < .01$, $\eta_p^2 = 0.18$]. The post-hoc t-tests ($\alpha_{Bonfer.} = 0.025$) showed that FR participants could detect emotion categories of normal FR sentences significantly better than normal JP sentences [$t(19) = 3.60$, $p < .001$, $d = 0.75$], and JP participants could detect emotions in normal JP sentences significantly better than in normal FR sentences [$t(20) = 3.0$, $p < .01$, $d = 0.60$; see Figure 4.3, Left]. These results indicate a clear LFE.

This LFE was observed in a quasi-identical manner in all three degraded stimulus conditions. When participants were tasked to decode the same emotional cues on reversed JP and FR stimuli, average hit rate was around chance level for both participant groups. However, unbiased hit rates showed a significant interaction between Participant group and Stimulus language [$F(1{,}39) = 14.05$, $p < .001$, $\eta_p^2 = 0.26$], with French participants better in reverse FR than reverse JP sentences ($t(1{,}19) = 2.53$, $p = .010$, $d = 0.41$), and Japanese participants better in the reverse JP than reverse FR sentences [$t(1{,}20) = 2.78$, $p < .01$, $d = 0.54$; Figure 4.3, Right].

Average hit rate of the FR participants was at chance level for both jabberwocky and shuffled sentences, while JP participants showed better-than-chance performance in jabberwocky JP, shuffled JP, and shuffled FR sentences ($ps < .01$). Unbiased hit rates again showed a significant interaction of Language × Participant group in both jabberwocky sentences [$F(1{,}39) = 12.06$, $p = .001$, $\eta_p^2 = 0.24$], and shuffled sentences [$F(1{,}39) = 9.37$, $p = .004$, $\eta_p^2 = 0.19$]. FR participants were better in the jabberwocky FR than jabberwocky JP sentences [$t(1{,}19) = 3.23$, $p = .002$, $d = 0.58$], and JP participants were better in shuffled JP than shuffled FR sentences [$t(1{,}20) = 2.32$, $p = .016$, $d = 0.57$]. Language differences were marginal in jabberwocky sentences for JP participants [$t(1{,}19) = 2.04$, $p = .028$, $d = 0.35$], and in shuffled sentences for FR participants [$t(1{,}20) = 1.42$, $p = .086$, $d = 0.26$].

Finally, we analyzed the three emotion categories (happy, sad, afraid) separately by averaging all of the four sentence types (normal, reverse, jabberwocky, shuffled), since all of these conditions showed an interaction effect of language and participants group (Figure 4.4). For FR participants, the LFE affected all three emotions identically. A two-way rmANOVA of Emotion category × Speaker language showed a significant main effect of Emotion category [$F(1{,}19) = 22.97$, $p < .001$, $\eta_p^2 = 0.55$], as well as a significant main effect of Speaker language [$F(2{,}38) = 3.63$, $p = .003$, $\eta_p^2 = 0.16$], but no interaction between emotion and language. For JP participants however, we found a significant main effect of Emotion type [$F(1{,}20) = 28.05$, $p =$

.003, $\eta_p^2 = 0.58$], main effect of Speaker language[$F(2,40) = 8.78$, $p < .001$, $\eta_p^2 = 0.31$], and an interaction [$F(2,40) = 9.45$, $p < .001$, $\eta_p^2 = 0.32$]. Further analyses showed simple main effects of Speaker language in Happy and Afraid emotions (*ps* $< .001$), but not in Sad emotion ($p > .05$).

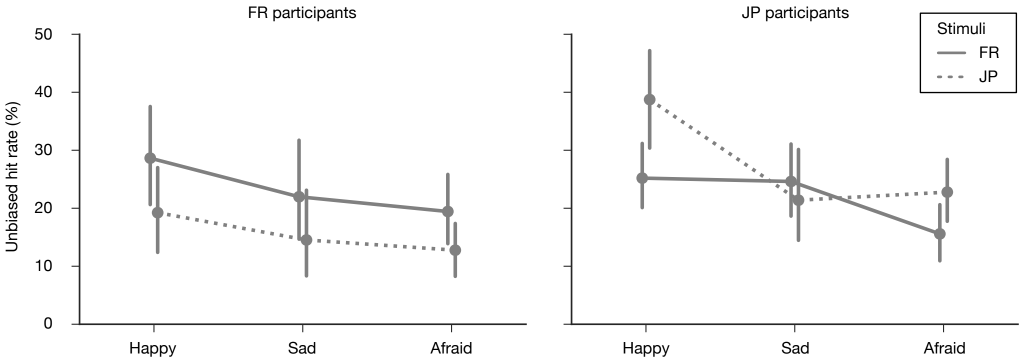


Fig. 4.4: Unbiased hit rates for each emotion category for both FR (Left) and JP participants (Right), averaged across normal, reversed, jabberwocky, and shuffled conditions. Solid line shows unbiased hit rate of French sentences, while the dashed line shows that of Japanese sentences. Error bars represent 95% confidence intervals.

4.1.3 Discussion

The results of this study provide univocal evidence that production differences (or “speaker dialects” - Scherer et al., 2001) are not the sole drivers of in-group advantages in cross-cultural emotion perception. Even when we controlled e.g. happiness to be expressed with a precisely-calibrated pitch change of +50 cents in both languages, participants more accurately recognized such changes when they occurred in their native language. Critical to this manipulation is the fact that both groups reached identical performance in their native language, showing that the computer-generated cues were equally discriminative in both languages. It could have been that computer manipulations differed in saliency depending on the phonetic characteristics of the language on which they were applied (e.g. vibrato needing relatively large consonant/vowel ratio, see Rachman et al., 2018), but this does not seem to be the case for the cues and the languages considered here. This native-language advantage can therefore only be explained by an interaction between the processes of encoding the linguistic or paralinguistic features of the familiar and non-familiar language and the processes responsible for extracting the emotional features important for the task - in short, by a language-familiarity effect of the kind already known for other-race face or speaker recognition (Meissner and Brigham, 2001; Perrachione et al., 2011), and strongly reminiscent of the auditory advantage to recognizing emotions in the self-voice seen in Chapter 2.

It is of particular note that, when we contrasted normal stimuli with three degraded conditions, breaking semantics (grammatical sentences with non-words, or jabberwocky - Hahne and Jescheniak, 2001), syntax (shuffled jabberwocky), and suprasegmental patterns (reversed speech, see e.g. Fleming et al., 2014), we found clear LFEs in all three conditions. This suggests that low-level auditory/phonological basis for the effect, i.e. that it is the listeners' lack of familiarity with the individual speech sounds of the other language that primarily impairs their processing of higher-level emotional cues. This finding is consistent with LFEs in speaker recognition, which are preserved with reversed-speech (Fleming et al., 2014) but absent in participants with impaired phonological representations (Perrachione and Wong, 2007), but less intuitive in the case of emotional prosody, which processing is in well-known interaction with syntax (Eckstein and Friederici, 2006) and semantics (Kotz and Paulmann, 2007). As for the self-voice, emotional cue extraction may be facilitated in the case of native-language because participants have better auditory and/or articulatory-motor representations of what is phonological, and are therefore better able to distinguish what is incidental and expressive, in their own language. In short, other-voice effects on emotional voice perception extends not only to other members of the in-group (me vs others), but also to other-language members of the out-group (us vs them). Because these effects have likely a low-level, phonological nature, one would predict that they would extend to basic pitch processing, i.e. that native participants would process non-semantic variations of pitch more accurately on their own language. Study 6 below will provide evidence that this is indeed the case.

Beyond the self, Study 5 has broader significance for cross-cultural research in how emotions are communicated. First, the percentage accuracy effect size of the native-language advantage in this study (FR: +7.2%, JP:+7.9% unbiased hit rate; Figure 4.4) was comparable with that of meta-studies of the in-group advantage in cross-cultural emotion recognition (+9.3%, Elfenbein et al., 2002, p.216). The fact that studies included in this meta-analysis typically subsumed both speaker- and listener-level influences on emotion recognition, and in particular differences in how the emotions were displayed by actors cross-culturally, suggests that the perceptual LFE uncovered here is by no means a minor contributor in cross-cultural misperceptions, but may rather explain a large proportion of such effects.

Second, irrespective of listener language, happiness was more distinct in its recognizability across languages (M=+12.1% unbiased hit rate) than afraid (M=+7.1%) and sadness (M=+3.3%). This is consistent with previous meta-studies of in-group advantage in the voice modality (happiness: +17.5%; afraid: +12.9%; sadness: +7.6%; Elfenbein et al., 2002, p.225), and confirms the general notion that expressions of happiness are a lot less universal in the vocal modality (Pell et al., 2009a; Juslin and Laukka, 2003) than they are in the (overwhelmingly smile-driven) facial modality (Elfenbein et al., 2002; Jack et al., 2012), possibly here because they rely on cues that require sufficiently accurate phonological representations of the target language to be extracted successfully. Interestingly, this would be the case, e.g., of the smiling oro-facial gesture which, universal as it may be in the visual domain, translates acoustically to fine formant-related variations of vowel spectra (Ponsot et al., 2018b).

Finally, we found that the performance of Japanese participants did not differ between native- and foreign-language stimuli for the sad emotion, although it did for happy and afraid. It is possible that, while the computer-generated cues used here were generally appropriate for all emotions in both cultures, they were comparatively further away from the cultural norm of how sadness is expressed in the Japanese language. This data may also result from different boundaries between emotional terms: it is possible that the cues manipulated in e.g. the happy effect spanned a larger proportion of the vocal expressions referred to as “sad” (*kanashimi*) in Japanese than the proportion of expressions referred to as “sad” (*triste*) in French.

Several aspects of the current study make a complete analogy with the other-voice effects reported in Chapter 2 and 3 somewhat challenging. First, while hit rates in both participant groups were significantly above chance level in the native language, participants’ performance remained relatively low. This makes it difficult to assess the effect of stimulus degradation on participants’ performance as this may be limited by floor effects. The next study, Study 6, will use changes in pitch rather than emotional transformations, allowing to normalize participants’ performance in their native language with an adaptive procedure, similar to Study 4. Second, rather than emotion recognition, Study 6 will test participants’ ability to detect sudden pitch changes in continuous sentences, thus allowing to conclude on the low-level, auditory basis of the effect, as well as to measure how it manifests itself on both response accuracies and (as before in Study 4) response times.

4.2 Study 6 - Pitch change detection in native and foreign language

4.2.1 Methods

Participants

Twenty-four native French (FR) speakers (12 female, 21 right-handed, mean age=22.3, SD=2.9 years) and twenty right-handed native Japanese (JP) speakers (8 female, mean age=23.6, SD=6.2 years) participated in this study. The French participants had also participated in Study 4 one week prior to the current study.

Stimuli

The base stimuli used in this experiment were the neutral French and Japanese stimuli used in Study 5. We had four sentence types (normal, Jabberwocky, shuffled, reversed), produced by a male and a female speaker in each language. Instead of being modified by DAVID to create emotional variants and asking participants to discriminate them, we used DAVID to create sudden pitch changes at random time points and ask participants whether they could detect their occurrence (see procedure below).

Procedure

Participants performed two separate adaptive staircase procedures to equalize their performance. The order of the two staircase procedures was counterbalanced across male and female participants separately. The stimuli used in this protocol were the same normal sentences in the participants' native language that were also used in Study 5, which therefore served as the reference condition in this study. The adaptive procedure involved a two-interval forced choice (2IFC) paradigm with a one-up, two-down progression rule to converge to a 70.7% level on the psychometric function (Leek, 2001; Levitt, 1971). For each trial, a normal type sentence in the participant's native language was randomly selected and presented twice. A pitch shift was applied at a random point in one of the two sentences and participants were asked to respond in which sentence they heard the pitch change. Responses were made with the left middle and index fingers by pressing one of two buttons on a keyboard. The initial pitch shift was set at 200 Hz and the initial step size fixed at 40 Hz and halved after each two reversals until a 5 Hz step size that was maintained until the end of the procedure. The pitch shifts were controlled by pyDAVID (see Figure 4.5 and Chapter 5 for further details). The onset of the pitch shift in each trial was randomly set at 500 ± 100 ms. After twelve reversals, the adaptive procedure was terminated and an average of the pitch shift values of the last six reversals were was taken to estimate the value targeting a 70.7% response accuracy.

In a subsequent pitch shift detection task, participants performed the same paradigm that was used in the adaptive staircase procedure, with the fixed pitch shift magnitude defined in the previous phase. The pitch shift was applied on all stimuli of the speakers of the same respective sex across the whole experiment. Participants took part in three blocks in total. In each block, they were presented with the 16 different sentences in both languages, and produced by both speakers in each language. These 64 sentences were randomized in repeated in each of the three blocks.

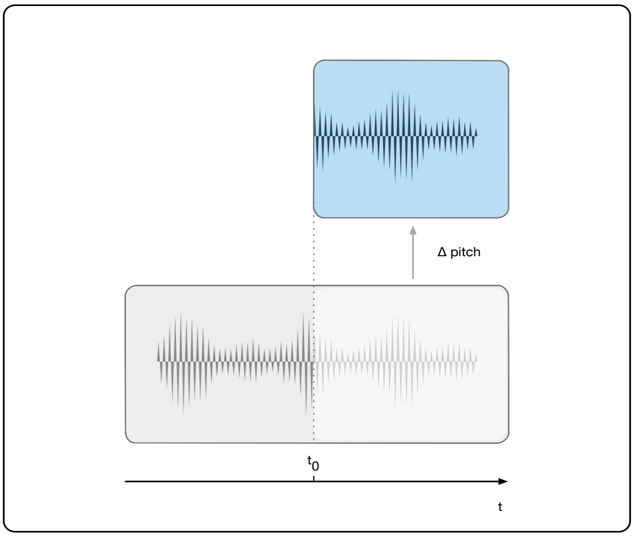


Fig. 4.5: Schematic representation of the use of DAVID to apply parametrically controlled changes of pitch in running speech in Study 6. This set-up is similar to the one used in Study 4.

All statistical tests were two-tailed and used an alpha-level set at .05. Where appropriate, Greenhouse-Geisser corrections for non-sphericity were applied. Uncorrected degrees or freedom and corrected p-values are reported. Significant effects were followed up by paired *t*-tests using Bonferroni corrections for multiple comparisons.

4.2.2 Results

An independent *t*-test showed that the magnitudes of the pitch shift applied in this study, based on the adaptive staircase procedure, did not differ between the two participant groups (FR: 70 cents; JP: 81 cents; p > .05).

Native vs. foreign language A mixed 2×2×4 (Participant language × Speaker language × Condition) ANOVA was conducted on the response accuracies. There were main effects of Participant language [$F(1, 42) = 7.8$, $p < .01$], Speaker language [$F(1, 42) = 5.4$, $p < .05$], and Condition [$F(3, 126) = 4.5$, $p < .01$]. Additionally, there was a significant interaction between Participant language and Speaker language [$F(1, 42) = 7.9$, $p < .01$]. Post-hoc pairwise *t*-tests for the main effect of condition showed that the accuracy for the normal sentences was significantly higher than for the reversed sentence. There were no significant differences in accuracy between any of the other sentence types. Follow-up *t*-tests for the interaction between Participant language and Speaker language showed that accuracies in the French participant group were higher for French stimuli than for Japanese stimuli ($p< .01$). The accuracies in the Japanese participant group did not differ between French and Japanese stimuli (see Figure 4.6).

A mixed three-way ANOVA of Participant language × Speaker language × Condition on the response times only showed a significant main effect of Condition [$F(3, 126) = 23.5$, $p < .001$], but no other significant main or interaction effects. Post-hoc pairwise *t*-tests for the main effect of condition showed that response times to Jabberwocky and Shuffled sentences did not differ significantly, but that in both of these conditions response times were slower than in the Normal condition, and faster than in the Reversed condition ($ps < .01$).

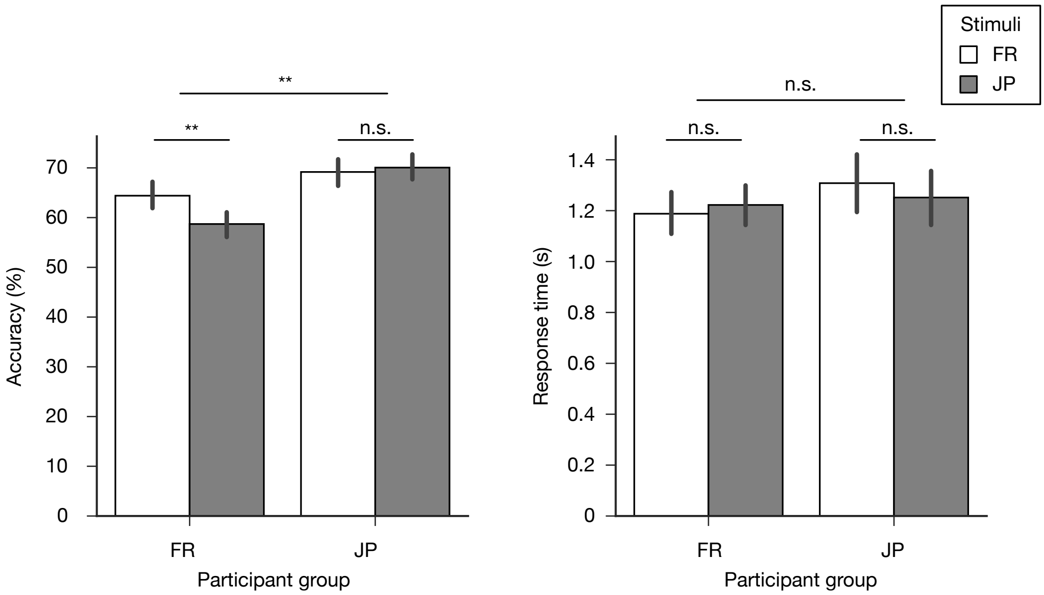


Fig. 4.6: Results of Study 6. Accuracies in percentage (Left) and response times in seconds (Right) averaged over the 3 non-neutral emotion categories, and the 4 sentence types. In each participant group, accuracies for French (white) and Japanese (grey) stimuli are shown. $**p < .01$. Error bars represent 95% confidence intervals.

Same-sex vs. opposite-sex speakers In a secondary exploratory analysis, responses on same-sex speakers and opposite-sex speakers were analyzed separately. Figure 4.7 shows the differences of performance in the native language and the foreign language (FR participants: accuracy$_{FR}$ - accuracy$_{JP}$; JP participants: accuracy$_{JP}$ - accuracy$_{FR}$) for each sentence type. Here, the French participants showed a stronger language familiarity effect reflected in higher values of Δ accuracy in same-sex speakers. Moreover, in same-sex stimuli this difference deviates from zero in both the normal and jabberwocky sentence types, whereas difference in accuracy deviates from zero only in the normal sentences types for stimuli of the opposite-sex.

To further examine the effect of speaker sex independently of language, accuracy scores of each participant's native language stimuli were entered into a mixed 2×2 (Participant sex × Speaker sex) ANOVA. The results showed no main effects of Participant sex or Speaker sex, but the interaction between these two factors was marginally significant [$F(1,42) = 3.6$, $p = .066$]. Post-hoc *t*-tests for this interaction showed that female participants responded more accurately for female voices than male voices (p<.05). Accuracy for male and female voices did not differ in the male participants (see Figure 4.8, Left). Similarly, response times were entered into a mixed two-way ANOVA of Participant sex × Speaker sex. These results showed a significant interaction between Participant sex and Speaker sex [$F(1,42) = 6.0$, $p <$ .05]. Post-hoc t-tests for this interaction showed that female participants had shorter response times to female voice compared to male voices ($p < .01$). Response times in male participants did not differ significantly between male and female voices (see Figure 4.8, Right).

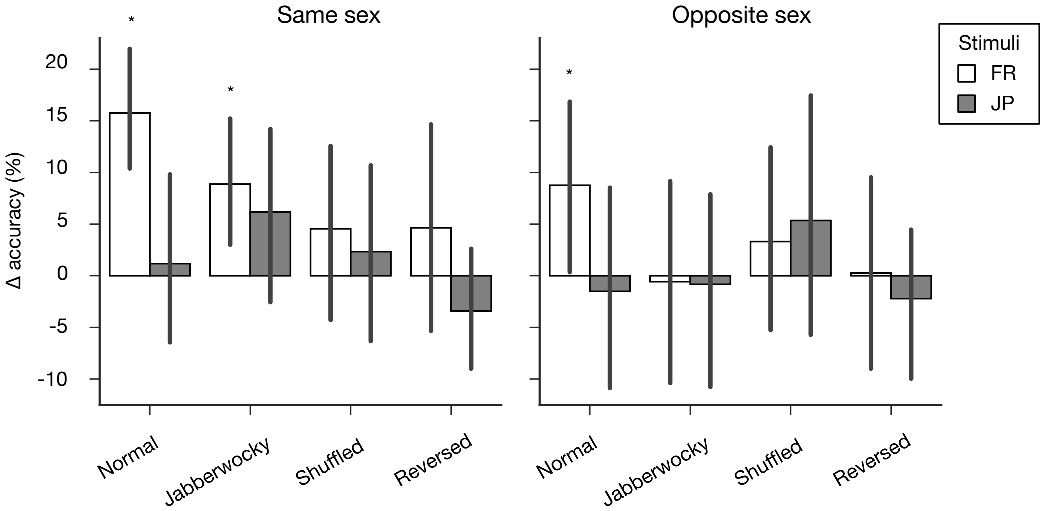


Fig. 4.7: The effect of language familiarity on accuracies (native - foreign) averaged over the 3 non-neutral emotion categories per sentence type (normal, jabberwocky, shuffled, and reversed), grouped according to speakers were of the same or different sex as the observer. Error bars represent 95% confidence intervals, *indicate Δ accuracies of which the confidence intervals do not cross zero.

4.2.3 Discussion

The results of Study 6 globally replicate those of Study 5 by showing a language-familiarity effect to detecting sudden pitch changes in sentences, with more accuracy in participants' native language. While the effect was mostly driven by the French participant group, this suggests that the in-group effect found in Study 5 for vocal emotion recognition is at least partly driven by low-level auditory processing skills, in this case pitch perception. Because pitch change detection is a lower-level, more robust auditory process than the extraction and categorization of emotional cues, it is possible that accuracy effects at that level are more difficult to observe experimentally (Study 4, Study 6) than for emotions (Study 1, Study 5). Nevertheless, it is quite remarkable that something as basic as the ability to detect a +30 cent increase of pitch in running speech is influenced by one observer's familiarity with the sounds of a given language, and such an effect is bound to have important consequences down the auditory processing line on all judgments based on pitch variation: not only

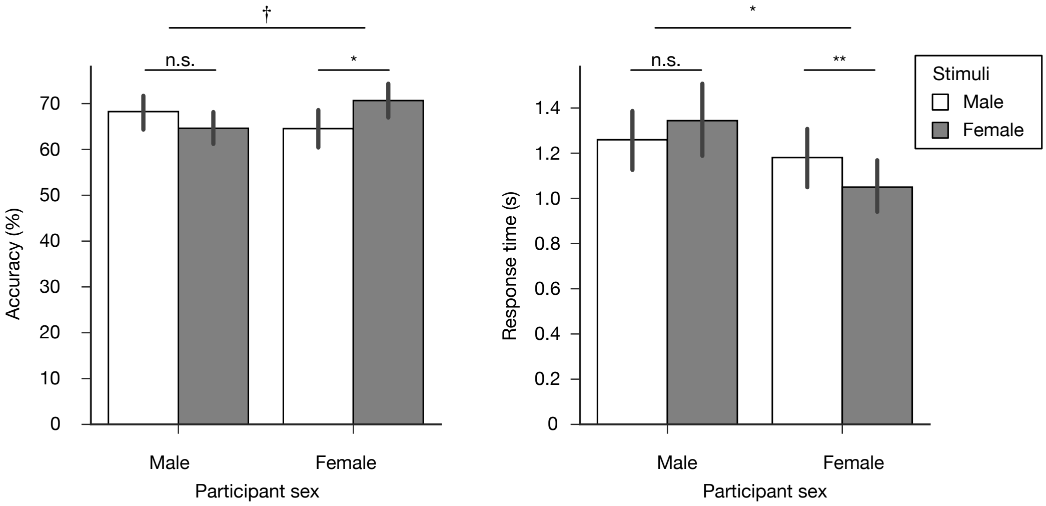


Fig. 4.8: The effect of speaker and listener sex. Accuracies in percentage (Left) and response times in seconds (Right) for male and female participants, averaged over the 3 non-neutral emotion categories and the 4 sentence types of each participants' native language. In each participant group, accuracies for male (white) and female (grey) speakers are shown. † p=.066, *p < .05, **p < .01. Error bars represent 95% confidence intervals.

emotional expression (Study 5), but also syntactic or sentence mode information (e.g. whether a sentence is interrogative or declarative), stress (e.g. on what word is the sentence's focus) or attitudinal content (e.g. whether a speaker is confident or doubtful) (Gussenhoven, 2004). When already pitch changes are difficult to process in a foreign language, a lot more can expected to be *lost in translation*.

Because an exploratory analysis showed that the LFE on pitch change detection was more prominent when participants listened to same-sex compared to opposite-sex speakers (Figure 4.7, we then re-analyzed the subpart of data that corresponded only to native language stimuli, i.e. French participants on normal-condition French sentences, and Japanese participants on normal-condition Japanese sentences, and seemingly also found evidence for an other-sex effect on the accuracy of detecting pitch changes (Figure 4.8, Left): independent of language, female participants were significantly more accurate when detecting pitch changes on samples of other female speakers than of male speakers, and male participants showed the opposite trend. It is tempting to interpret this effect in the same line as the other-voice and other-language effects found earlier: that greater familiarity (through the self-voice) with the vocal characteristics (pitch and phonology) of a speaker's own sex creates an auditory advantage to process pitch changes in other same-sex speakers. It is interesting to note that, when opposite-sex effects were found for social signals in previous research, they were often in the direction of facilitating other-sex stimuli, and interpreted in evolutionary terms suggesting that people's attention is biased towards potential mates (Feinberg, 2008). Such effects were notably reported for the perception of voice gender (both males and females categorize the gender of voices of the opposite sex more accurately than the voice of their own sex - Junger et al., 2013) and voice attractiveness (men report stronger attraction than women to feminized women's voices and women showed stronger attraction than men to masculinized men's voices - Jones et al., 2010). However, other-sex biases are not observed throughout the spectrum of auditory judgments, and notably in tasks that may not be directly related to mate selection, such as dominance (Jones et al., 2010) or emotions (Schirmer et al., 2007). This suggests that there is no general opposite-sex bias in sensitivity to manipulated voice pitch (Jones et al., 2010). When same-sex processing advantages were found, for instance in recognizing degraded facial identities or facial elements (Cellerino et al., 2004; Yamaguchi et al., 1995), these effects were generally interpreted as a result of familiarity and differential socialization (e.g., boys and girls tend to spend more time with individuals of the same-sex, hence the higher efficiency for recognition of same-sex faces - Cellerino et al., 2004).

The pitch detection paradigm in Study 6 also allowed us to measure participants' response times. In Study 4, we found that response times to pitch changes in other-voice stimuli were faster than on self-voice stimuli, a result consistent with an earlier onset of the MMN to emotional deviants in sequences of other-voice sounds in Study 3. Together with self-voice advantages in response accuracy on the same tasks, these results account for a shallower but faster processing of other-voice stimuli - possibly because these are more socially relevant and warrant faster responses. Study 6 provides additional evidence for this interpretation of response times: there was a significant effect of same-sex stimuli on response times to pitch changes, in which same-sex voices were processed faster than opposite-sex voices (Figure 4.8, Right), and an identical trend, albeit not statistically significant, was found for language familiarity, with faster pitch detection in stimuli of one's own language (Figure 4.6, Right). If differences in response times are to be interpreted as an effect of stimulus relevance, as we did in Study 3 and 4, this overall pattern of result would suggest that first, it is more socially important to process expressive signals of others than the self (Study 3 and 4) and second, that it is also generally more socially relevant to process signals of the in-group (same sex, same language) than the out-group (Study 6). Outside of processes directly involved in mate selection, such as judgments of facial attractiveness, in which faster opposite-sex decisions are generally observed (Duncan and Barrett, 2007; Spreckelmeyer et al., 2013), same-sex stimuli, and notably emotional displays, may indeed have greater relevance, either in terms of biological (e.g., male survival against male aggression) or social relevance (e.g., empathy or closeness). This increased relevance of same-sex stimuli is consistent with previous literature, both in reaction time data (e.g. angry male faces detected significantly more rapidly by male than female observers - Williams and Mattingley, 2006), but also in electrophysiology (e.g. emotional female faces elicited larger P300 than male faces in female participants - Oliver-Rodriguez et al., 1999) and brain imaging (e.g. stronger right amygdala activation for memory of same-sex fearful faces - Armony and Sergerie, 2007) (for a review, see Kret and De Gelder, 2012).

5 Voice transformation software

5.1 Why do we need transformation tools?

In this final chapter, we take a step back from this thesis' main theoretical contribution and reflect on the use of voice transformation software in the studies reported here. Voice transformation tools such as DAVID are a relatively new methodology for the study of voice processing, which - as illustrated by the previous chapters - holds considerable promise, but they also introduce their own methodological questions. We take the time to discuss some of these questions here, before concluding on the whole thesis in the next chapter.

In recent years, important progress has been made in the domain of face perception with the development of several transformation tools. While standardized stimulus databases, such as the NimStim (Tottenham et al., 2009) or the Radboud Faces Database (Langner et al., 2010), are often used in psychology and cognitive science research, tools now exist that can manipulate personality traits (Todorov et al., 2013) and emotional expressions (Roesch et al., 2011) by transforming specific facial cues. For instance, FaceGen (Oosterhof and Todorov, 2008; Todorov et al., 2013) has been used to independently manipulate skin color and facial cues of computer-generized faces to investigate racial categorization in children (Dunham et al., 2016). In another study, Korb and colleagues (2014) studied the facial cues that convey smile authenticity by precisely manipulating facial action units of avatars with FACSGen (Roesch et al., 2011, see Figure 5.1). Moreover, a recent multimodal transformation tool is available to manipulate facial and vocal smiles in video recordings of real persons (Arias et al., 2018b).

Transformation tools have two important advantages over stimulus databases. First, such tools typically allow for parametric control over various characteristics that can in addition be manipulated independently from each other. Second, the original stimulus to which the transformations are applied can be kept constant. This is an important aspect because studies on face perception for instance have shown how different lighting, different camera angles, or different recording equipment can affect facial identity perception and recognition (e.g. Jenkins et al., 2011). Even seemingly slight changes may reduce performance on facial matching tasks. In such cases, transformation tools have an advantage because researchers can depart from a common stimulus to which they apply different transformations, reducing possible interfering effects of variance that may not be of interest for the specific research question. Transformation tools may also provide an advantage over synthesis when the original stimulus is a recording of a real face or voice. Creating naturally appearing stimuli from scratch remains more challenging compared to transforming recordings of faces or voices from real people in which characteristics of the original signal are retained.

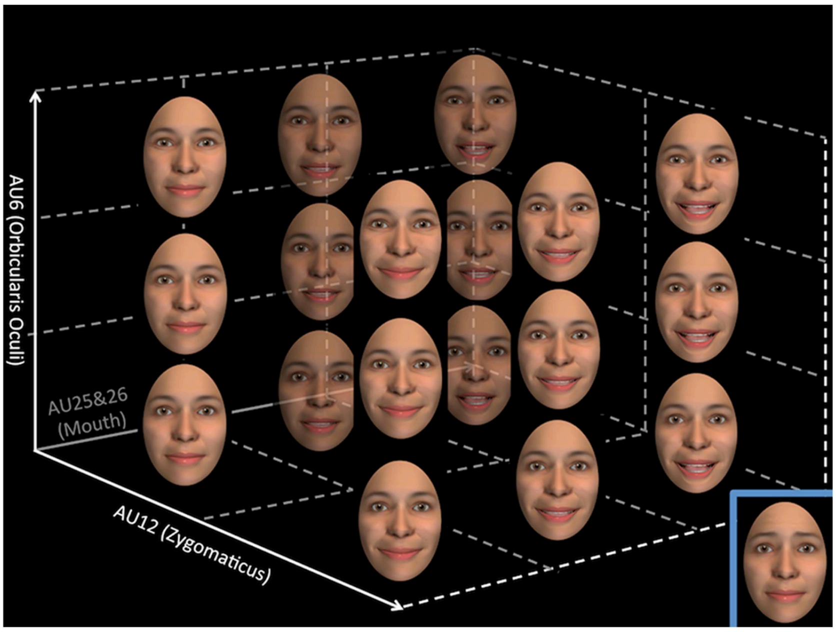


Fig. 5.1: Example of FACSGen where different action units are independently manipulated to investigate judgments of smile authenticity. Figure adapted from Korb et al., 2014

In the context of this thesis, the voice transformations offered by DAVID formed a common methodology in the creation of the stimulus material in all studies. DAVID is designed as a collection of building blocks, or "audio effects", that can be combined in different configurations to create emotion transformations. Each audio effect corresponds to a frequently identified correlate of emotional voices in the literature (see reviews by Scherer, 2003; Juslin and Laukka, 2003; Patel and Scherer, 2013). For instance, fear is often associated with fluctuations in the voice pitch (Laukka et al., 2005; Dromey et al., 2015), an effect that is implemented in DAVID as vibrato (see below). However, individual effects are not associated with individual emotions (e.g. vibrato ↔ fear), because of a large degree of overlap and/or contradicting claims in previous works. For instance, Laukka et al., 2005 observe that a low mean pitch is a correlate of positive valence, but also of negative arousal, casting doubt on what should be associated with a state of joy. Rather, audio effects in DAVID are best described as "things that often happen to one's voice when in an emotional situation". How these effects map to emotions depends on the way the effects are quantified, the way emotions are represented (words, multidimensional scales, etc.), and possibly other factors such as context or culture (Elfenbein and

Ambady, 2002), and elucidating this mapping has not been done in the development of this software.

5.2 Emotional voice transformations in this thesis

Several functionalities of DAVID have been applied in the different studies presented in this thesis. This section will give a brief overview of the results from each study and of the contribution of voice transformations in each specific paradigm. Additionally, some alternative approaches for stimulus creation will be discussed.

In **Study 1**, DAVID was used to create emotional transformations of the self-voice and of unfamiliar voices. Several words produced by each participant were recorded and transformed into a happy and sad version (Figure 5.2A). In an emotion categorization task, participants were better at identifying happy voices when the effects were applied on their own voice than on tranformations of unfamiliar voices. The use of emotional voice transformations allowed us to investigate the specific contribution of the base stimulus (self or other) on the perception and categorization of vocal emotional cues. The use of natural voice recordings would have rendered this assessment difficult as different speakers may use different cues to express emotions. If natural recordings of emotional utterances had been used, it would not have been clear whether this difference in performance was due to the recognition of the specific cues that were used to express the emotion or whether the perception of a particular speaker identity modulated recognition.

In **Study 2**, the same emotional transformations as in Study 1 were used (Figure 5.2A). Instead of investigating the influence of speaker identity on the categorization of the emotional effects, this study was aimed at testing this influence in the opposite direction. Here, we assessed whether characteristics of speaker identity were preserved by the DAVID transformations. The results of this study suggest that identity preservation by DAVID is speaker-specific to some extent. For example, when recordings of utterances produced by the same speaker were presented in pairs with one original version and one happy transformation, the two voices were more often judged as belonging to the same speaker when participants made the judgments on their own voice as opposed to unfamiliar voices. This study provides both an assessment of an additional feature of DAVID that had not been studied in our original validation study (Rachman et al., 2018), as well as a new way of studying implicit processing of the self and the resulting auditory processing advantages.

In **Study 3**, emotional voice transformations were used in an EEG oddball paradigm with the purpose of investigating the interaction effect of speaker identity and emotional cue processing on the mismatch negativity (MMN). Because the MMN response reflects an automatic change detection process that is sensitive to low-level acoustic changes, auditory control of the stimulus is of the utmost importance in oddball paradigms. Similar to Study 1, the aim here was to assess how speaker identity affects the processing of emotional cues. To do so we created two different sequences that were presented to participants, a sequence of the self-voice and a sequence of an unfamiliar voice. One recording per speaker identity (self and other) was presented as the standard stimulus and happy and sad transformations of this neutral utterance were presented as the deviant stimuli (Figure 5.2C). In this design we were able to study the effect of speaker identity on the MMN because the difference of acoustic cues in the standard stimulus and deviant stimuli was precisely controlled. This was an important prerequisite as the MMN is modulated by the differences in acoustics between standard and deviant stimuli. An alternative approach using real recordings of emotional utterances would have introduced more variance in the stimulus acoustics, making it difficult to disentagle effects of speaker identity and effects driven by low-level acoustic differences.

The goal of **Study 4** was to examine the explicit detection of auditory changes. In this behavioral study, only the pitch module of DAVID was used. The results of this study showed that responses to pitch changes on unfamiliar voices were faster than responses to pitch changes on the self-voice. During the experiment, DAVID was used in a more dynamic manner compared to the previous studies in which the voice transformations were applied to offline recordings. In the first part of this study, participants performed an adaptive staircase procedure to estimate detection thresholds for pitch changes in runing speech. For this purpose, a Python module was developed to control DAVID ("pyDAVID") in order to apply pitch changes on a trail-by-trial basis (Figure 5.2B). Besides the magnitude of the pitch, pyDAVID also allows researchers to precisely control the timing of when the effects are prompted. This feature was important to reduce the predictability in the task.

In **Study 5** we aimed to investigate the effect of language-familiarity on vocal emotion recognition. Similar to Study 1, stimuli were created by taking recordings of neutral sentences and transform them to create versions of voices that sounded happy, sad, and afraid (Figure 5.2A). The specific advantage of using voice transformations in cross-cultural emotion research is the control researchers have over the acoustic cues, providing them with the possibility to disentangle effects driven by the culutral specificity of expressing certain emotions and effects due to differences in the perception of emotional expressions. The distinction between these two effects would not have been possible if the stimuli had been recorded by actors or people in general in each respective language. Even when manipulating the same acoustic cues in French and Japanese speech samples, the results of Study 5 show that participants were more accurate in an emotion categorization task in their own language compared to the non-native language.

Study 6 extended the findings of Study 5 by using the same paradigm as Study 4. Here, French and Japanese participants performed a pitch shift detection task on French and Japanese sentences to investigate a language familiarity effect in low-level auditory processing. Similar to Study 4, pyDAVID allowed us to control the magnitude of pitch changes on a trial-by-trial basis. The results of this study show that even low-level auditory processing skills, such as pitch change detection, are susceptible to a language familiarity effect and possibly a same-sex bias as well.

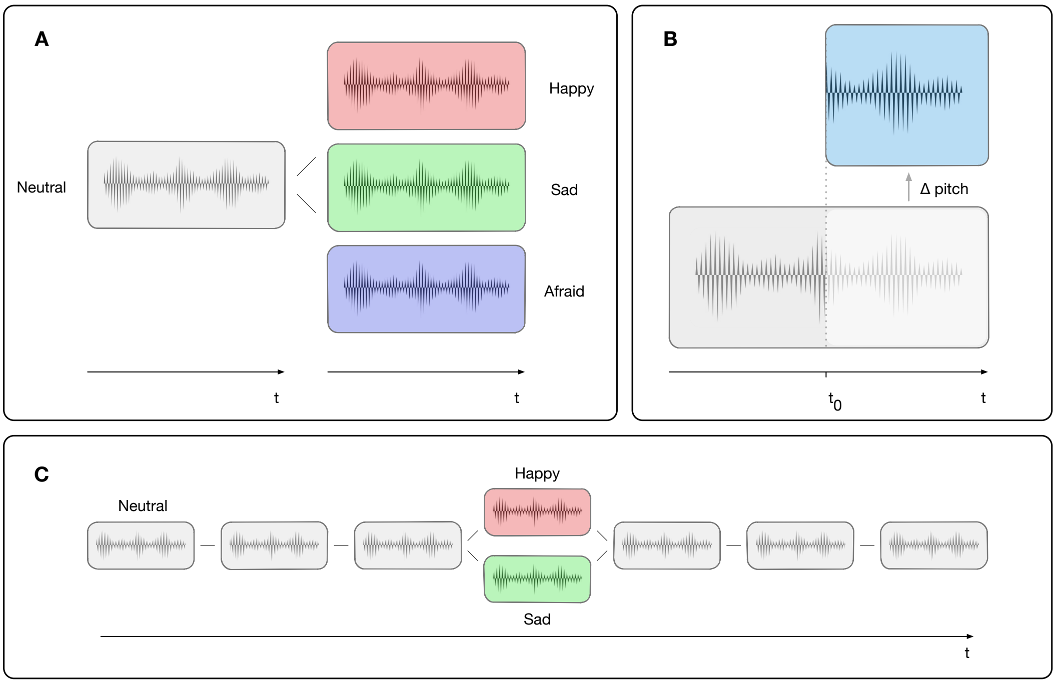


Fig. 5.2: An overview of the several ways voice transformations were used in this thesis. A: Creating several emotional variants of single words or sentences (Study 1,2,5). B: Imposing a sudden pitch shift in the middle of running speech (Study 4,6). C: Creating controlled emotional deviants in a sequence of standard pseudo-words, in order to measure the MMN response (Study 3).

Taken together, the use of voice transformations in the six studies allowed for comparisons between different conditions in which the same voice effects were applied. This way, the influence of speaker identity and language familiarity on emotion perception could be addressed in a manner that would have been challenging, to say the least, if recordings of natural emotional expressions produced by actors or non-actors had been used.

5.3 Validating DAVID

5.3.1 Requirements

The previous section described the advantage of using a voice transformation tool in the various experimental paradigms presented in this thesis. However, in order to test the suitability of this tool in the research domain of psychology and cognitive science, some requirements should be defined against which the software can be tested, comparable to the validation of stimulus databases. This section will therefore discuss the assessment of the following requirements for DAVID:

1. The emotional transformations should be recognizable.

2. The transformations should be natural and should not be perceived as synthetic.

3. The software user should be able to control the emotional intensity of the transformations.

4. The transformations should make sense in several languages, making the tool applicable in multiple research environments, as well as in cross-cultural research approaches.

5. The transformations should preserve speaker identity.

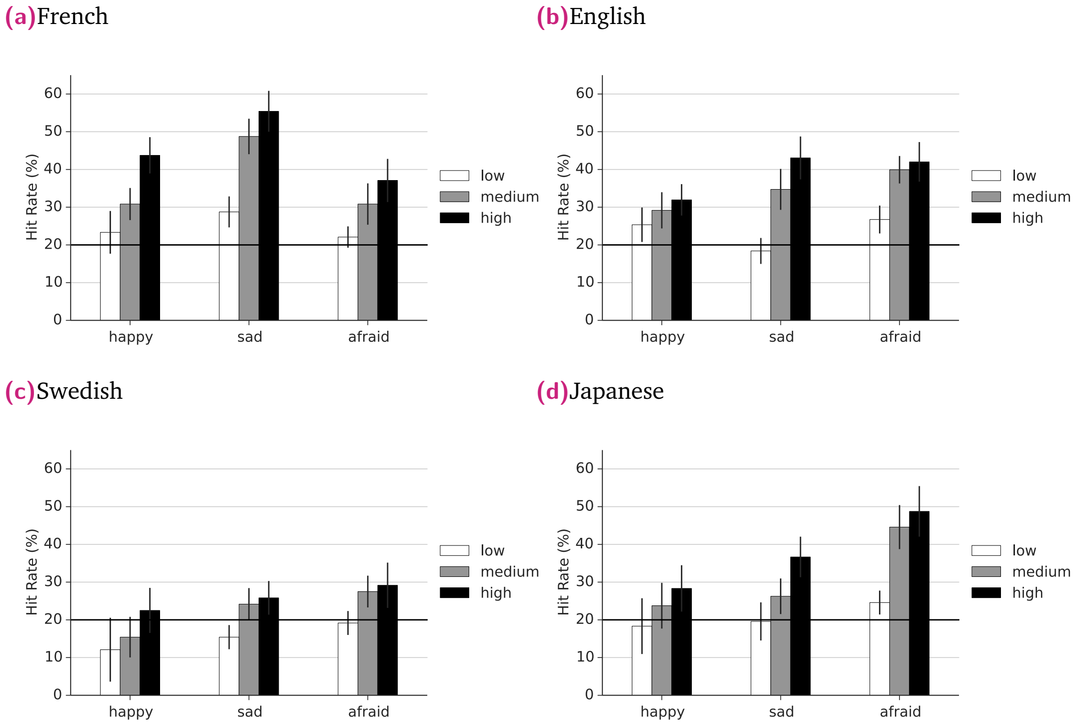
To test the criteria of recognizability, naturalness, and control of intensity, participants performed three consecutive tasks: a naturalness rating task, an emotion recognition task, and an intensity rating task. The validation experiments were conducted in four different languages - French, English, Swedish, and Japanese - to address the requirement of multicultural validity and each participant performed the tasks with the stimuli in his or her own language. Nevertheless, validating a transformation tool such as DAVID is not straightforward and, in addition, I will provide some critical views on the validation experiment that was conducted and discuss alternative ways to assess the abovementioned requirements. The results of the different tasks of the validation study are summarized below, but for more details of the experiments the reader is referred to the publication of the study (Rachman et al., 2018).

Emotion recognition The recognition of the emotional voice transformations was assessed with a forced-choice paradigm using expressions of happiness, sadness, and fear at three different intensities. In each trial, participants listened to two utterances of the same sentence and the same speaker. The first utterance was always the neutral recording and the second utterance was either the same recording unprocessed (neutral condition), or processed with one of the emotional transformations. Participants compared the two utterances in order to indicate in a forced choice task whether the second extract, compared to the first, sounded happy, sad, afraid, neutral. Additionally, a "none of the above" label was included and participants were asked to choose this option whenever they heard a difference that did not fit one of the other response labels (e.g. because the voice did not sound emotional at all, or because it sounded more like another emotion or a mixture of different emotions).

Performance of the participants was indicated in mean accuracy scores (%). To take possible response biases in the recognition task into account, we calculated the unbiased hit rates (H_u) and individual chance proportions (p_c) for each participant (Wagner, 1993). Unbiased hit rates take a value between zero and one and take into account how often an emotion is identified correctly, as well as the total number of times that an emotion label is used. H_u therefore comprises a measure of both the sensitivity and the specificity of each participant's performance.

Raw hit rates for all intensity levels and all languages are shown in Figure 5.3, where chance performance is 20%. Analysis of the unbiased hit rates showed that the three emotional effects were recognited above individual chance levels in all four participant groups (see Rachman et al., 2018 for further details). Overall, the results show that French, English, Swedish and Japanese participants were able to decode these three emotional transformations with accuracies above chance level, with sad (39.4%) and afraid (38.4%) better recognized than happy (33.3%).

The fact that some transformations are more easily recognized as others could always be explained by algorithmic differences, in which one effect could be e.g., a "better-engineered" simulation of human-produced expression than another. However, because the happy and sad transformations largely rely on the same building blocks (pitch shift up or down, high- or low-shelf filter), we find this explanation unsatisfactory, and suggest that this difference is due to cognitive effects that would occur identically with human-produced expressions. It is well documented that, irrespective of language, some emotion displays are recognized more accurately than others, with negative emotions (such as anger or fear) often being more easily recognized than happiness (see e.g. Pell et al., 2009b; Paulmann and Uskul, 2014). It has been argued that recognizing potential danger (such as, here, the afraid transformation) is more adaptive than a non-threatening situation (see Öhman, 2002, in the facial domain), whereas vocally-expressed joy or happiness is especially strongly modulated by culture differences, even within a language group (Juslin and Laukka, 2003).


(a)French

(b)English

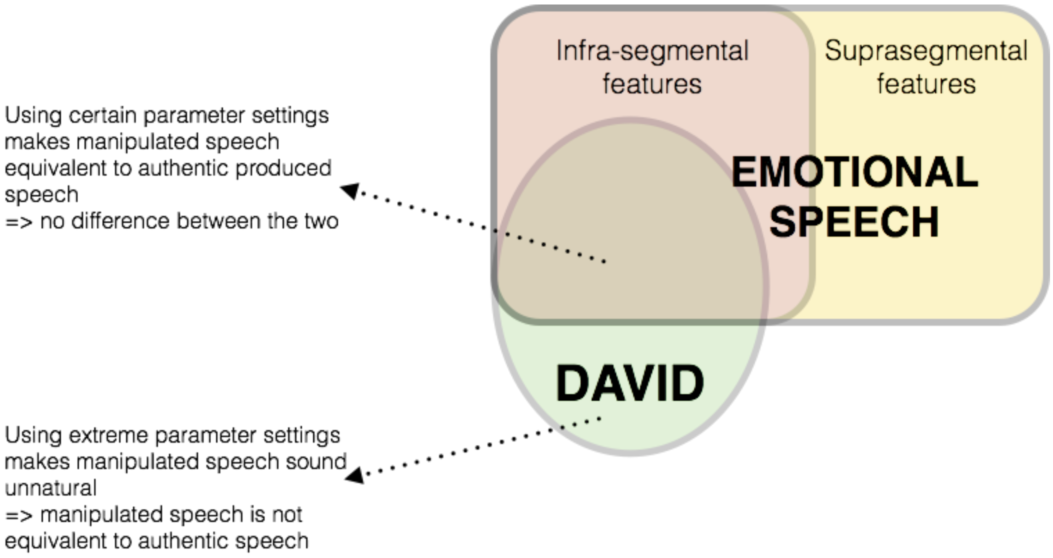
(c)Swedish

(d)Japanese

Fig. 5.3: Emotion recogntion raw hit rates. (a) French, (b) English, (c) Swedish and (d) Japanese raw accuracy scores for three emotions at the nominal level ('high') and two lower intensity levels, error bars represent SEM, black line represents chance level (20%).

While the accuracies obtained here obeyed the same type of pattern, and roughly fell within the range of decoding rates reported in other studies of human-produced speech (see e.g. the meta-study by Juslin and Laukka (2003). *pi*(happy) = 0.66 (this study, all languages averaged) vs 0.51-1.0 (Juslin & Laukka); *pi*(sad) = 0.71 vs 0.80-1.0; *pi*(afraid) = 0.70 vs 0.65-1.0), they were still relatively low compared to typical performance (e.g. the mean hit rates reported in Scherer et al. (2011), H(happy) = 54%, H(sad) = 69%, H(afraid) = 62.4%). Moreover, neutral (unmodified) expressions were labeled correctly more often than any of the transformed emotions.

Several factors may explain these results. First, the difference between emotion recognition accuracies in this study and other studies using acted emotional expressions are likely a consequence of the tool's operating on only infra-segmental speech features (and not e.g. on speech rate and pitch contour, see Figure 5.4). The emotional tone of the transformed voices is therefore more subtle – and expressed with a more restricted set of cues – than acted emotional expressions. It is therefore in line with expectation that, by manipulating only a subset of the acoustic markers involved in the vocal expression of emotions, the decoding accuracy should be reduced and biased towards the neutral label.



Using certain parameter settings makes manipulated speech equivalent to authentic produced speech
=> no difference between the two

Using extreme parameter settings makes manipulated speech sound unnatural
=> manipulated speech is not equivalent to authentic speech

Fig. 5.4: What can DAVID do?

Second, the use of forced-choice paradigms in emotion research has been subject to criticism. A commonly mentioned weakness of this design is that it may bias performance because of the limited response options. As such, recognition may be overestimated in comparison to spontaneous labeling of emotions (for further discussion on this topic see e.g., Banse and Scherer, 1996; Scherer et al., 2003). However, the decision to use a forced-choice test was motivated by the objective to compare results across several languages. The response option *"None of the above"* was added to avoid forcing participants too much towards a certain emotion label. Additionally, results were analyzed as unbiased hit rates to control for possible asymmetries between response categories.

Third, the data of all languages show a confusion between the afraid and sad labels, where an afraid voice is often identified as a sad one. Because the vibrato effect is a particularly salient component of the afraid transformation, we could speculate that this may have been perceived as a trembling voice of someone who is on the verge of crying, which would explain the frequent use of the sad label. This confusion between "cold" and "hot" sadness (low or high arousal) has in particular already been noted in the Japanese language (Takeda et al., 2013), and could explain parts of our results.

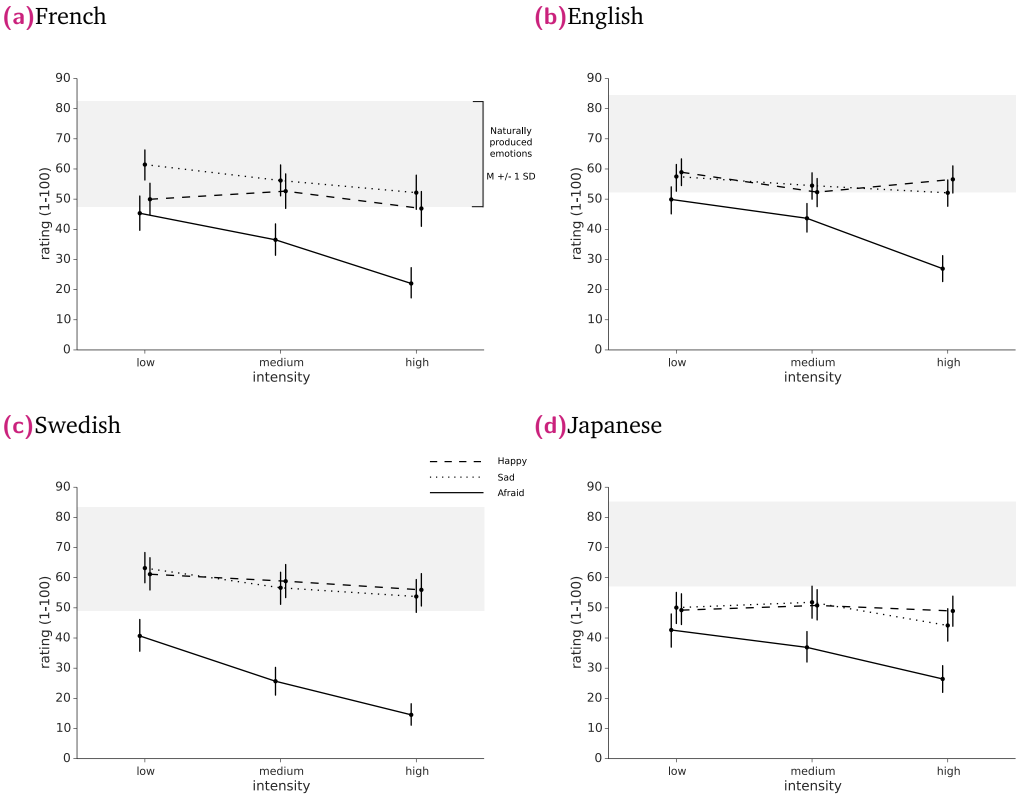
Fourth, the high performance for neutral utterances is likely due to both the subtlety of the emotional effects and the fact that each trial comprised a neutral reference voice. As a result the response strategy is slightly different for the neutral vocalizations, which would involve reporting the absence of any auditory transformation. Conversely, when a transformation is perceived, the next decision to be made is then more subtle because the appropriate label for the transformation should be chosen out of four options. We would argue that this could lead to a decrease in performance. Furthermore, the use of a neutral reference voice brings up another issue worthy of discussion, because studies using a similar paradigm (i.e., comparing a neutral voice to a pitch-shifted voice) found that pitch influences the perception of physical dominance and traits such as leadership capacity and trustworthiness for example (Klofstad et al., 2012; Klofstad et al., 2015; Vukovic et al., 2011). Because some of the emotional transformations in DAVID also use pitch-shifting, we cannot be certain that these acoustic manipulations are exclusively changing the emotional tone of the voice. So even though the instructions in this experiment involved a mapping of acoustic features onto emotions, we cannot rule out that participants perceived differences in personality traits or physical characteristics as well.

Finally, we cannot exclude the possibility that the semantic content has influenced the recognizability of the emotional transformations. In this study we included only semantically neutral sentences because we wanted to use the same sentence for each of the emotional transformation, trying to avoid a bias towards a certain emotion caused by the content of the phrase. However, it could be that a neutral sentence such as *"the plane is almost full"* does create a mismatch when pronounced with a voice that sounds afraid. Indeed, it has been shown that a mismatch between sentence content and the voice quality (e.g., negative sentence content and a voice quality expressing positive valence) can render utterances to be perceived as emotionally ambiguous (Aylett et al., 2013).

Naturalness An important requirement for any kind of transformation tool used in experimental psychology is that the manipulation should appear realistic. To test this requirement in DAVID participants listened to both human-produced emotional utterances and recordings of neutral sentences modified by DAVID and rated the naturalness of the voice on a continuous scale anchored by "very artificial/not at all natural" and "not at all artificial/very natural" (1-100). Mean naturalness ratings for each emotional transformation at the three intensity levels are presented in Figure 5.5a for all four languages, compared to ratings of human-produced voices. The results of this experimental task showed that naturalness ratings increased with increasing parameter settings for each emotion.

It is however dificult to determine when DAVID has "passed the test" of naturalness. The degree of naturalness required in any given study is likely dependent on the experimental setting. For example, lower levels of naturalness may be acceptable in cases where participants are likely to attribute strange or unnatural characteristics in the sound to recording equipment instead of the emotional expression of the voice.

In an effort to position the naturalness ratings of the transformed speech samples to those of natural speech, we presented the effect sizes and the probability of inferiority


(a)French

(b)English

(c)Swedish

(d)Japanese

Fig. 5.5: Naturalness ratings. (a) French, (b) English, (c) Swedish and (d) Japanese naturalness ratings for three emotions at three intensity levels compared to unmodified voices (grey: mean ± 1 SD), error bars represent 95% confidence intervals.

for each emotional transformation compared to the three human-produced emotions grouped together in Table 5.1. The probability of inferiority (POI) is calculated by subtracting the common language effect size statistic (McGraw and Wong, 1992) from 100% and it represents the probability that an utterance drawn at random from the set of transformed voices has a higher naturalness rating than an utterance drawn at random from the set of human-produced emotional voices.

At the nominal level, the mean natural ratings were 46.9 for happy, 52.2 for sad, and 22.0 for afraid, with 95% confidence intervals [39.5, 54.3], [46.5, 57.9], and [15.2, 28.8], respectively (in the French language, see Figure 5.5a for complete results). The mean naturalness rating of the sad transformation fell within one standard deviation of the mean naturalness rating for the human-produced emotions (M=64.9, SD=17.5), meaning that POI=27.4% of the human-produced stimuli were judged less natural than the effect (at nominal level). The mean rating for happy fell just outside of this range, with POI=22.2% at nominal level. The afraid effect was rated as least natural (mean=22.0, POI=3%).

Regarding the paradigm and data of the validition study several remarks should be made. While the effects were generally rated as less natural than human-produced

Tab. 5.1: Cohen's *d* and probability of inferiority (POI) of the naturalness ratings for each emotional transformation compared to natural emotional voices.

		French		English		Swedish		Japanese	
		Cohen's *d*	POI *(%)*	Cohen's *d*	POI *(%)*	Cohen's *d*	POI *(%)*	Cohen's *d*	POI *(%)*
Happy	low	0.86	27.6	0.54	35.1	0.29	41.9	1.51	14.3
	med	0.81	28.3	0.92	25.8	0.45	37.5	1.40	16.1
	high	1.08	22.2	0.75	29.8	0.64	32.5	1.42	15.8
Sad	low	0.21	44.1	0.67	31.8	0.18	44.9	1.44	15.4
	med	0.57	34.3	0.77	29.3	0.58	34.1	1.30	17.9
	high	0.85	27.4	1.04	23.1	0.78	29.1	1.69	11.6
Afraid	low	1.20	19.8	1.18	20.2	1.58	13.2	1.68	11.7
	med	1.71	11.3	1.43	15.6	2.77	2.5	2.31	5.1
	high	2.66	3.0	2.43	4.3	3.82	0.3	2.87	2.1

speech, naturalness ratings for happy and sad still fell within one standard deviation of the mean ratings for authentic speech, with one fourth to one third of our human-produced stimuli being rated as *less* natural than our effects. Moreover, naturalness ratings of these two emotions did not differ significantly across the four different languages and across the three intensity levels. Naturalness for the afraid effect was more problematic, and behaved like happy and sad only at the weakest intensity levels. In all four languages, stronger intensity levels significantly lowered the naturalness ratings of the afraid effect.

The interpretation of these results deserves caution. First, the purpose of this task was not to test for people's maximum capacity to recognize the manipulation, but to assess typical performance. In our view, there is always a situation where DAVID will fail. For instance, when one can compare an original recording with several of its transformations, it would not be hard to notice that transformations are bound to be the outcome of the system when the original prosody is exactly reproduced. What these data show is that, at least in some situations, some natural sentences will be judged as equally or less natural than the transformations produced by DAVID. In our experience, the acceptance of DAVID-transformed speech as authentic cases of emotional speech is heavily dependent on context. For instance, in a recent study of vocal feedback where participants were instructed to read a text out loud while the effect was gradually increased without their knowing, only 14% of the participants reported detecting an artifact with their voice (Aucouturier et al., 2016). In contrast, had participants been instructed before the experiment about a potential voice manipulation, it is likely that this proportion would have been larger.

Second, it should be noted that the naturalness ratings of human-produced voices are not concentrated around the high end of the scale, showing that even authentic speech can be judged as not very natural. The relatively low ratings of human-produced voices in our study are likely due to the fact that participants were informed that some of the presented voices were computer manipulated. While it could be argued that such framing artificially reduced the baseline to which we compare the transformations, we believe it would be very difficult to elicit reliable responses without explicit instructions of what we mean by naturalness (i.e. here "not artificially manipulated"). Indeed, judging a recording as "natural" could be construed alternatively as evaluating the authenticity of the emotion ("is it sincere or simulated"), the match between an emotion and the sentence's verbal content ("is it natural to be happy about an alarm clock"), or a rating of how well-acted the emotion was. Besides, it is not clear why such a paradigm should not also reduce the naturalness ratings of the manipulated recordings themselves.

Finally, the low naturalness scores of the afraid transformation at high intensity deserves a special mention. It is possible that this is a consequence of the vibrato effect used in this transformation, which may have provided a salient cue when compared to non-manipulated voices, either because it was too systematic or because it created occasional artefacts. It is to be noted that, in an alternative A/B testing paradigm reported by Aucouturier et al. (2016), the same effect was not discriminated from human-produced speech above chance-level. Rather than arguing whether one specific effect is “natural” or not, the hope is that, by presenting effect sizes and probabilities of inferiority in each configuration, each potential user can judge for themselves whether the tool is useful and appropriate for their own research.

Intensity Another important advantage of transformation tools is the flexibility it provides to manipulate different stimulus features independently and parametrically. To test whether intensity differences of the emotional presets of DAVID are perceived, participants performed a naturalness rating task when listening to modified or a human-produced emotional voices. Participants indicated the emotional intensity of the stimuli on a continuous rating scale (1-100) anchored by “not at all happy/sad/afraid” and “very happy/sad/afraid”. In addition, participants rated the loudness (subjective sound intensity) of the utterance to avoid confusion between the emotional intensity and other perceived acoustic characteristics that are not necessarily related to the intensity of the emotion. Loudness ratings were not further analyzed here.

The mean intensity ratings (Figure 5.6) show that, irrespective of language, both the angry and sad transformations were rated as more intense as parameter levels were increased. On the other hand, the intensity of the happy transformation did not seem to change for different parameter levels, for neither language. More generally, all transformations show a clear inverse relation between naturalness and intensity (the more intense, the less acceptable as authentic speech), and the choice of one particular configuration should follow which of these two factors is most important in each experimental context.

The lack of change for the happy effect is interesting as the different intensity levels do change recognition rates: it appears that, as we increase the depth of the pitch change and the amount of high frequencies in voice, transformed speech becomes more recognizably, but not more strongly, happy. This is especially surprising as this seems to hold in all four languages tested here, and the same effect does not seem to occur in the sad transformation, which yet uses symmetrical manipulations of pitch and spectral shape. Human actors have notorious difficulty manipulating the intensity of their emotional expression without a confounding variation of acoustic intensity or pitch (Juslin and Laukka, 2001; Ethofer et al., 2006). Consequently, the psychoacoustics of emotional intensity (e.g., what makes a happy voice happier) is still unknown to a large degree. It would be interesting, with DAVID, to selectively manipulate the acoustical depth of various parameters (e.g. pitch shift independently

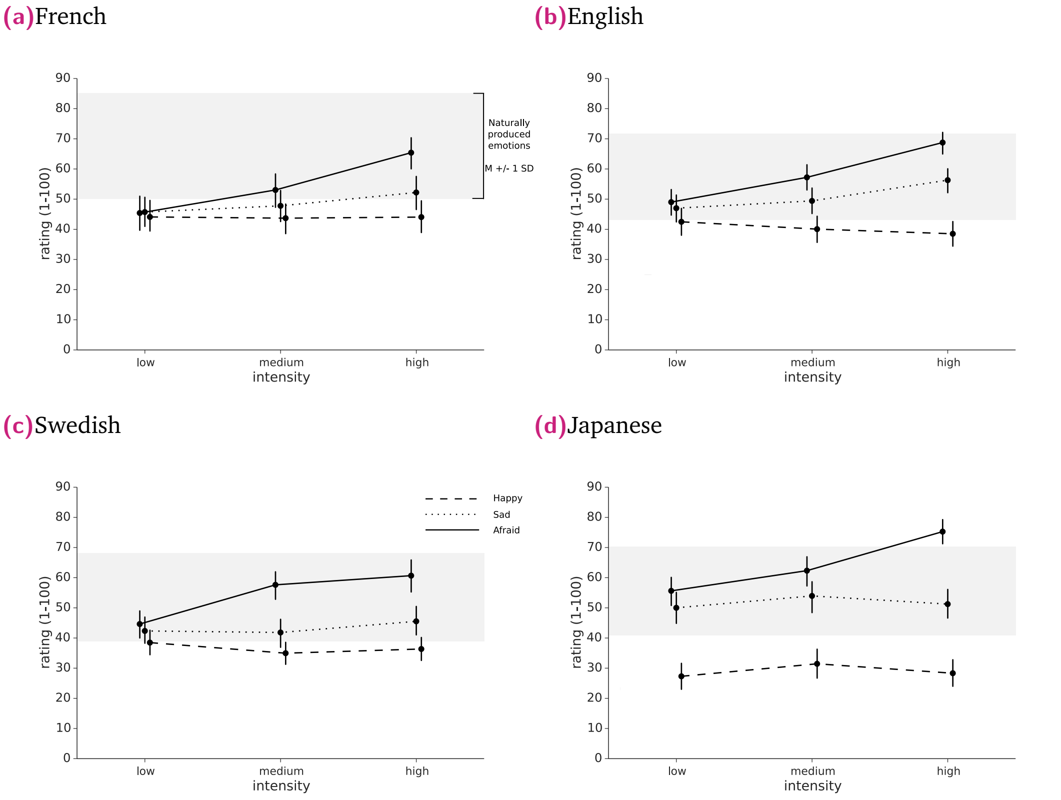

(a)French (b)English (c)Swedish (d)Japanese

Fig. 5.6: Intensity ratings. (a) French, (b) English, (c) Swedish and (d) Japanese intensity ratings for three emotions at three intensity levels compared to unmodified voices (grey: mean ± 1 SD), error bars represent 95% confidence intervals.

from RMS energy), and examine how these parameter changes influence perceived emotional intensity.

One methodological limitation in this task is the fact that sound levels were normalized across stimuli so that the stimuli were perceived with the same loudness for each intensity level and across the whole experiment. Previous studies have shown that loudness is an important cue of arousal in speech and nonverbal vocalizations (e.g. Lima et al., 2013; Juslin and Laukka, 2001) and it is likely that changing this parameter would have an effect on the intensity ratings.

Taken together, these results warrant further investigation of the respective contribution of different acoustical characteristics to emotional intensity. One conservative conclusion is that the tool does not appear ideally suited to controlling the emotional intensity of happy vocalizations, in its current form. As such, the studies presented in this thesis did not use different intensity levels for the voice transformations.

Speaker identity preservation Initial validation of DAVID did not include any specific task to asses the preservation of speaker-specific characteristics. However, several studies in this thesis addressed the issue of speaker identity preservation. Regarding the recognition of self-produced voice recordings, the voice transformations seem to affect self-voice recognition to a certain extent (see Studies 1-3). While accuracy was typically lower for transformed self-voices compared to neutral self-voices, performance remained above chance-level in several cases. In the identity matching task (Study 2), pairs of voices are often correctly attributed to the same speaker when one of the voice samples transformed with the happy effect, but not with the sad effect. In this study, contrary to the self/other discrimination experiment, samples of the self-voice were more often correctly attributed to the same speaker compared to voice samples of an unknown speaker. These results illustrate that, similar to the assessment of naturalness, validating identity preservation requirements of a software tool is challenging as it may be modulated by various factors, such as familiarity or task demands. As such, rather than validating strict requirements of identity preservation, a more interesting use of DAVID is perhaps the control it offers on different acoustic parameters in order to address new questions. In the examples offered here, DAVID has been used to control within-speaker variability to investigate how familiarity influences emotional voice processing and vice versa.

Intercultural applicability Intercultural differences in the perception of vocal emotions, and emotional expression in general, are widely documented (for a review see e.g. Elfenbein and Ambady, 2002; Scherer et al., 2011). The first three tasks presented here, conducted in four languages, depart a little from standard paradigms in that they are neither a test of cross-cultural universality, because the stimuli used in the four participant groups are not the same (e.g. Biehl et al., 1997), nor a test of intercultural differences (e.g. Elfenbein and Ambady, 2002), because both speakers and decoders belong to the same cultural group. What these results address is the cross-cultural validity of the acoustic cues on which DAVID operates: participants in each cultural group listened to voices produced in their own language, albeit transformed with a unique algorithm applied identically to all languages.

These results, like most previous studies, point at the co-existence of both universal cues of emotional expression and culturally-learned display rules. On the one hand, the three emotional transformations were recognized above chance levels in all languages. On the other hand, language had an influence on performance in all of the three tasks. In the recognition task, Swedish participants scored lower than French and Japanese participants, irrespective of emotion. In the naturalness task, ratings for afraid were lower in the Swedish population than in the English. Finally, in the intensity task, happy was rated as less intense in Japan compared to all the other languages. Swedish intensity ratings of happy were also lower than French.

The fact that the same transformations were decoded above chance in four languages shows that the emotional cues manipulated in DAVID are not purely cultural. This may be a blessing of having to rely only on infra-segmental cues (for real-time constraints) and not manipulating supra-segmental aspects of speech such as intonation contours and rhythm, which Schröder (2001) have found can vary across language families and be difficult for outsiders to decode. Manipulating only pitch and spectral shape as we do here, if arguably accountable for relatively low recognition rates, at least appears to be a cross-culturally valid way to simulate emotional expression.

The amount of cross-cultural differences seen in the data here in both recognition hit rates and intensity ratings is typical of other cross-cultural decoding studies with human-produced stimuli. Even on the same stimuli, different cultures perform differently and give different levels of intensity: e.g. in Matsumoto and Ekman (1989), Americans gave higher absolute intensity ratings on facial expressions of happiness, anger, sadness and surprise than Japanese; in Biehl et al. (1997), non-western cultures gave higher intensity for fear, western cultures gave higher intensity for happy, and Japanese were worse in recognition of fear and sadness. Cross-cultural ratings of the perceived intensity of our transformations appear consistent with this pattern, with Japanese participants giving higher intensity for the afraid transformation, and English, French and Swedish participants giving higher intensity for the happy transformation.

Several factors may explain such differences in the agreement and intensity levels across cultures. First, the display rules of a given culture shape its members' mental representations of emotions, such as the intensity level of emotional prototypes (Engelmann and Pogosyan, 2013) or the accuracy of their decoding (Biehl et al., 1997). For instance, it is possible that lower intensity levels for fear and higher intensity levels for happiness are the cultural norm in Japan (which some have indeed argued, see e.g. Kitayama et al., 2000; Kitayama et al., 2006) and therefore that a given amount of expressivity (i.e., given parameter values) for these two emotions is judged, respectively, as higher and lower intensity by Japanese participants than by English, French or Swedish decoders.

Second, different cultures may have different cognitive strategies for judging the same emotion. For instance, when asked to judge the intensity of an emotion, Americans were found to rate the intensity of the external display of affect, while Japanese rated their perceived intensity of the subjective experience of the poser (Matsumoto, 1999). Because the scale used in the intensity tasks confounded both constructs, it is possible that different cultures have in fact rated different aspects of the same stimuli, rather than differed in their rating of a common aspect.

Third, difference in the level of agreement across cultures may also be explained by the translation of terms used as response categories (happy: joyeux, glad, yorokobi; sad: triste, ledsen, kanashimi; afraid: inquiet, rädd, osore. Even when terms are derived through back-translation, they may not be equivalent to the original, and in particular may not refer to the same facial or vocal expression. For example, shame and its common translation into Spanish (vergüenza), sadness and its common translation into Arabic (huzn), do not refer to emotions with identical features (de Mendoza et al., 2010; Kayyal and Russell, 2012). In the present data, Swedish participants were overall less accurate than French and Japanese participants, and notably mistook an afraid voice for a sad one more often than Japanese participants did. It is possible that these differences result from different boundaries between terms, and that the cues manipulated in the afraid effect spanned a larger proportion of the vocal expressions referred to as *ledsen* in Swedish than that referred to as “sad” or *triste* in French.

Finally, it remains possible that, while the cues manipulated in the transformations are cross-culturally valid, their algorithmic manipulation differed in salience depending on the phonetic characteristics of the language on which it is applied. Effects like vibrato for instance rely on the availability of relatively long voiced portions in phonemes (e.g., 250 ms for 2 cycles of a 8 Hz-vibrato to be perceived), and may not operate well on languages with a relatively large consonant/vowel ratio such as Swedish (Haspelmath, 2005). Similarly, inflection added with the happy or afraid transformations may be more salient in languages that display comparatively flatter prosodic variations such as French. More research will be needed to understand how acoustic parameters that are optimal for emotion recognition depend on the phonetic characteristics of a given language, or even a given speaker. Until then, users are encouraged to experiment with parameters beyond the values in the validation study, and to adapt them to the specificities of their experimental context.

5.3.2 A note on applicability to other stimuli

DAVID has been developed to transform a neutral voice into an emotional voice and not to transform an already emotional vocal expression into another type of emotion. While it would be interesting and of great use to transform a sad voice into a happy voice, this possibility was not addressed during the development of this tool and the execution of the validation study. Notably, because DAVID does not operate on supra-segmental cues, it is possible that sad voices made happier with DAVID present a conflicting mix of cues - with high pitch and increased high frequencies in the short-term but slow speech rate and decreasing prosody in the long term, and may not impart the intended emotion successfully.

Moreover, while the DAVID transformations have initially been developed to be used on speech, they have now also been applied to both singing voices and instrumental music (Bedoya, 2018). In this experiment, different acoustic cues (among which pitch and vibrato effects of DAVID) were manipulated in both speech and music to compare valence and arousal ratings in the various conditions, allowing to test the hypothesis by Juslin and Laukka (2008) that emotional responses to music are largely driven by their similarities to expressive speech. For instance, acoustic cues that convey sadness in speech should convey that same emotion when perceived in music. The validation of this hypothesis has been challenging and support comes mainly from correlational data (Ilie and Thompson, 2006; Juslin and Laukka, 2003). A software tool like DAVID however, makes it possible to test this hypothesis in a causal manner. Indeed, in accordance with Juslin and Laukka's hypothesis, Bedoya (2018) found that various static and dynamic acoustic cues of emotional expression do in fact translate from voice to music stimuli (see Figure 5.7).

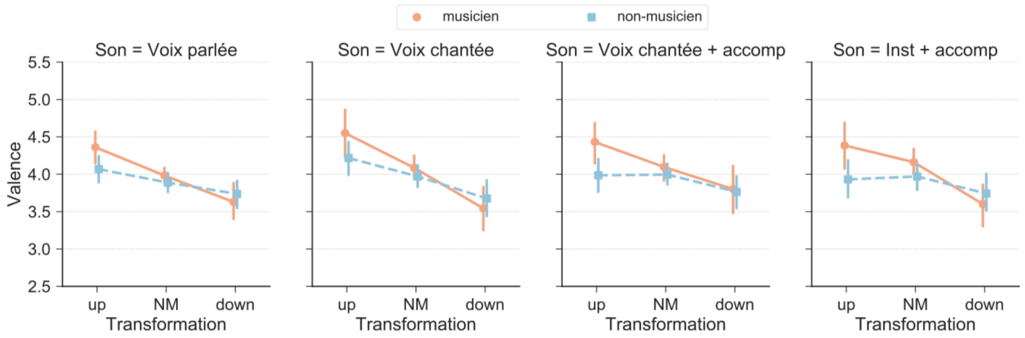


Fig. 5.7: DAVID transformations applied to music. Example of valence ratings of the DAVID pitch effect applied in four different conditions for musicians (orange) and non-musicians (blue), from left to right: speaking voice, singing voice, singing voice with background, and instrumental music with background. Error bars represent 95% confidence intervals. Figure adapted from Bedoya (2018).

5.3.3 Lessons from DAVID and the limitations of transformation tools

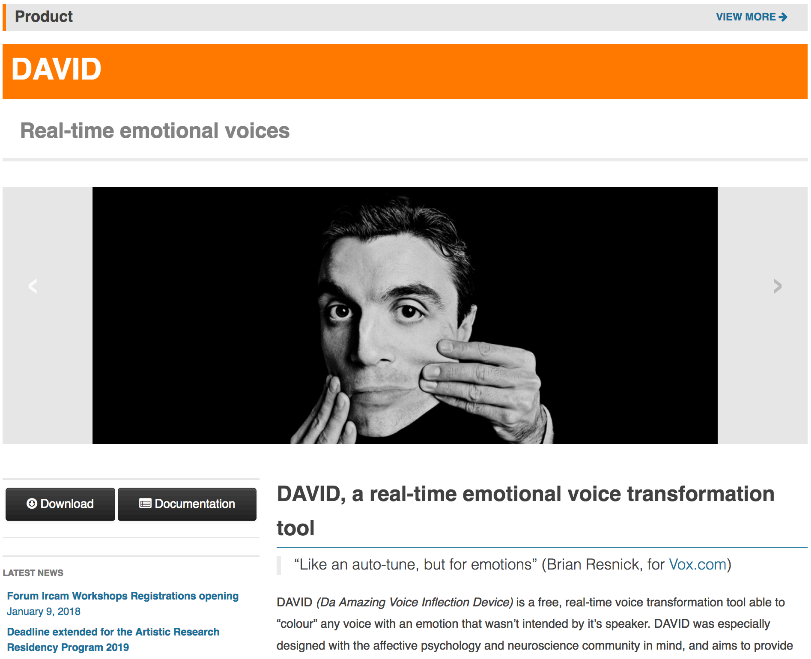
Taking together the results from the DAVID validation study, it becomes clear that validating transformation tools is not always straightforward and is often dependent on the context in which the software is going to be used. The main objective when developing DAVID has been to transform continuous, running speech and the validation study was conducted in the context we believed the tool to be most useful in. As a result, one should be conscious that certain parameter settings that are acceptable in continuous speech may have a more pronounced effect in unconventional speech production, for example during the production of sustained vowels. Users are therefore encouraged to experiment with parameters beyond the values suggested here, and to adapt them to the type of stimuli used in their work.

Perhaps in addition to validation experiments of DAVID, the more relevant contributions for the domain of experimental psychology and cognitive science are examples of research paradigms in which the tool can be used. When the possibilites and limitations of available tools are clearly defined, researchers can decide accordingly on the appropriateness of available tools. In the case of DAVID, one important characteristic is the short latency of the transformations. This prevents researchers from manipulating speech rhythm or speech rate, making the tool inappropriate to address this parameter in emotional speech perception, but it makes the tool particularly suitable for emotional feedback paradigms. Similarly, DAVID does not manipulate suprasegmental speech features which form an important part of emotional speech (Figure 5.4). This is also likely an important reason why overall accuracy is on the low side in Study 5. However, the parameters used in DAVID are arguably closely linked to physiological changes that are related to emotional voice production and less susceptible to culture-specific influences. Because of this, we judged the tool pertinent to use it to investigate these cues in the context of cross-cultural emotion recogntion. Overall, it is unlikely that one single tool would answer to all requirements in experimental psychology and emotion research. With more tools being developed, they can hopefully supplement each other as well as supplement other existing methodological approaches.

As a final note on the lessons learned from using DAVID, rather than being put down by the software's limitations, a more interesting and constructive approach would be to think about the questions that it raises when requirements are not completely fulfilled and to use the tool to address them. This thesis is to a certain extent shaped by this approach. It would have been too harsh to write off DAVID arguing that it does not preserve speaker identity based on the findings of Study 1, showing that the transformations lead to less accurate discrimination of self-produced voice recordings. Instead, DAVID has been applied in subsequent studies to investigate different conditions under which speaker identity may or may not be affected by the transformations, allowing us to address the influence of familiarity in emotional voice perception. As such, in some cases, "limitations" of experimental tools may just offer an entry point to address new questions.

5.4 A note on open science

In recent years, initiatives promoting open science have gained momentum. The availability of open-source software is an important factor in facilitating replication studies. DAVID is implemented as an open-source patch in the Max environment (Cycling74), a programming software developed for music and multimedia. The DAVID software, the pyDAVID module to control parameters via Python, and accompanying documentation can be downloaded under the MIT license from `http://forumnet.ircam.fr`. Forumnet can be accessed upon creating a user account, after which DAVID can be downloaded for free (Figure 5.8). Using DAVID first requires to install the Max environment, which is provided in free versions for Windows and Mac systems. DAVID comes with the parameter presets used in the validation studies described below, but users also have full control over the values of each audio effect to create their own transformations and store them as presets for future use. Updates and new functionalities will be made available on IRCAM Forumnet as well. In addition, all sound samples used in the validation study were made available for download from `https://archive.org/details/DAVIDAudioFiles`.



DAVID

Real-time emotional voices

LATEST NEWS

Forum Ircam Workshops Registrations opening
January 9, 2018

Deadline extended for the Artistic Research Residency Program 2019

DAVID, a real-time emotional voice transformation tool

> “Like an auto-tune, but for emotions” (Brian Resnick, for Vox.com)

DAVID *(Da Amazing Voice Inflection Device)* is a free, real-time voice transformation tool able to “colour” any voice with an emotion that wasn’t intended by it’s speaker. DAVID was especially designed with the affective psychology and neuroscience community in mind, and aims to provide

Fig. 5.8: DAVID software available on http://forumnet.ircam.fr.

Several research groups, such as those of Robert Zatorre (Montreal), Boris Kleber and Elvira Brattico (Aarhus), Marcel Zentner (Innsbruck), Petter Johansson (Lund), Katsumi Watanabe and Yuko Yotsumoto (Tokyo) and Tanzeem Choudhury (Ithaca, NY), have now started to use DAVID in different experimental settings. The Python module of DAVID makes the tool suitable to study emotional voice effects in social interaction paradigms by allowing trial-by-trial parameter control. In one line of research in Japan, the approachable user interface allowed for participants to manipulate the different parameters themselves to calibrate stimulus characteristics on an individual and personal level. For instance, Kimura and colleagues (2018) asked participants to manipulate recurdings of their own voices to create a stimulus that sounded as similar as possible to the way they would hear their own voice in daily life. The availability of DAVID as open-source software also facilitates the points discussed in Section 5.3.3 suggesting that a useful way to validate the software is to use it in different experimental settings to explore its possibilities and boundaries. In another study, Costa et al. (2018) used DAVID to present participants with voice feedback with a calmer tone during relationship conflicts, and found that they felt less anxious.

5.5 Future perspectives

Both the results presented in this thesis and methodological aspects of using voice transformations give rise several questions that may be addressed in future studies. Several of these future directions are pointed out below.

fMRI of voice effects In Study 3, an effort was made to investigate the neural sources driving the differences in MMN onset latency upon the processing of emotional cues on the self-voice and the other-voice. While EEG provides an excellent means to investigate the temporal dynamics of emotional voice processing, source reconstruction remains challenging and does not provide conclusive evidence of underlying mechanisms. For instance, an fMRI study of the processing of the transformations on different speaker identites should be conducted in order to verify whether the left-lateralized somatomotor areas found in the MMN study may indeed be driving factors in the shorter MMN onset latency found in unfamiliar voices.

Speech production The first four studies in this thesis use recordings of the self-voice as stimuli. In daily life however, people hear their own voice while speaking. A relevant direction for future research would be to study the processes involved in emotional self-voice perception in the context of speech production as well. During speech production, the auditory system differentiates internally generated sounds from externally generated sounds by comparing an efference copy (an internal signal), which is generated by the motor command and predicts its sensory outcome, with the sensory feedback (an auditory signal) (Heinks-Maldonado et al., 2006). When the sensory feedback does not match the expected feedback, the auditory signal is attributed to an external source. A match between the efference copy and the auditory feedback is thought to be reflected in a suppression of activity in the auditory cortex (e.g. Greenlee et al., 2011), while a mismatch has been shown to cancel this suppression (e.g. Behroozmand et al., 2011). A study by Niziolek et al. (2013) suggests that the auditory cortical suppression that is seen during speech production might be related to sensory goals, rather than predictions, as less cortical suppression was seen when speech production was less prototypical. With the short latency of DAVID, it woud be possible to manipulate the auditory feedback during speech production, in order to further investigate to what extent sensory goals and expectations affect processes involved in speech production in the context of emotional speech.

Problems with self-voice recordings A common methodological issue in studies of the self-voice is the strange feeling of hearing recordings of one's own voice. As briefly discussed in Chapter 2, Box 2.1, perception of the self-voice in daily life results from combining the auditory signal that is transmitted through the air, which follows the same pathway when hearing one's own voice and hearing the voice of another speaker, and the signal transmitted through bone conduction, which is specific to hearing the self-voice. The question of how to best present self-voice recordings in such a way that listeners have the same experience hearing recordings of their voice through headphones as they have when hearing their own voice in real-life settings, while interesting, is orthogonal to the ones addressed in this thesis. Still, tackling this issue is of practical value and can refine experimental paradigms and findings on the specific role of the self-voice.

To this date, and despite various efforts (e.g. Kimura and Yotsumoto, 2018; Won et al., 2014), no single sollution for this problem with self-voice recordings has been found. One possible strategy to find a suitable correction of self-voice recordings is to use a more data-driven approach. Recently, another voice transformation tool that has been developed in our team has been used to identify mental representations of prosodic patterns of personality traits, such as dominance and trustworthiness (Ponsot et al., 2018a), and spectral cues conveying perceived smiles in vocal signals (Ponsot et al., 2018b). By presenting a multitude of voice stimuli to which random variations of a certain parameter (e.g. pitch or timbre changes) are applied, reverse-correlation techniques can be employed to reconstruct the specific parameters that underlie the mental representation of the voice characteristic being studied. It would be interesting to use this paradigm in the context of self-voice perception. However, the study of Kimura and Yotsumoto (2018) suggests that interindividual differences are to be expected. In that case, since the paradigm used by Ponsot et al. (2018a) is rather time-consuming, it might not be the most efficient way to optimize correction strategies on an individual basis in studies of self-voice perception. Nevertheless, this approach might offer valuable additional insights in the relevant auditory characteristics specific to self-voice perception.

6 Conclusion

Emotion perception forms an important part of our social interactions. A large part of emotional communication is conveyed through the voice. Throughout this thesis, several studies have been conducted to investigate the influence of familiarity on emotional voice perception. Using different experimental paradigms (see Table 6.1), this thesis provided a stream of evidence for an "other-voice" effect on the processing of expressive voices, which spans both the individual (self vs other) and the collective level (same vs opposite sex, native vs foreign language).

Tab. 6.1: Overview of experimental paradigms

Paradigm	Self/other	Native/foreign language
Speaker discrimination	Studies 1 & 2	
Speaker identity matching	Study 2	
Emotion identification	Study 1	Study 5
Emotional/auditory change detection	Studies 3 & 4	Study 6

This "other-voice" effect seems consistent with two competing driving forces. The first one is linked to auditory familiarity, which warrants self-produced and in-group stimuli with richer sensory and/or motor representations. As a result, this familiarity seems to lead to accuracy advantages in recognizing emotions in one's own voice compared to an another person's voice (Study 1), one's native language compared to a foreign language (Study 5), and in detecting pitch changes in same-sex compared to opposite-sex voices (Study 6). Such facilitation plausibly results from increased perceptual learning or familiarity for self and in-group stimuli with greater exposure (consistent e.g. with the other-race effect on face recognition, Figure 4.1), and/or from the additional recruitment of articulatory representations for self-produced stimuli that are easier to simulate *with the self* (consistent e.g. with the facilitated prediction of self-produced actions, Figure 2.1). A second driving force of the "other-voice" effect is linked to stimulus relevance, which warrants other-voice, but also in-group (same sex, native language) stimuli with more social/biological relevance, leading to faster response times when detecting pitch changes on another person's voice than on the self-voice (Study 4), on voices of the same sex than the opposite sex (Study 6) and, to some extent, in one's native language compared to a foreign language (Study 6), as well as to an earlier MMN onset for expressive deviants in the other-voice sequences (Study 3). Such stimuli may be prioritized in auditory processing because of their biological relevance, as it may be more urgent to attend to external, potentially threatening stimuli, or social relevance, as it may be more important to decode social signals in the in-group rather than the out-group.

Taken together, these two forces either work together or against each other. In the first case, people seem to be both more accurate and faster on voices of their own sex than the opposite sex, or on voices of their native language than a foreign language. In the latter case, processing of the self-voice compared to the other voice seems to be both richer on the one hand, but slower on the other hand. An overview of these two driving forces is presented in Figure 6.1.

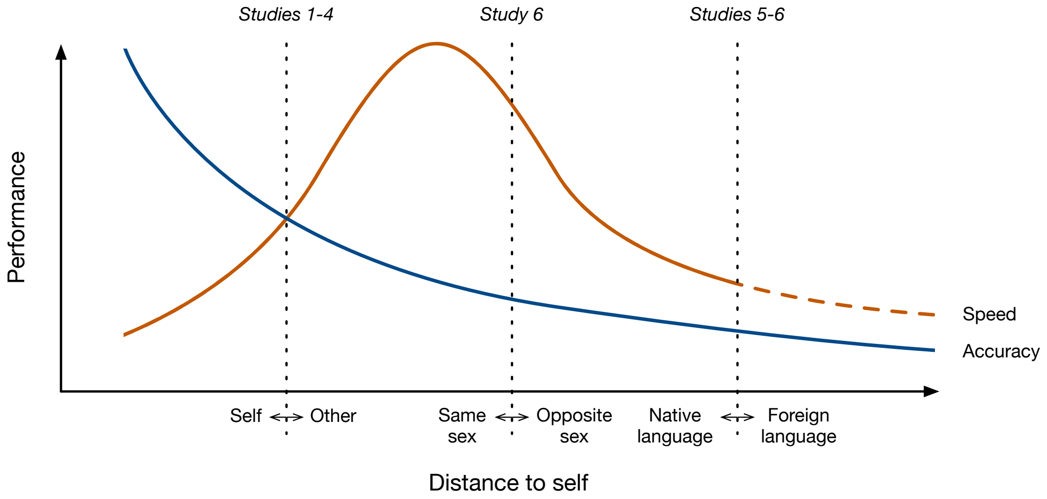


Fig. 6.1: Schematic representation of the sumarized results presented in this thesis. The progression from personal familiarity to socio-cultural familiarity is presented along the x-axis. The performance measures are presented along the y-axis in terms of accuracy (blue) and response time (orange). The dashed part of the orange curve represents an absence of significant results in this thesis.

It would be interesting to test these effects by systematically varying familiarity in the same study, for example by also including personally familiar voices (such as family members, or friends, see e.g. Graux et al., 2015). In the studies presented here, the effect of personal familiarity was assessed by contrasting self and non-self voices. However, inclusion of familiar non-self voices would allow for more precise investigation of the effect of familiarity on the two proposed driving forces. Our results would predict faster processing (response times and MMN onset) of familiar voices with respect to both the self-voice and unfamiliar voices. Additionally, familiar voices would be expected to lead to accuracy advantages with respect to unfamiliar voices but not to the same extent as the self-voice.

Regarding the influence of familiarity on emotional voice recognition and the processing of lower-level acoustic cues, it would be interesting to further investigate driving mechanisms of LFEs on emotion recognition. For instance, Orena and colleagues (2015) found that not just linguistic competence, but also mere exposure to non-native languages facilitated speaker discrimination, reducing an LFE. Other studies however failed to find such an effect of passive experience of foreign language (Perrachione and Wong, 2007; Xie and Myers, 2015). If indeed mere exposure to an unfamiliar language aids speaker identification, this effect may also be mirrored in emotion recognition paradigms and is worth investigating in future studies.

Finally, fMRI and TMS paradigms using DAVID would allow for a more precise investigation of the underlying neural processes (fMRI) and their causal role (TMS) in the perception of emotional speech produced by in-group and out-group speakers. The results of Study 4 warrant further investigation of the involvement of motor regions in the processing of emotional cues on non-self voices. The recruitment of motor areas may be due to the perceived distance between the self and non-self voices as suggested by Bartoli and colleagues (2015) or alternatively may be modulated by the social relevance of the speaker. In future neuroimaging studies it would therefore be interesting to take subjective judgments the difference between speaker and listener ino account and to manipulate social relevance by contrasting communicative and non-communicative contexts in which emotional expressions are presented to further investigate the mechanisms driving the results found in Study 4.

From a methodological point of view, the main advantage of DAVID in this thesis was to provide a tight control over the emotional cues used in different conditions. This allowed us to control within-speaker variance of the acoustic cues of interest in order to study the effects of between-speaker variance at different levels (individual and collective) on the processing of expressive voices. The use of emotional expressions produced by actors would have made this research more challenging due to the inevitable within-speaker variance that would likely blur the effects of voice familiarity discussed in this thesis. Additionally, the use of DAVID allowed us to link low-level auditory processing skills, assessed with an adaptive procedure, to effects of language familiarity on pitch perception using the same paradigm. A similar effort has been made by Xie and Myers (2015) who investigated the influence of musical expertise on speaker identification in the listeners' native language and in foreign languages. However, in their study, pitch processing was assessed in an independent task using different non-voice stimuli. As such, this design did not take into account that pitch processing by itself may already be influenced by language familiarity effects as shown in Study 6. A more direct comparison of these two different skills (pitch processing and emotional processing) is made possible by transformation tools such as DAVID.

In sum, the results presented in this thesis suggest an “other-voice” effect that modulates emotional voice perception. Investigation of potential underlying mechanisms driving this effect was made possible by using voice transformation software and future studies should aim to use this in combination with neuroimaging techniques to reveal the neural processes involved in the interplay between familiarity and emotional voice perception.

Bibliography

Alho, Kimmo and Nina Sinervo (1997). “Preattentive processing of complex sounds in the human brain”. In: *Neuroscience Letters* 233.1, pp. 33–36.

Anikin, Andrey (2018). “Soundgen: An open-source tool for synthesizing nonverbal vocalizations”. In: *Behavior research methods* Advance online publication.

Anikin, Andrey, Rasmus Bååth, and Tomas Persson (2018). “Human Non-linguistic Vocal Repertoire: Call Types and Their Meaning”. In: *Journal of nonverbal behavior* 42.1, pp. 53–80.

Arias, Pablo, Pascal Belin, and Jean-Julien Aucouturier (2018a). “Auditory smiles trigger unconscious facial imitation”. In: *Current Biology* 28.14, R782–R783.

Arias, Pablo, Catherine Soladie, Oussema Bouafif, et al. (2018b). “Realistic transformation of facial and vocal smiles in real-time audiovisual streams”. In: *IEEE Transactions on Affective Computing*.

Armony, Jorge L and Karine Sergerie (2007). “Own-sex effects in emotional memory for faces”. In: *Neuroscience Letters* 426.1, pp. 1–5.

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.

Aucouturier, Jean-Julien, Petter Johansson, Lars Hall, et al. (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.

Aylett, Matthew P, Blaise Potard, and Christopher J Pidcock (2013). “Expressive speech synthesis: Synthesising ambiguity”. In: *8th ISCA Speech Synthesis Workshop*, pp. 217–221.

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, pp. 614–636.

Barrett, Lisa Feldman (2006). “Are emotions natural kinds?” In: *Perspectives on psychological science* 1.1, pp. 28–58.

Bartoli, Eleonora, Alessandro D’Ausilio, Jeffrey Berry, et al. (2015). “Listener–speaker perceived distance predicts the degree of motor contribution to speech perception”. In: *Cerebral Cortex* 25.2, pp. 281–288.

Bayer, Mareike, Katja Ruthmann, and Annekathrin Schacht (2017). “The impact of personal relevance on emotion processing: Evidence from event-related potentials and pupillary responses”. In: *Social Cognitive and Affective Neuroscience* 12 (9), pp. 1470–1479.

Beauchemin, Maude, Louis De Beaumont, Phetsamone Vannasing, et al. (2006). “Electrophysiological markers of voice familiarity”. In: *European Journal of Neuroscience* 23.April, pp. 3081–3086.

Bedoya, Daniel (2018). “Les émotions sont-elles exprimées de la meme façon en musique que dans la voix parlée?” Unpublished master’s thesis, Sorbonne Université, Paris, France.

Behroozmand, Roozbeh, Hanjun Liu, and Charles R Larson (2011). “Time-dependent neural processing of auditory feedback during voice pitch error detection.” In: *Journal of cognitive neuroscience* 23, pp. 1205–1217.

Belin, Pascal and Marie-Hélène Grosbras (2010). “Before speech: cerebral voice processing in infants”. In: *Neuron* 65.6, pp. 733–735.

Belin, Pascal, Shirley Fecteau, and Catherine Bedard (2004). “Thinking the voice: neural correlates of voice perception”. In: *Trends in cognitive sciences* 8.3, pp. 129–135.

Biehl, Michael, David Matsumoto, Paul Ekman, et al. (1997). “Matsumoto and Ekman’s Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences”. In: *Journal of Nonverbal behavior* 21.1, pp. 3–21.

Boersma, Paul and David Weenink (2006). “Praat (Version 4.5)[Computer software]”. In: *Amsterdam: Institute of Phonetic Sciences*.

Breitenstein, Caterina, Diana Van Lancker, and Irene Daum (2001). “The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample”. In: *Cognition & Emotion* 15.1, pp. 57–79.

Bruce, Vicki (1988). *Recognising faces*. London, United Kingdom: Routledge.

Bruce, Vicki and Andy Young (1986). “Understanding face recognition”. In: *British journal of psychology* 77.3, pp. 305–327.

Campanella, Salvatore, C Gaspard, Damien Debatisse, et al. (2002). “Discrimination of emotional facial expressions in a visual oddball task: an ERP study”. In: *Biological psychology* 59.3, pp. 171–186.

Candini, Michela, Elisa Zamagni, Angela Nuzzo, et al. (2014). “Who is speaking? Implicit and explicit self and other voice recognition”. In: *Brain and Cognition* 92, pp. 112–117.

Carminati, Mathilde, Nicole Fiori-Duharcourt, and Frédéric Isel (2018). “Neurophysiological differentiation between preattentive and attentive processing of emotional expressions on French vowels”. In: *Biological psychology* 132, pp. 55–63.

Cellerino, Alessandro, Davide Borghetti, and Ferdinando Sartucci (2004). “Sex differences in face gender recognition in humans”. In: *Brain research bulletin* 63.6, pp. 443–449.

Charest, Ian, Cyril R Pernet, Guillaume A Rousselet, et al. (2009). “Electrophysiological evidence for an early processing of human voices”. In: *Bmc Neuroscience* 10.1, p. 127.

Chen, Bohan, Norihide Kitaoka, and Kazuya Takeda (2016). “Impact of acoustic similarity on efficiency of verbal information transmission via subtle prosodic cues”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2016.1, p. 19.

Chen, Chenyi, Yu-Hsuan Lee, and Yawei Cheng (2014). "Anterior insular cortex activity to emotional salience of voices in a passive oddball paradigm". In: *Frontiers in human neuroscience* 8, p. 743.

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.

Conde, Tatiana, Óscar F Gonçalves, and Ana P Pinheiro (2015). "Paying attention to my voice or yours: An ERP study with words". In: *Biological psychology* 111, pp. 40–52.

– (2016). "The effects of stimulus complexity on the preattentive processing of self-generated and nonself voices: An ERP study". In: *Cognitive, Affective, & Behavioral Neuroscience* 16.1, pp. 106–123.

Conway, Martin A and Christopher W Pleydell-Pearce (2000). "The construction of autobiographical memories in the self-memory system." In: *Psychological review* 107.2, p. 261.

Costa, Jean, Malte F Jung, Mary Czerwinski, et al. (2018). "Regulating Feelings During Interpersonal Conflicts by Changing Voice Self-perception". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, p. 631.

De Mendoza, Alejandra Hurtado, José Miguel Fernández-Dols, W Gerrod Parrott, and Pilar Carrera (2010). "Emotion terms, category structure, and the problem of translation: The case of shame and vergüenza". In: *Cognition & Emotion* 24.4, pp. 661–680.

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

Dromey, Christopher, Sharee O Holmes, J Arden Hopkin, and Kristine Tanner (2015). "The Effects of Emotional Expression on Vibrato". In: *Journal of Voice* 29.2, pp. 170–181.

Dunbar, Robin et al. (1998). "Theory of mind and the evolution of language". In: *Approaches to the Evolution of Language*, pp. 92–110.

Duncan, Seth and Lisa Feldman Barrett (2007). "Affect is a form of cognition: A neurobiological analysis." In: *Cognition & emotion* 21.764698767, pp. 1184–1211.

Dunham, Yarrow, Ron Dotsch, Amelia R Clark, and Elena V Stepanova (2016). "The development of White-Asian categorization: Contributions from skin color and other physiognomic cues". In: *PloS one* 11.6, e0158211.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, and Jacques Mehler (1999). "Epenthetic vowels in Japanese: A perceptual illusion?" In: *Journal of experimental psychology: human perception and performance* 25.6, p. 1568.

D'Ausilio, Alessandro, Ilaria Bufalari, Paola Salmas, and Luciano Fadiga (2012). "The role of the motor system in discriminating normal and degraded speech sounds". In: *Cortex* 48.7, pp. 882–887.

Eckstein, Korinna and Angela D Friederici (2006). "It's early: event-related potential evidence for initial interaction of syntax and prosody in speech comprehension". In: *Journal of Cognitive Neuroscience* 18.10, pp. 1696–1711.

Ekman, Paul and Wallace V Friesen (1971). "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2, p. 124.

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

Elfenbein, Hillary Anger (2013). “Nonverbal dialects and accents in facial expressions of emotion”. In: *Emotion Review* 5.1, pp. 90–96.

Elfenbein, Hillary Anger and Nalini Ambady (2002). “On the universality and cultural specificity of emotion recognition: A meta-analysis”. In: *Psychological Bulletin* 128.2, pp. 203–235.

Elfenbein, Hillary Anger, Manas K. Mandal, Nalini Ambady, Susumu Harizuka, and Surender Kumar (2002). “Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques”. In: *Emotion* 2.1, pp. 75–84.

Engelmann, Jan B and Marianna Pogosyan (2013). “Emotion perception across cultures: the role of cognitive mechanisms”. In: *Frontiers in psychology* 4, p. 118.

Escera, Carles, Kimmo Alho, István Winkler, and Risto Näätänen (1998). “Neural mechanisms of involuntary attention to acoustic novelty and change”. In: *Journal of cognitive neuroscience* 10.5, pp. 590–604.

Ethofer, Thomas, Silke Anders, Sarah Wiethoff, et al. (2006). “Effects of prosodic emotional intensity on activation of associative auditory cortex”. In: *Neuroreport* 17.3, pp. 249–253.

Evans, Samuel and Matthew H. Davis (2015). “Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis”. In: *Cerebral Cortex* 25.12, pp. 4772–4788.

Feinberg, David R (2008). “Are human faces and voices ornaments signaling common underlying cues to mate value?” In: *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews* 17.2, pp. 112–118.

Ferragne, Emmanuel and François Pellegrino (2010). “Formant frequencies of vowels in 13 accents of the British Isles”. In: *Journal of the International Phonetic Association* 40.1, pp. 1–34.

Fischer, Catherine, Frédéric Dailler, and Dominique Morlet (2008). “Novelty P3 elicited by the subject’s own name in comatose patients”. In: *Clinical neurophysiology* 119.10, pp. 2224–2230.

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

Fitch, W Tecumseh and Jay Giedd (1999). “Morphology and development of the human vocal tract: A study using magnetic resonance imaging”. In: *The Journal of the Acoustical Society of America* 106.3, pp. 1511–1522.

Flege, James Emil (1987). “The production of “new” and “similar” phones in a foreign language: Evidence for the effect of equivalence classification”. In: *Journal of phonetics* 15.1, pp. 47–65.

Fleming, David, Bruno L Giordano, Roberto Caldara, and Pascal Belin (2014). “A language-familiarity effect for speaker discrimination without comprehension”. In: *Proceedings of the National Academy of Sciences* 111.38, pp. 13795–13798.

Frassinetti, Francesca, Francesca Ferri, Manuela Maini, Maria Grazia Benassi, and Vittorio Gallese (2011). “Bodily self: An implicit knowledge of what is explicitly unknown”. In: *Experimental Brain Research* 212.1, pp. 153–160.

Frith, C. D. (2012). “The role of metacognition in human social interactions”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, pp. 2213–2223.

Furl, Nicholas, P Jonathon Phillips, and Alice J O’Toole (2002). “Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis”. In: *Cognitive Science* 26.6, pp. 797–815.

Gallese, Vittorio, Christian Keysers, and Giacomo Rizzolatti (2004). “A unifying view of the basis of social cognition”. In: *Trends in cognitive sciences* 8.9, pp. 396–403.

Garrido, Marta I, James M Kilner, Klaas E Stephan, and Karl J Friston (2009). “The mismatch negativity: a review of underlying mechanisms”. In: *Clinical neurophysiology* 120.3, pp. 453–463.

Giard, Marie-Hélène, François Perrin, Jacques Pernier, and Patrick Bouchet (1990). “Brain generators implicated in the processing of auditory stimulus deviance: a topographic event-related potential study”. In: *Psychophysiology* 27.6, pp. 627–640.

Gramfort, Alexandre, Théodore Papadopoulo, Emmanuel Olivi, and Maureen Clerc (2010). “OpenMEEG: opensource software for quasistatic bioelectromagnetics”. In: *Biomedical engineering online* 9.1, p. 45.

Graux, Jérôme, Marie Gomot, Sylvie Roux, et al. (2013). “My voice or yours? An electrophysiological study”. In: *Brain topography* 26.1, pp. 72–82.

Graux, Jérôme, Marie Gomot, Sylvie Roux, Frédérique Bonnet-Brilhault, and Nicole Bruneau (2015). “Is my voice just a familiar voice? An electrophysiological study”. In: *Social cognitive and affective neuroscience* 10.1, pp. 101–105.

Greenlee, Jeremy D W, Adam W. Jackson, Fangxiang Chen, et al. (2011). “Human auditory cortical activation during self-vocalization”. In: *PLoS ONE* 6.3.

Grèzes, Julie and Jean Decety (2001). “Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis”. In: *Human brain mapping* 12.1, pp. 1–19.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge, United Kingdom: Cambridge University Press.

Hahne, Anja and Jörg D Jescheniak (2001). “What’s left if the Jabberwock gets the semantics? An ERP investigation into semantic and syntactic processes during auditory sentence comprehension”. In: *Cognitive Brain Research* 11.2, pp. 199–212.

Haspelmath, Martin (2005). *The world atlas of language structures*. Vol. 1. Oxford, United Kingdom: Oxford University Press.

Heinks-Maldonado, Theda H, Srikantan S Nagarajan, and John F Houde (2006). “Magnetoencephalographic evidence for a precise forward model in speech production.” In: *Neuroreport* 17.13, pp. 1375–1379.

Hervais-Adelman, Alexis G, Robert P Carlyon, Ingrid S Johnsrude, and Matthew H Davis (2012). “Brain regions recruited for the effortful comprehension of noise-vocoded words”. In: *Language and Cognitive Processes* 27.7-8, pp. 1145–1166.

Hickok, Gregory and David Poeppel (2007). “The cortical organization of speech processing.” In: *Nature reviews. Neuroscience* 8.May, pp. 393–402.

Hodges-Simeon, Carolyn R, Steven JC Gaulin, and David A Puts (2010). “Different vocal parameters predict perceptions of dominance and attractiveness”. In: *Human Nature* 21.4, pp. 406–427.

Holmes, Emma, Ysabel Domingo, and Ingrid S Johnsrude (2018). “Familiar voices are more intelligible, even if they are not recognized as familiar”. In: *Psychological science* 29.10, pp. 1575–1583.

Hughes, Susan M. and Shevon E. Nicholson (2010). “The processing of auditory and visual recognition of self-stimuli”. In: *Consciousness and Cognition* 19.4, pp. 1124–1134.

Hung, An-Yi and Yawei Cheng (2014). “Sex differences in preattentive perception of emotional voices and acoustic attributes”. In: *Neuroreport* 25.7, pp. 464–469.

Ilie, Gabriella and William Forde Thompson (2006). “A comparison of acoustic cues in music and speech for three dimensions of affect”. In: *Music Perception: An Interdisciplinary Journal* 23.4, pp. 319–330.

Iyer, Parameswaran Mahadeva, Kieran Mohr, Michael Broderick, et al. (2017). “Mismatch Negativity as an Indicator of Cognitive Sub-Domain Dysfunction in Amyotrophic Lateral Sclerosis”. In: *Frontiers in Neurology* 8, p. 395.

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.

Jacobsen, Thomas and Erich Schröger (2001). “Is there pre-attentive memory-based comparison of pitch?” In: *Psychophysiology* 38, pp. 723–727.

Jacobsen, Thomas, Erich Schröger, István Winkler, and János Horváth (2005). “Familiarity affects the processing of task-irrelevant auditory deviance”. In: *Journal of Cognitive Neuroscience* 17.11, pp. 1704–1713.

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

Jenkins, Rob, David White, Xandra Van Montfort, and A Mike Burton (2011). “Variability in photos of the same face”. In: *Cognition* 121.3, pp. 313–323.

Jiang, Aishi, Jianfeng Yang, and Yufang Yang (2014). “MMN responses during implicit processing of changes in emotional prosody: an ERP study using Chinese pseudo-syllables”. In: *Cognitive neurodynamics* 8.6, pp. 499–508.

Johnson, Elizabeth K, Ellen Westrek, Thierry Nazzi, and Anne Cutler (2011). “Infant ability to tell voices apart rests on language experience”. In: *Developmental Science* 14.5, pp. 1002–1011.

Johnsrude, Ingrid S, Allison Mackey, Hélène Hakyemez, et al. (2013). “Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice”. In: *Psychological Science* 24.10, pp. 1995–2004.

Jones, Benedict C, David R Feinberg, Lisa M DeBruine, Anthony C Little, and Jovana Vukovic (2010). “A domain-specific opposite-sex bias in human preferences for manipulated voice pitch”. In: *Animal Behaviour* 79.1, pp. 57–62.

Junger, Jessica, Katharina Pauly, Sabine Bröhr, et al. (2013). “Sex matters: Neural correlates of voice gender perception”. In: *Neuroimage* 79, pp. 275–287.

Jürgens, Rebecca, Annika Grass, Matthis Drolet, and Julia Fischer (2015). “Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected”. In: *Journal of nonverbal behavior* 39.3, pp. 195–214.

Juslin, Patrik N and Petri Laukka (2001). “Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion.” In: *Emotion* 1.4, pp. 381–412.

– (2003). “Communication of emotions in vocal expression and music performance: different channels, same code?” In: *Psychological bulletin* 129.5, pp. 770–814.

Juslin, Patrik N and Daniel Västfjäll (2008). “Emotional responses to music: the need to consider underlying mechanisms.” In: *The Behavioral and brain sciences* 31.5, 559–75; discussion 575–621.

Kaplan, Jonas T., Lisa Aziz-Zadeh, Lucina Q. Uddin, and Marco Iacoboni (2008). “The self across the senses: An fMRI study of self-face and self-voice recognition”. In: *Social Cognitive and Affective Neuroscience* 3, pp. 218–223.

Kawahara, Hideki and Hisami Matsui (2003). “Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation”. In: *In Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*. Vol. 1. IEEE, pp. I–I.

Kayyal, Mary H and James A Russell (2012). “Language and emotion: certain English–Arabic translations are not equivalent”. In: *Journal of Language and Social Psychology*, p. 0261927X12461004.

Keenan, Julian Paul, Bruce McCutcheon, Stefanie Freund, et al. (1999). “Left hand advantage in a self-face recognition task”. In: *Neuropsychologia* 37.12, pp. 1421–1425.

Keenan, Julian Paul, Aaron Nelson, Margaret O’connor, and Alvaro Pascual-Leone (2001). “Neurology: Self-recognition and the right hemisphere”. In: *Nature* 409.6818, p. 305.

Kelly, David J, Paul C Quinn, Alan M Slater, et al. (2007). “The other-race effect develops during infancy: Evidence of perceptual narrowing”. In: *Psychological Science* 18.12, pp. 1084–1089.

Kiesel, Andrea, Jeff Miller, Pierre Jolicœur, and Benoit Brisson (2008). “Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods”. In: *Psychophysiology* 45.2, pp. 250–274.

Kimura, Marino and Yuko Yotsumoto (2018). “Auditory traits of" own voice"”. In: *PloS one* 13.6, e0199443.

Kitayama, Shinobu, Hazel Rose Markus, and Masaru Kurokawa (2000). “Culture, emotion, and well-being: Good feelings in Japan and the United States”. In: *Cognition & Emotion* 14.1, pp. 93–124.

Kitayama, Shinobu, Batja Mesquita, and Mayumi Karasawa (2006). “Cultural affordances and emotional experience: socially engaging and disengaging emotions in Japan and the United States.” In: *Journal of personality and social psychology* 91.5, p. 890.

Klofstad, Casey A, Rindy C Anderson, and Susan Peters (2012). “Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women”. In: *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20120311.

Klofstad, Casey A, Rindy C Anderson, and Stephen Nowicki (2015). “Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices”. In: *PloS one* 10.8, e0133779.

Knoblich, Günther and Rüdiger Flach (2001). “Predicting the effects of actions: Interactions of perception and action”. In: *Psychological science* 12.6, pp. 467–472.

Korb, Sebastian, Paula Niedenthal, Susanne Kaiser, Didier Grandjean, et al. (2014). “The perception and mimicry of facial movements predict judgments of smile authenticity”. In: *PloS one* 9.6, e99194.

Kotlewska, Ilona and Anna Nowicka (2015). “Present self, past self and close-other: Event-related potential study of face and name detection”. In: *Biological Psychology* 110, pp. 201–211.

Kotz, Sonja A and Silke Paulmann (2007). “When emotional prosody and semantics dance cheek to cheek: ERP evidence”. In: *Brain research* 1151, pp. 107–118.

Kovarski, Klara, Marianne Latinus, Judith Charpentier, et al. (2017). “Facial Expression Related vMMN: Disentangling Emotional from Neutral Change Detection”. In: *Frontiers in human neuroscience* 11, p. 18.

Kreitewolf, Jens, Samuel R Mathias, and Katharina von Kriegstein (2017). “Implicit Talker Training Improves Comprehension of Auditory Speech in Noise”. In: *Frontiers in psychology* 8, p. 1584.

Kret, Mariska E and Beatrice De Gelder (2012). “A review on sex differences in processing emotional signals”. In: *Neuropsychologia* 50.7, pp. 1211–1221.

Kuhl, Patricia K, Karen A Williams, Francisco Lacerda, Kenneth N Stevens, and Bjorn Lindblom (1992). “Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age”. In: *Science* 255, pp. 606–608.

Kujala, Teija, Mari Tervaniemi, and Erich Schröger (2007). “The mismatch negativity in cognitive and clinical neuroscience: theoretical and methodological considerations”. In: *Biological psychology* 74.1, pp. 1–19.

Laird, James D and Katherine Lacasse (2014). “Bodily influences on emotional feelings: Accumulating evidence and extensions of William James’s theory of emotion”. In: *Emotion Review* 6.1, pp. 27–34.

Lamm, Claus, Jean Decety, and Tania Singer (2011). “Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain”. In: *Neuroimage* 54.3, pp. 2492–2502.

Langner, Oliver, Ron Dotsch, Gijsbert Bijlstra, et al. (2010). “Presentation and validation of the Radboud Faces Database”. In: *Cognition and emotion* 24.8, pp. 1377–1388.

Lappe, Claudia, Markus Lappe, and Christo Pantev (2016). “Differential processing of melodic, rhythmic and simple tone deviations in musicians-an MEG study”. In: *Neuroimage* 124, pp. 898–905.

Latinus, Marianne, Phil McAleer, Patricia E G Bestelmeyer, and Pascal Belin (2013). “Norm-based coding of voice identity in human auditory cortex”. In: *Current Biology* 23.12, pp. 1075–1080.

Laukka, Petri, Patrik Juslin, and Roberto Bresin (2005). “A dimensional approach to vocal expression of emotion”. In: *Cognition & Emotion* 19.5, pp. 633–653.

Laukka, Petri, Tuomas Eerola, Nutankumar S Thingujam, Teruo Yamasaki, and Grégory Beller (2013). “Universal and culture-specific factors in the recognition and performance of musical affect expressions.” In: *Emotion (Washington, D.C.)* 13.3, pp. 434–49.

LeDoux, Joseph E and Richard Brown (2017). “A higher-order theory of emotional consciousness”. In: *Proceedings of the National Academy of Sciences* 114.10, E2016–E2025.

Leek, Marjorie R (2001). “Adaptive procedures in psychophysical research”. In: *Perception & psychophysics* 63.8, pp. 1279–1292.

Levenson, Robert W (2003). “Blood, sweat, and fears: The autonomic architecture of emotion”. In: *Annals of the New York Academy of Sciences* 1000.1, pp. 348–366.

Lévêque, Yohana, Neil Muggleton, Lauren Stewart, and Daniele Schön (2013). “Involvement of the larynx motor area in singing-voice perception: A TMS study”. In: *Frontiers in Psychology* 4, p. 418.

Levi, Susannah V, Stephen J Winters, and David B Pisoni (2011). “Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible?” In: *The Journal of the Acoustical Society of America* 130.6, pp. 4053–4062.

Levitt, H (1971). “Transformed up-down methods in psychoacoustics”. In: *The Journal of the Acoustical society of America* 49.2, pp. 467–477.

Lieberman, Philip, Jeffrey T Laitman, Joy S Reidenberg, and Patrick J Gannon (1992). “The anatomy, physiology, acoustics and perception of speech: essential elements in analysis of the evolution of human speech”. In: *Journal of Human Evolution* 23.6, pp. 447–467.

Lima, César F, São Luís Castro, and Sophie K Scott (2013). “When voices get emotional: a corpus of nonverbal vocalizations for research on emotion processing.” In: *Behavior research methods* 45.4, pp. 1234–45.

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

Martínez-Montes, Eduardo, Heivet Hernández-Pérez, Julie Chobert, et al. (2013). “Musical expertise and foreign speech perception”. In: *Frontiers in systems neuroscience* 7, p. 84.

Matsumoto, David (1999). “American-Japanese cultural differences in judgements of expression intensity and subjective experience”. In: *Cognition & Emotion* 13.2, pp. 201–218.

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.

McDougall, Kirsty (2004). “Speaker-specific formant dynamics: an experiment on Australian English/a”. In: *International Journal of Speech, Language and the Law* 11, pp. 103–130.

McGraw, Kenneth O. and S. P. Wong (1992). “A common language effect size statistic”. In: *Psychological Bulletin* 111.2, pp. 361–365.

Meissner, Christian A and John C Brigham (2001). “Thirty Years Of Investigating The Own-race Bias In Memory For Faces”. In: *Psychology, Public Policy and Law* 7.1, pp. 3–35.

Mullennix, John W, Keith A Johnson, Meral Topcu-Durgun, and Lynn M Farnsworth (1995). “The perceptual representation of voice gender”. In: *The Journal of the Acoustical Society of America* 98.6, pp. 3080–3095.

Näätänen, Risto (1990). “The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function”. In: *Behavioral and Brain Sciences* 13.2, pp. 201–233.

Näätänen, Risto, Anthony WK Gaillard, and S Mäntysalo (1978). “Early selective-attention effect on evoked potential reinterpreted”. In: *Acta psychologica* 42.4, pp. 313–329.

Näätänen, Risto, Petri Paavilainen, Teemu Rinne, and Kimmo Alho (2007). “The mismatch negativity (MMN) in basic research of central auditory processing: a review”. In: *Clinical neurophysiology* 118.12, pp. 2544–2590.

Nakamura, Katsuki, Ryuta Kawashima, Motoaki Sugiura, et al. (2001). “Neural substrates for recognition of familiar voices: a PET study”. In: *Neuropsychologia* 39.10, pp. 1047–1054.

Neuloh, Georg and Gabriel Curio (2004). “Does familiarity facilitate the cortical processing of music sounds?” In: *Neuroreport* 15.16, pp. 2471–2475.

Niedenthal, Paula M (2007). “Embodying emotion”. In: *science* 316.5827, pp. 1002–1005.

Niziolek, Caroline A., Srikantan S. Nagarajan, and John F. Houde (2013). “What Does Motor Efference Copy Represent? Evidence from Speech Production”. In: *Journal of Neuroscience* 33.41, pp. 16110–16116.

Öhman, Arne (2002). “Automaticity and the amygdala: Nonconscious responses to emotional faces”. In: *Current directions in psychological science* 11.2, pp. 62–66.

Oliver-Rodriguez, Juan C, Zhiqiang Guan, and Victor S Johnston (1999). “Gender differences in late positive components evoked by human faces”. In: *Psychophysiology* 36.2, pp. 176–185.

Oostenveld, Robert, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen (2011). “FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data”. In: *Computational Intelligence and Neuroscience* 2011.

Oosterhof, Nikolaas N and Alexander Todorov (2008). “The functional basis of face evaluation”. In: *Proceedings of the National Academy of Sciences* 105.32, pp. 11087–11092.

Orena, Adriel John, Rachel M Theodore, and Linda Polka (2015). “Language exposure facilitates talker learning prior to language comprehension, even in adults”. In: *Cognition* 143, pp. 36–40.

Pakarinen, Satu, Laura Sokka, Marianne Leinikka, et al. (2014). “Fast determination of MMN and P3a responses to linguistically and emotionally relevant changes in pseudoword stimuli”. In: *Neuroscience letters* 577, pp. 28–33.

Pampalk, Elias (2004). “A Matlab Toolbox to Compute Music Similarity from Audio.” In: *ISMIR*.

Patel, Sona and Klaus R Scherer (2013). “Vocal behaviour”. In: *Handbook of nonverbal communication*, pp. 167–204.

Paulmann, Silke and Ayse K Uskul (2014). “Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners”. In: *Cognition & emotion* 28.2, pp. 230–244.

Peirce, Jonathan W (2007). “PsychoPy—psychophysics software in Python”. In: *Journal of neuroscience methods* 162.1, pp. 8–13.

Pell, Marc D and Sonja A Kotz (2011). “On the time course of vocal emotion recognition”. In: *PLoS ONE* 6.11, e27256.

Pell, Marc D, Silke Paulmann, Chinar Dara, Areej Alasseri, and Sonja A Kotz (2009a). “Factors in the recognition of vocally expressed emotions: A comparison of four languages”. In: *Journal of Phonetics* 37.4, pp. 417–435.

Pell, Marc D, Laura Monetta, Silke Paulmann, and Sonja A Kotz (2009b). “Recognizing emotions in a foreign language”. In: *Journal of Nonverbal Behavior* 33.2, pp. 107–120.

Perrachione, Tyler K and Patrick CM Wong (2007). “Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex”. In: *Neuropsychologia* 45.8, pp. 1899–1910.

Perrachione, Tyler K, Stephanie N Del Tufo, and John DE Gabrieli (2011). “Human voice recognition depends on language ability”. In: *Science* 333.6042, pp. 595–595.

Perrin, Fabien, Luis Garcıa-Larrea, François Mauguière, and Hélène Bastuji (1999). “A differential brain response to the subject’s own name persists during sleep”. In: *Clinical Neurophysiology* 110.12, pp. 2153–2164.

Pinheiro, Ana P, Neguine Rezaii, Paul G Nestor, et al. (2016). “Did you or I say pretty, rude or brief? An ERP study of the effects of speaker’s identity on emotional word processing”. In: *Brain and language* 153, pp. 38–49.

Pinheiro, Ana P, Carla Barros, Margarida Vasconcelos, Christian Obermeier, and Sonja A Kotz (2017). “Is laughter a better vocal change detector than a growl?” In: *Cortex* 92, pp. 233–248.

Pisanski, Katarzyna, Paul J Fraccaro, Cara C Tigue, Jillian JM O’connor, and David R Feinberg (2014). “Return to Oz: Voice pitch facilitates assessments of men’s body size.” In: *Journal of Experimental Psychology: Human Perception and Performance* 40.4, p. 1316.

Pisoni, David B (1993). “Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning”. In: *Speech communication* 13.1, pp. 109–125.

Ponsot, Emmanuel, Juan José Burred, Pascal Belin, and Jean-Julien Aucouturier (2018a). “Cracking the social code of speech prosody using reverse correlation”. In: *Proceedings of the National Academy of Sciences* 115.15, pp. 3972–3977.

Ponsot, Emmanuel, Pablo Arias, and Jean-Julien Aucouturier (2018b). “Uncovering mental representations of smiled speech using reverse correlation”. In: *The Journal of the Acoustical Society of America* 143.1, EL19–EL24.

Puts, David Andrew, Steven JC Gaulin, and Katherine Verdolini (2006). “Dominance and the evolution of sexual dimorphism in human voice pitch”. In: *Evolution and human behavior* 27.4, pp. 283–296.

Quené, Hugo, Gün R Semin, and Francesco Foroni (2012). “Audible smiles and frowns affect speech comprehension”. In: *Speech Communication* 54.7, pp. 917–922.

Rachman, Laura, Marco Liuni, Pablo Arias, et al. (2018). “DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech”. In: *Behavior Research Methods* 50.1, pp. 323–343.

Repp, Bruno H (1987). “The sound of two hands clapping: An exploratory study”. In: *The Journal of the Acoustical Society of America* 81.4, pp. 1100–1109.

Rizzolatti, Giacomo, Leonardo Fogassi, and Vittorio Gallese (2001). “Neurophysiological mechanisms underlying the understanding and imitation of action”. In: *Nature reviews neuroscience* 2.9, p. 661.

Roesch, Etienne B, Lucas Tamarit, Lionel Reveret, et al. (2011). “FACSGen: a tool to synthesize emotional facial expressions through systematic manipulation of facial action units”. In: *Journal of Nonverbal Behavior* 35.1, pp. 1–16.

Rogers, Timothy B, Nicholas A Kuiper, and William S Kirker (1977). “Self-reference and the encoding of personal information.” In: *Journal of personality and social psychology* 35.9, p. 677.

Rogier, Ophelie, Sylvie Roux, Pascal Belin, Frederique Bonnet-Brilhault, and Nicole Bruneau (2010). “An electrophysiological correlate of voice processing in 4-to 5-year-old children”. In: *International Journal of psychophysiology* 75.1, pp. 44–47.

Rohr, Lana and Rasha Abdel Rahman (2015). “Affective responses to emotional words are boosted in communicative situations”. In: *NeuroImage* 109, pp. 273–282.

Rosa, Christine, Maryse Lassonde, Claudine Pinard, Julian Paul Keenan, and Pascal Belin (2008). “Investigations of hemispheric specialization of self-voice recognition”. In: *Brain and Cognition* 68.2, pp. 204–214.

Roswandowitz, Claudia, Claudia Kappes, Hellmuth Obrig, and Katharina von Kriegstein (2018). “Obligatory and facultative brain regions for voice-identity recognition”. In: *Brain* 141.1, pp. 234–247.

Roye, Anja, Thomas Jacobsen, and Erich Schröger (2007). “Personal significance is encoded automatically by the human brain: an event-related potential study with ringtones”. In: *European Journal of Neuroscience* 26.3, pp. 784–790.

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.

Sams, M, P Paavilainen, K Alho, and R Näätänen (1985). “Auditory frequency discrimination and event-related potentials”. In: *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 62.6, pp. 437–448.

Santiesteban, Idalmis, Sarah White, Jennifer Cook, et al. (2012). “Training social cognition: from imitation to theory of mind”. In: *Cognition* 122.2, pp. 228–235.

Sato, Marc, Pascale Tremblay, and Vincent L Gracco (2009). “A mediating role of the premotor cortex in phoneme segmentation”. In: *Brain and language* 111.1, pp. 1–7.

Sauter, Disa (2013). “The role of motivation and cultural dialects in the in-group advantage for emotional vocalizations”. In: *Frontiers in psychology* 4, p. 814.

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.

– (2015). “Emotional vocalizations are recognized across cultures regardless of the valence of distractors”. In: *Psychological science* 26.3, pp. 354–356.

Schall, Sonja, Stefan J. Kiebel, Burkhard Maess, and Katharina von Kriegstein (2015). “Voice identity recognition: Functional division of the right STS and its behavioral relevance”. In: *Journal of Cognitive Neuroscience* 27.2, pp. 280–291.

Scherer, Klaus R (2003). “Vocal communication of emotion: A review of research paradigms”. In: *Speech communication* 40.1, pp. 227–256.

Scherer, Klaus R and James S Oshinsky (1977). “Cue utilization in emotion attribution from auditory stimuli”. In: *Motivation and emotion* 1.4, pp. 331–346.

Scherer, Klaus R, Rainer Banse, and Harald G Wallbott (2001). “Emotion inferences from vocal expression correlate across languages and cultures”. In: *Journal of Cross-cultural psychology* 32.1, pp. 76–92.

Scherer, Klaus R, Tom Johnstone, and Gundrun Klasmeyer (2003). “Vocal expression of emotion”. In: *Handbook of affective sciences*, pp. 433–456.

Scherer, Klaus R, Elizabeth Clark-Polner, and Marcello Mortillaro (2011). “In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion”. In: *International Journal of Psychology* 46.6, pp. 401–435.

Schirmer, Annett and Sonja A Kotz (2006). “Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing”. In: *Trends in Cognitive Sciences* 10.1, pp. 24–30.

Schirmer, Annett, Tricia Striano, and Angela D Friederici (2005). “Sex differences in the preattentive processing of vocal emotional expressions.” In: *Neuroreport* 16.6, pp. 635–9.

Schirmer, Annett, Elizabeth Simpson, and Nicolas Escoffier (2007). “Listen up! Processing of intensity change differs for vocal and nonvocal sounds”. In: *Brain research* 1176, pp. 103–112.

Schröder, Marc (2001). “Emotional speech synthesis: A review”. In: *Seventh European Conference on Speech Communication and Technology*.

Scott, Sophie K and Ingrid S Johnsrude (2003). “The neuroanatomical and functional organization of speech perception”. In: *Trends in neurosciences* 26.2, pp. 100–107.

Sel, Alejandra, Bettina Forster, and Beatriz Calvo-Merino (2014). “The emotional homunculus: ERP evidence for independent somatosensory responses during facial emotional processing”. In: *Journal of Neuroscience* 34.9, pp. 3263–3267.

Sel, Alejandra, Rachel Harding, and Manos Tsakiris (2016). “Electrophysiological correlates of self-specific prediction errors in the human brain”. In: *Neuroimage* 125, pp. 13–24.

Sell, Aaron, Gregory A Bryant, Leda Cosmides, et al. (2010). “Adaptations in humans for assessing physical strength from the voice”. In: *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20100769.

Shapiro, Peter N and Steven Penrod (1986). “Meta-analysis of facial identification studies.” In: *Psychological Bulletin* 100.2, p. 139.

Shuster, Linda I and John D Durrant (2003). “Toward a better understanding of the perception of self-produced speech”. In: *Journal of Communication Disorders* 36.1, pp. 1–11.

Silani, Giorgia, Claus Lamm, Christian C Ruff, and Tania Singer (2013). “Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments”. In: *Journal of Neuroscience* 33.39, pp. 15466–15476.

Skipper, Jeremy I., Joseph T. Devlin, and Daniel R. Lametti (2017). “The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception”. In: *Brain and Language* 164, pp. 77–105.

Smith, David RR and Roy D Patterson (2005). “The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age”. In: *The Journal of the Acoustical Society of America* 118.5, pp. 3177–3186.

Spengler, Stephanie, D Yves von Cramon, and Marcel Brass (2009). “Control of shared representations relies on key processes involved in mental state attribution”. In: *Human brain mapping* 30.11, pp. 3704–3718.

Spreckelmeyer, Katja N, Lena Rademacher, Frieder M Paulus, and Gerhard Gründer (2013). “Neural activation during anticipation of opposite-sex and same-sex faces in heterosexual men and women”. In: *Neuroimage* 66, pp. 223–231.

Steinbeis, Nikolaus (2016). “The role of self–other distinction in understanding others’ mental and emotional states: neurocognitive mechanisms in children and adults”. In: *Phil. Trans. R. Soc. B* 371.1686, p. 20150074.

Sugiura, Motoaki, Ryuta Kawashima, Katsuki Nakamura, et al. (2000). “Passive and active recognition of one’s own face”. In: *Neuroimage* 11.1, pp. 36–48.

Sui, Jie and Glyn W. Humphreys (2013). “The boundaries of self face perception: Response time distributions, perceptual categories, and decision weighting”. In: *Visual Cognition* 21.4, pp. 415–445.

Sui, Jie and Glyn W Humphreys (2017). “The ubiquitous self: what the properties of self-bias tell us about the self”. In: *Annals of the New York Academy of Sciences* 1396.1, pp. 222–235.

Sulykos, István, Krisztina Kecskés-Kovács, and István Czigler (2015). “Asymmetric effect of automatic deviant detection: the effect of familiarity in visual mismatch negativity”. In: *Brain research* 1626, pp. 108–117.

Symons, Cynthia S and Blair T Johnson (1997). “The self-reference effect in memory: a meta-analysis.” In: *Psychological bulletin* 121.3, p. 371.

Tacikowski, Pawel and Anna Nowicka (2010). “Allocation of attention to self-name and self-face: An ERP study”. In: *Biological Psychology* 84.2, pp. 318–324.

Tadel, François, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy (2011). “Brainstorm: a user-friendly application for MEG/EEG analysis”. In: *Computational intelligence and neuroscience* 2011, p. 8.

Takeda, Shoichi, Yoshiki Kabuta, Tomohiro Inoue, and Masashi Hatoko (2013). “Proposal of a Japanese-speech-synthesis Method with Dimensional Representation of Emotions based on Prosody as well as Voice-quality Conversion”. In: *International Journal of Affective Engineering* 12.2, pp. 79–88.

Tamietto, Marco and Beatrice De Gelder (2010). “Neural bases of the non-conscious perception of emotional signals”. In: *Nature Reviews Neuroscience* 11.10, p. 697.

Thompson, William Forde and Laura-Lee Balkwill (2006). “Decoding speech prosody in five languages”. In: *Semiotica* 2006.158, pp. 407–424.

Tian, Xing and David Poeppel (2014). “Dynamics of self-monitoring and error detection in speech production: evidence from mental imagery and MEG”. In: *Journal of cognitive neuroscience* 27.2, pp. 352–364.

Titze, Ingo R and William J Strong (1975). “Normal modes in vocal cord tissues”. In: *The Journal of the Acoustical Society of America* 57.3, pp. 736–744.

Todorov, Alexander, Ron Dotsch, Jenny M Porter, Nikolaas N Oosterhof, and Virginia B Falvello (2013). “Validation of data-driven computational models of social perception of faces.” In: *Emotion* 13.4, p. 724.

Tonndorf, J (1976). “Bone conduction”. In: *Auditory System*. Berlin, Heidelberg: Springer, pp. 37–84.

Tottenham, Nim, James W Tanaka, Andrew C Leon, et al. (2009). “The NimStim set of facial expressions: judgments from untrained research participants”. In: *Psychiatry research* 168.3, pp. 242–249.

Tye-Murray, Nancy, Brent P Spehar, Joel Myerson, Sandra Hale, and Mitchell S Sommers (2015). “The self-advantage in visual speech processing enhances audiovisual speech recognition in noise”. In: *Psychonomic bulletin & review* 22.4, pp. 1048–1053.

Ulrich, Rolf and Jeff Miller (2001). “Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs”. In: *Psychophysiology* 38.5, pp. 816–827.

Uttl, Bob and Alain Morin (2010). “Ceiling effects make Hughes and Nicholson’s data analyses and conclusions inconclusive”. In: *Consciousness and Cognition* 19.4, pp. 1135–1137.

Valentine, Tim (1991). “A unified account of the effects of distinctiveness, inversion, and race in face recognition”. In: *The Quarterly Journal of Experimental Psychology* 43.2, pp. 161–204.

Van Zuijen, Titia L, Veerle L Simoens, Petri Paavilainen, Risto Näätänen, and Mari Tervaniemi (2006). “Implicit, intuitive, and explicit knowledge of abstract regularities in a sound sequence: an event-related brain potential study”. In: *Journal of Cognitive Neuroscience* 18.8, pp. 1292–1303.

Von Kriegstein, Katharina and Anne-Lise Giraud (2004). “Distinct functional substrates along the right superior temporal sulcus for the processing of voices”. In: *Neuroimage* 22.2, pp. 948–955.

Vukovic, Jovana, Benedict C Jones, David R Feinberg, et al. (2011). “Variation in perceptions of physical dominance and trustworthiness predicts individual differences in the effect of relationship context on women’s preferences for masculine pitch in men’s voices”. In: *British Journal of Psychology* 102.1, pp. 37–48.

Wagenmakers, E-J, Titia Beek, Laura Dijkhoff, et al. (2016). “Registered Replication Report: Strack, Martin, & Stepper (1988)”. In: *Perspectives on Psychological Science* 11.6, pp. 917–928.

Wagner, Hugh L (1993). “On measuring performance in category judgment studies of nonverbal behavior”. In: *Journal of nonverbal behavior* 17.1, pp. 3–28.

Warren, JD, SK Scott, CJ Price, and TD Griffiths (2006). “Human brain mechanisms for the early analysis of voices”. In: *Neuroimage* 31.3, pp. 1389–1397.

Williams, Mark A and Jason B Mattingley (2006). “Do angry men get noticed?” In: *Current Biology* 16.11, R402–R404.

Wilson, Stephen M and Marco Iacoboni (2006). “Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception”. In: *Neuroimage* 33.1, pp. 316–325.

Won, Sook Young, Jonathan Berger, and Malcolm Slaney (2014). “Simulation of One‘s Own Voice in a Two-parameter Model”. In: *Proc Int Conf Music Percept Cogn*.

Xie, Xin and Emily Myers (2015). “The impact of musical training and tone language experience on talker identification”. In: *The Journal of the Acoustical Society of America* 137.1, pp. 419–432.

Xu, Mingdi, Fumitaka Homae, Ryu-Ichiro Hashimoto, and Hiroko Hagiwara (2013). “Acoustic cues for the recognition of self-voice and other-voice.” English. In: *Frontiers in psychology* 4, p. 735.

Yadav, Manuj, Densil Cabrera, and Dianna T Kenny (2014). “Towards a method for loudness-based analysis of the sound of one’s own voice”. In: *Logopedics Phoniatrics Vocology* 39.3, pp. 117–125.

Yamaguchi, Masami K, Tastu Hirukawa, and So Kanazawa (1995). “Judgment of gender through facial parts”. In: *Perception* 24.5, pp. 563–575.

List of Figures

List of Tables

Scientific communications

Peer-reviewed publications

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., Richardson, D., Watanabe, K., Dubal, S., & 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.

Ticini, L.F., **Rachman, L.**, Pelletier, J., & Dubal, S. (2014). Enhancing aesthetic apprectiation by priming canvases with actions that match the artist's painting style. *Frontiers in Human Neuroscience, 8*, 391.

Under review

Rachman, L., Dubal, S., & Aucouturier, J.J. (2nd revision). Happy me, happy you: expressive changes on a stranger's voice recruit faster implicit processes than self-produced expressions.

In preparation

Nakai, T., **Rachman, L.**, Arias, P., Okanoya, K., & Aucouturier, J.J. (in prep). Algorithmic voice transformations reveal the phonological basis of language-familiarity effects in cross-cultural emotion judgments.

Conference presentations

Rachman, L., Dubal, S., & Aucouturier, J.J. (2018). Automatic processing of expressive changes on self-produced and other-produced speech, European Society for Cognitive and Affective Neuroscience (ESCAN), Leiden, Netherlands, 19-22 July 2018.

Rachman, L. & Aucouturier, J.J. (2018). Dispatch from the ERC CREAM Project: Four New Software Tools to Manipulate Audio Emotional Signals in Cognitive Psychology and Neuroscience Research, Consortium of European Research on Emotion (CERE), Glasgow, Scotland (UK), 4-5 April 2018.

Rachman, L. (2016). Effects of speaker identity on emotion-related auditory change detection, Tokyo Workshop on Music cognition, emotion and audio technology, Tokyo, Japan, 7 November 2016.

Rachman, L., Liuni, M., Arias, P., & Aucouturier, J.J. (2015). Synthesizing speech-like emotional expression onto music and speech signals, 4th International Conference on Music & Emotion (ICME4), Geneva, Switzerland, 12-16 October 2015.

Poster presentations

Rachman, L., Dubal, S., & Aucouturier, J.J. (2017). Processing of emotion-related vocal cues in self-produced and non-self produced speech, International Conference for Cognitive Neuroscience (ICON), Amsterdam, the Netherlands, 5-8 August 2017.

Rachman, L., Dubal, S., & Aucouturier, J.J. (2017). Processing of emotion-related vocal cues in self-produced and non-self produced speech, Salzburg Mind-Brain Annual Meeting (SAMBA), Salzburg, Austria, 13-14 July 2017.

Rachman, L., Dubal, S., & Aucouturier, J.J. (2016). Effects of speaker identity on emotion-related auditory change detection: An ERP study, Journées Jeunes Chercheurs en Acoustique, Audition et Signal (JJCAAS), Paris, France, 23-25 November 2016.

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., ... & Aucouturier, J.J. (2016). Validation of a new open-source platform for emotional speech transformation, Workshop on Auditory Neuroscience, Cognition and Modelling, London, UK, 17 February 2016.

Conference organisation

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

Master student supervision

Seropian, L. (2017). Contrôle de sa propre voix pendant la production de discours: une étude EEG. IRCAM.

Colophon

This thesis was typeset with $\LaTeX\,2_{\varepsilon}$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.