

Vocal emotions on the brain:

The role of acoustic parameters and musicality



Dissertation
zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena

von M. Sc. Christine Nussbaum
geboren am 20.03.1994 in Freiberg

GutachterInnen:

1. Prof. Dr. Stefan R. Schweinberger (Department for General Psychology and Cognitive Neuroscience, Institute of Psychology, Friedrich Schiller University Jena, Germany)
2. Prof. Dr. Annett Schirmer (Institute of Psychology, University of Innsbruck, Austria)
3. Prof. Dr. César F. Lima (Department of Psychology at Iscte, University Institute of Lisbon, Portugal)

Datum der Verteidigung: 03.07.2023

Contents

1. Zusammenfassung (Summary)	11
2. Introduction	13
2.1. The expression and perception of emotions in voices	14
2.1.1. Vocal expression – deciphering the acoustic code of emotional prosody . .	15
2.1.2. Electrophysiological correlates of vocal emotional processing	17
2.2. Music and vocal expression	19
2.2.1. Emotions in voices and music – a tale of joint evolution	19
2.2.2. Links between musicality and non-musical skills	21
2.2.3. Musicality and vocal emotion perception	23
2.3. Voice morphing	24
2.3.1. Voice morphing using Tandem-STRAIGHT – implementation, requirements, and assumptions	24
2.3.2. Applications of (parameter-specific) voice morphing	26
2.4. Research questions	28
2.5. Research outline	30
3. Links between musicality and vocal emotion perception	33
3.1. Introduction	33
3.2. Systematic literature search	35
3.3. Review of identified literature	36
3.3.1. Differences in vocal emotion perception between musicians and non-musicians	36
3.3.2. Impairments of vocal emotion perception in individuals with congenital amusia	43
3.3.3. Correlation of vocal emotion perception with musical interests or psychoa- coustic abilities	44
3.3.4. Effects of musical training interventions on vocal emotion perception . . .	45
3.4. Discussion	46
3.4.1. Active engagement in musical activities versus receptive training	47
3.4.2. The role of different acoustic cues and supramodal processes	47
3.4.3. Nature and nurture	48
3.4.4. Identification of relevant topics for future research	50
3.5. Conclusion and outlook	50
4. Perceived naturalness of emotional voice morphs	51
4.1. Introduction	51
4.1.1. The potentials and limits of parameter-specific voice morphing in vocal emotional research	52
4.1.2. Perspectives on voice naturalness across different research domains	53
4.1.3. Aims of the present studies	54
4.2. Experiment 1	54
4.2.1. Method	54
4.2.2. Results	58
4.2.3. Short summary	62

4.3.	Experiment 2	63
4.3.1.	Method	63
4.3.2.	Results	64
4.3.3.	Short summary	64
4.4.	Discussion	65
4.4.1.	The role of naturalness in the perception of emotion and other social signals	65
4.4.2.	Is there an uncanny valley for voices?	66
4.4.3.	Emotional voice morphing – a tool of unlimited possibilities?	66
4.4.4.	Directions for future research	68
4.5.	Summary and conclusion	68
5.	Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates	69
5.1.	Introduction	69
5.1.1.	The role of different acoustic parameters in vocal emotion perception . . .	70
5.1.2.	Electrophysiological correlates of vocal emotion perception	71
5.1.3.	Aims of the present study	72
5.2.	Method	72
5.2.1.	Listeners	72
5.2.2.	Stimuli	72
5.2.3.	Design	74
5.2.4.	Data processing and analysis	75
5.3.	Results	76
5.3.1.	Behavioral data – proportion of correct classifications	76
5.3.2.	ERP data	77
5.4.	Discussion	80
5.4.1.	The unique contribution of F0 and timbre in vocal emotion processing . .	80
5.4.2.	Electrophysiological correlates of F0 vs timbre processing	81
5.4.3.	Directions for future research	82
5.5.	Summary and conclusion	83
6.	Musicality - tuned to the melody of vocal emotions	85
6.1.	Introduction	85
6.1.1.	What are the acoustic features of emotions and what is shared between music and voice?	86
6.1.2.	How does musicality benefit vocal emotion perception?	87
6.1.3.	Methodological challenges and aims of the present study	88
6.2.	Method	88
6.2.1.	Participants	88
6.2.2.	Stimuli	89
6.2.3.	Design	90
6.2.4.	Data analysis	92
6.3.	Results	93
6.3.1.	Demographic, musicality, and personality characteristics of participants .	93
6.3.2.	Emotion classification performance	93
6.3.3.	Links between musical skills and vocal emotion perception	96
6.4.	Discussion	98
6.4.1.	The musicality benefit for vocal emotion perception – a matter of auditory sensitivity?	98
6.4.2.	Emotional communication in music and voice – same code, same task? . .	99
6.4.3.	The relative importance of pitch contour (F0) and timbre	100

6.4.4.	Constraints on generality, and future directions	101
6.5.	Summary and outlook	102
7.	Electrophysiological correlates of vocal emotional processing in musicians and non-musicians	103
7.1.	Introduction	103
7.1.1.	Auditory evoked potentials related to musical expertise	104
7.1.2.	Rationale of the study	105
7.1.3.	Hypotheses and analysis plan	106
7.2.	Method	107
7.2.1.	Participants	107
7.2.2.	Stimuli	107
7.2.3.	Design	107
7.2.4.	EEG-data processing and analysis	108
7.3.	Results	109
7.3.1.	Analysis of the P200, the N200, and the LPP	109
7.3.2.	Correlations between ERP amplitudes and the PROMS	109
7.3.3.	Nonparametric cluster-based permutation tests	111
7.3.4.	Behavioral classification task	113
7.4.	Discussion	113
7.4.1.	Electrophysiological correlates of F0 and timbre processing	114
7.4.2.	Electrophysiological correlates of musical expertise	115
7.4.3.	Future research	117
7.4.4.	Summary	117
8.	General Discussion	119
8.1.	The role of F0 and timbre for the processing of vocal emotions	119
8.1.1.	The role of F0 and timbre for emotion recognition	120
8.1.2.	Electrophysiological correlates of emotional F0 and timbre cues	122
8.1.3.	Open questions	123
8.1.4.	Summary and conclusion: What is the contribution of different acoustic cues to vocal emotion perception?	124
8.2.	Links between musicality and vocal emotion perception	124
8.2.1.	Sensitive to melodies? – How musicality benefits the processing of vocal emotions	125
8.2.2.	The role of musical training vs natural auditory sensitivity	126
8.2.3.	Electrophysiological correlates	127
8.2.4.	Limitations and open questions	129
8.2.5.	Summary and conclusion: How does musicality affect the processing of emotional voice cues?	130
8.3.	Reflections on parameter-specific voice morphing	130
8.3.1.	Perceived naturalness of parameter-specific voice morphs	131
8.3.2.	The role of stimulus naturalness for emotional voice processing	132
8.3.3.	Emotional voice morphing – practical recommendations and future applications	133
8.3.4.	Summary and conclusion: Is parameter-specific voice morphing a suitable tool to study the processing of vocal emotions?	135
8.4.	Directions for future research	135
8.5.	Summary and conclusion	138
A.	Supplemental Tables	141

B. Supplemental Figures	151
C. Supplemental Rating Data	157
D. References	163
E. Danksagung	183
F. Ehrenwörtliche Erklärung	185

List of Figures

2.1. Time- and frequency anchors	25
4.1. Illustration of parameter-specific voice morphs based on two different references .	56
4.2. Interaction of Morph Type and Reference Type on perceived Naturalness	60
4.3. Interaction of Morph Type and Emotion on perceived Emotionality	61
4.4. Relationship between mean ratings of perceived naturalness and emotionality (A), or emotional intensity (B)	62
4.5. Response tendency towards the more natural reference category as a function of Morph Type	64
5.1. Schematic illustration of the different parameter-specific voice morphs	73
5.2. Mean proportion of correct responses per Emotion and Morph Type	77
5.3. Confusion matrices for each Emotion separately for the three Morph Types . . .	77
5.4. Scalp topographies of the contrast between the difference waves $\text{Diff}_{\text{Full-F0}}$ and $\text{Diff}_{\text{Full-Timbre}}$ for each emotion separately from 50 to 500 ms	78
5.5. ERPs separately for Emotion and Morph Type, averaged across nine channels . .	79
6.1. Schematic depiction of the voice averaging process	90
6.2. Morphing matrix for stimuli with averaged voices as reference	91
6.3. Boxplots depicting correct responses per Morph Type separately for musicians and non-musicians	95
6.4. Mean proportion of correct classifications per Emotion and Morph Type	96
6.5. Confusion matrices for each Emotion for the three Morph Types	96
7.1. ERPs separately for Emotion and Morph Type – fronto-central ROI	110
7.2. ERPs separately for Emotion and Morph Type – centro-parietal ROI	110
7.3. Relationship between music perception abilities (PROMS) and ERP amplitudes .	111
7.4. Happiness – Scalp topographies of the contrast between F0 and timbre	111
7.5. Sadness – Scalp topographies of the contrast between musicians and non-musicians, averaged across morph types	112
7.6. Happiness – Scalp topographies of the contrast between musicians and non- musicians for F0-Timbre difference waves	112
7.7. Mean proportion of correct responses per Emotion and Morph Type	113
B.1. Channel locations of the 64-channel setup used for the EEG experiments reported in Chapter 5 and 7	151
B.2. Effect sizes of the F0 vs. Timbre contrast for the P200 and N400 in different subsets of trials	152
B.3. Confusion matrices for each emotion for the three morph types – musicians only	153
B.4. Confusion matrices for each emotion for the three morph types – non-musicians only	153
B.5. Correlation between emotion classification performance and music perception abilities (PROMS)	154
B.6. Correlation between emotion classification performance and self-rated music skills (Gold-MSI)	155
B.7. ERPs separately for Emotion and Morph Type - centro-parietal ROI [-200, 1000]	156

C.1. Illustration of created voice morphs	157
C.2. Proportion of correct Classifications for each Emotion at different Morph Levels .	159
C.3. Patterns of misclassifications for each Emotion separately for the four Morph Levels	160
C.4. Mean Intensity Ratings for each Emotion at different Morph Levels	161

List of Tables

3.1. Summary of Empirical Articles on Musicality and Vocal Emotion Perception that met the Selection Criteria.	37
4.1. Acoustic properties of the stimulus material used in this study	57
4.2. Results of the 3 x 4 x 3 x 2 mixed-effects ANOVAs on mean ratings of Naturalness and Emotionality	59
4.3. Results of the regression analyses using cumulative link mixed models	62
6.1. Characteristics of participants - Demography, personality, and musicality	94
6.2. Correlations between vocal emotion recognition and music perception performance	97
6.3. Correlations between vocal emotion recognition and self-rated musicality	97
7.1. Results of the 3 x 4 x 3 mixed-effects ANOVAs on mean amplitudes of the P200, the N400 and the LPP	109
A.1. Summary of the acoustic characteristics of female voice morphs separately for each Emotion, Morph Type and Reference Type	141
A.2. Summary of the acoustic characteristics of male voice morphs separately for each Emotion, Morph Type and Reference Type	142
A.3. Summary of key mappings to (a) emotions and (b) pseudowords. Participants were assigned to key mappings via their participation number	143
A.4. Descriptive data of questionnaires	144
A.5. Pearson correlations between questionnaire data and vocal emotion recognition performance and for each Morph Type separately	145
A.6. List of reported instruments by musicians and non-musicians	146
A.7. Socioeconomic background of participants	147
A.8. Characteristics of participants - Demography, personality, and musicality (EEG Study - Chapter 7)	148
A.9. Summary of response key mappings to emotions	149
A.10. Participant assignment to the different response key mappings	149

1. Zusammenfassung (Summary)

Unsere Emotionen werden durch den Klang unserer Stimmen hörbar. Die **emotionale Prosodie** während des Sprechens ist dabei mehr als nur unterstützendes Beiwerk zum gesprochenen Inhalt, sondern ein wichtiger Transmitter non-verbaler Signale. Bei der Aussage *“Da bist du ja endlich!”* macht es beispielsweise einen entscheidenden Unterschied, ob sie mit einer fröhlichen oder wütenden Stimme gesprochen wird. Eine adäquate und effiziente Wahrnehmung von Emotionen in der menschlichen Stimme ist daher von großer Bedeutung im alltäglichen Miteinander.

Vorangegangene Untersuchungen haben gezeigt, dass Menschen in der Lage sind, die emotionale Qualität von vokalen Äußerungen scheinbar mühelos zu erkennen. Dieser Erkennungsleistung liegt eine sehr effiziente und automatische Verarbeitung der **akustischen Stimmenparameter** zugrunde, die sich in Abhängigkeit von emotionalen Zuständen verändern. Die wichtigsten Stimmenparameter sind dabei die Tonhöhe/Melodie, die Klangfarbe, die Lautstärke und der zeitliche Verlauf einer Äußerung. Trotz großer empirischer Anstrengungen waren jedoch bisherige Versuche, verschiedene Emotionen durch distinkte “akustische Profile” zu beschreiben, nur teilweise erfolgreich. Zum einen sind die Befunde sehr heterogen, zum anderen lassen sich auf Basis der untersuchten Stimmenparameter meist eher Aussagen zur unspezifischen emotionalen Erregung als zur spezifischen Differenzierung verschiedener Emotionen treffen. Darüber hinaus ist bisher unzureichend bekannt, wie verschiedene Klangparameter von Hörenden tatsächlich genutzt werden, um Emotionen in der Stimme zu erkennen. Bei dieser Frage setzt die vorliegende Dissertation an.

Der Fokus dieser Arbeit liegt auf dem relativen Einfluss der **Tonhöhe/Melodie** und der **Klangfarbe** auf die neuronale Verarbeitung und Erkennung von Emotionen in der Stimme. Moderne Stimmenmorphing-Technologie ermöglicht seit kurzem eine präzise Kontrolle dieser Stimmenparameter und wurde daher eingesetzt, um Aufnahmen von kurzen Äußerungen zu erstellen, die vier Emotionen (Freude, Genuss, Angst und Trauer) nur durch die Tonhöhe, nur durch die Klangfarbe oder durch beides ausdrücken. Diese wurden anschließend nicht nur für Verhaltensexperimente, sondern auch für Untersuchungen mit Elektroenzephalogramm (EEG) eingesetzt. Die Ergebnisse zeigen, dass sowohl die Tonhöhe als auch die Klangfarbe wichtige Informationen über die emotionale Qualität von Stimmen signalisieren, wobei jedoch die Tonhöhe insgesamt der dominantere Parameter zu sein scheint. Der spezifische Einfluss hängt jedoch von der emotionalen Kategorie ab: Bei Emotionen mit hoher Erregung – Freude und Angst – ist die Dominanz der Tonhöhe stärker, während bei Emotionen mit geringerer Erregung – Genuss und Trauer – die Beiträge von Tonhöhe und Klangfarbe ausgewogener sind. Die EEG-Daten zeigen parameter-spezifische Modulationen von neuronalen Prozessen, welche mit der Extraktion von

Emotionalität aus dem akustischen Signal und kognitiven Aspekten wie z.B. expliziter Entscheidungsbildung assoziiert sind.

In einem zweiten Schritt wurden individuelle Unterschiede untersucht, speziell im Hinblick auf auditorische Expertise und **Musikalität**. Dabei wurde zunächst der aktuelle Forschungsstand zum Zusammenhang zwischen Musikalität und Emotionsverarbeitung in Stimmen in einem systematischen Review zusammengefasst. Anschließend wurden die Erkennungsleistung und die EEGs einer Gruppe (semi-)professioneller MusikerInnen mit einer Gruppe von Nicht-MusikerInnen verglichen. Die Daten weisen auf eine besondere Bedeutung der Tonhöhe/Melodie für MusikerInnen hin: MusikerInnen zeigten besser Erkennungsleistungen als Nicht-MusikerInnen, wenn die Emotionen nur durch die Tonhöhe ausgedrückt wurden, aber nicht, wenn sie durch die Klangfarbe ausgedrückt wurden. Darüber hinaus zeigte sich, dass der Zusammenhang zwischen musikalischem Hörvermögen und der Emotionserkennung in Stimmen unabhängig von formaler Musikausbildung bestehen bleibt, was auf eine Prädisposition zur effizienten Nutzung melodischer Muster sowohl in der Musik als auch in Stimmen bei musikalisch begabten Personen hindeutet. Obwohl die EEG-Muster weniger schlüssig waren, deuten sie darauf hin, dass Musikalität auch die neuronale Antwort auf Emotionen in der Stimme zu modulieren scheint.

Abschließend widmete sich diese Arbeit noch der kritischen Auseinandersetzung mit der Stimmenmorphing-Technologie und daraus resultierenden Implikationen für die akustische Qualität des Stimmenmaterials. Es hat sich gezeigt, dass die akustische Manipulation der emotionalen Stimmen auch deren **wahrgenommene Natürlichkeit** beeinflusst, da sie verzerrt und weniger menschenähnlich klingen. Diese akustischen Verzerrungen können die ökologische Validität der empirischen Befunde einschränken, besonders wenn es systematische Unterschiede zwischen der Tonhöhe- und der Klangfarbe-Bedingung gibt. Um diesem Problem zu begegnen, wurden verschiedene Arten des Stimmenmorphings verglichen. Anschließend wurde in einer Ratingstudie die wahrgenommene Natürlichkeit erhoben und ihr Einfluss auf die Emotionswahrnehmung untersucht. Dabei zeigte sich, dass das Stimmenmorphing die Natürlichkeit der produzierten Stimmen zwar beeinflusst, sich die Wahrnehmung der Emotionen jedoch als bemerkenswert robust gegenüber diesen Verzerrungen erweist. Die EEG-Daten könnten davon hingegen stärker beeinflusst werden. Insgesamt präsentiert diese Arbeit damit überzeugende Befunde, dass Stimmenmorphing ein valides Instrument für die Erforschung von Emotionen in der Stimme darstellt, wenn es mit einem kritischen Bewusstsein für seine Grenzen und Probleme zum Einsatz gebracht wird.

Die vorliegende Dissertation liefert wichtige und neue Erkenntnisse über die Wahrnehmung von Emotionen in der menschlichen Stimme, sowohl auf empirischer als auch auf konzeptioneller Ebene. Die zentralen Beiträge beziehen sich dabei auf die Rolle der zugrundeliegenden akustischen Parameter, die elektrophysiologischen Korrelate, und den Einfluss individueller Unterschiede mit einem spezifischen Fokus auf Musikalität. Auf diese Weise trägt diese Arbeit zum Verständnis mehrerer komplexer und wichtiger Eigenschaften bei, welche uns als Menschen ausmachen: dem Gebrauch unserer Stimmen, dem Ausdruck unserer Emotionen und unserer Fähigkeit, zu Musizieren.

2. Introduction

“The voice is one of the prime channels for the expression of emotion, a fact that has been commented upon ever since the ancient school of rhetoric. It can be reasonably argued that the phylogenetic continuity of vocalization as a medium of emotion expression provides important information for the emergence of speech and music in the human species.” (Scherer, 2018, p. 61)

Emotions form an essential part of human experience and behavior. They are commonly understood as prompt, intense, and multilayered reactions to relevant environmental changes (Juslin & Laukka, 2003; Rothmund & Eder, 2011). These reactions include cognitive appraisal, physiological changes, subjective feeling, and behavioral outcomes. In principle, human experience and expression of emotion is not tied to the presence of others (Bachorowski & Owren, 2003; Scherer, 1986). For example, being alone outside during a thunderstorm can result in an intense feeling of fear and trigger verbal exclamations such as screams. However, without doubt, emotional communication and regulation between individuals are fundamental to human co-existence and social interaction in complex societies (Bachorowski & Owren, 2003; Scherer, 1986). A smile can help to distinguish between friend and foe, a scream in terror can alert immediate attention to a source of threat, and crying can trigger empathetic and caring behavior in others. To this end, emotional expressions ensure human survival by helping to navigate through a complex social world.

A powerful transmitter of emotions is **the human voice**. Sounds can travel long distances and do not require visual contact (Schirmer & Adolphs, 2017). The human voice therefore provides an emergency warning system that presumably shaped vocal expression and perception of emotions very early in the phylogenetic development of humans (Scheiner & Fischer, 2011; Scherer, 2018). Today, humans express and perceive vocal emotions seemingly effortlessly and automatically in everyday life (Bachorowski & Owren, 2003; Lima et al., 2019). Vocal emotions are further recognized across cultures, suggesting an innate predisposition shared by mankind (Laukka et al., 2016; Scheiner & Fischer, 2011; Scherer, 2018). Therefore, most humans are considered “experts for vocal emotions”, who do not seem to require conscious effort for accurate emotion perception (Chartrand et al., 2008; Lima et al., 2019). However, the tremendous importance and computational complexity of this seemingly effortless processes become painfully apparent in individuals whose vocal emotional processing is disrupted, for example as a result of hearing loss, brain damage, or a variety of mental disorders (Belin et al., 2011; J. A. Christensen et al., 2019; Nilsson & Sundberg, 1985). Deficits in the decoding of vocal emotional signals have been consistently linked to depression, reduced well-being, poor social-emotional adjustment and interpersonal difficulties (Blonder et al., 2012; Carton et al., 1999; Naranjo et al., 2011; Neves

et al., 2021). This great impact on an individuals' quality of life has motivated decades of research to face the key challenge in understanding vocal emotion perception: to uncover the complex mechanisms behind an apparently "easy everyday task".

2.1. The expression and perception of emotions in voices

The scientific investigation of vocal emotions goes back to Charles Darwin's "The expression of emotions in man and animals" (Darwin, 1872), and has been of great interest to researchers from various disciplines ever since. Today, it is widely acknowledged that humans can recognize emotions within and across cultures (Scherer, 2018; Thompson & Balkwill, 2006), with consensus across different emotion theories including the basic emotion, the dimensional and the appraisal accounts (Bachorowski, 1999; Banse & Scherer, 1996; Juslin & Laukka, 2003; Russell, 1980). An important distinction has to be made between non-verbal vocalization (e.g. laughter, cries, or moans) and emotional prosody of speech (Pell et al., 2015). In this work, I will focus on the latter.

Emotional prosody is not only a supplementary byproduct of speech, but itself a carrier of important information (Brück et al., 2011). For example, whether the utterance "There you are!" is spoken in a happy or an angry tone will make a substantial difference to the listener. In the following two paragraphs, I will elaborate in more detail on two important research branches that strive to understand the expression and perception of emotional prosody. The first covers the manifold efforts to decipher its "acoustic code" (Bachorowski, 1999). The relative ease with which humans seem to perceive vocal emotions has driven the idea that different emotional categories may be expressed by **distinct profiles of acoustic patterns**. This search for acoustic profiles has motivated more than 30 years of research, which I will summarize in paragraph 2.1.1. The second considers the time course of emotional prosody processing in the brain. Unlike other speaker characteristics such as age or sex ¹, emotional quality of an utterance can change rapidly within an encounter (A. W. Young et al., 2020). Hence, an efficient perceptual system has to be very sensitive to the unfolding of emotional cues over time (Paulmann & Kotz, 2018). Electroencephalography (EEG) is an excellent tool to study the time course of **vocal emotional processing in the brain**, because of its high temporal resolution. Therefore, in paragraph 2.1.2, I will summarize how EEG research has shed light on the brain mechanisms that transform sounds into emotional significance, and subsequently lead to cognitive and behavioral responses.

¹Note that "sex" and "gender" are frequently used synonymously in the voice perception literature. However, latest APA publication guidelines recommend using sex when referring to the biological construct and gender when referring to the social one (American Psychological Association, 2020). Research on voice perception and production usually focuses on biological aspects, because of the strong sexual dimorphism in the vocal production system which affects the acoustics of voices (Ladefoged, 1996). Therefore, I will refer to vocal sex in this thesis throughout, although some cited publications may have used the term vocal gender.

2.1.1. Vocal expression – deciphering the acoustic code of emotional prosody

“The cause of widely different sounds being uttered under different emotions and sensations is a very obscure subject. Nor does the rule always hold good that there is any marked difference.” (Darwin, 1872, p. 85)

Speaking involves well-coordinated movements of numerous muscles in the human body (Schweinberger et al., 2014). A sound is produced by air that is pressed out of the lungs through a narrow opening in the larynx, causing the vocal folds to vibrate in a quasi-periodic manner. This sound is further modified in the vocal tract where fast moving articulators including the tongue, velum, teeth, and lips affect its spectral features (Fant, 1970; Ladefoged, 1996). The resulting speech output, which is transmitted through the air and becomes perceivable by others, can be described by means of different acoustic parameters (Ladefoged, 1996; Scherer, 2018): The vocal fold vibration can be measured as **fundamental frequency (F0)** and is commonly perceived as pitch. The F0 depends on the tension, mass, and length of the vocal folds, which varies within and between speakers (Ladefoged, 1996). For neutral voices, average F0 is about 100 to 120 Hz in male speakers and 200 to 240 Hz in female ones. However, the F0 of emotional voices can be much higher (> 250 Hz in both sexes, refer to Tables A.1 and A.2). F0 unfolds over time, resulting in a dynamic pitch contour, also referred to as voice melody. The **amplitude** of the sound relates to its perceived loudness, while **temporal characteristics** relate to the perception of time-related features such as duration, speech rate, or pausing. Finally, vocal **timbre** includes all spectral features that relate to the perception of voice quality, e.g. a “harsh” or “soft” tone of voice. Vocal timbre is affected by vocal features like formant frequencies, high-frequency energy or harmonics-to-noise (HNR) ratio (Juslin & Laukka, 2003).

The emotional state of the speaker becomes audible because emotion-related physiological and cognitive changes affect virtually all components of the speech production system and its acoustic output, respectively. In a state of raging anger, for example, the heart rate and blood flow would rise, and increased muscle tension would affect the pressure in the lungs, the vocal fold vibration and the movements of the vocal tract (Scherer, 1986). This would result in a loud, harsh and high-pitched voice. In addition to such physiological factors, there is a voluntary aspect in the acoustic modification of uttered sounds that is considered unique in humans, and that has been shaped through socialization (Juslin & Laukka, 2003; Scherer, 1986). Hence, speakers have some degree of control to adjust emotional expressions according to their intentions (Darwin, 1872), e.g. using a soft and high voice to appear friendly. In short, emotions change the sound of our voices, as they are imprinted in the acoustic features of an utterance. These acoustic features are in turn picked up by listeners to infer the emotional quality - in a rapid, mostly accurate, automatic, and seemingly effortless manner (Lima et al., 2019; Paulmann & Kotz, 2018).

This observation has led to the assumption that discrete emotions may be expressed by distinct profiles of acoustic cues (Bachorowski, 1999; Scherer, 1986, 2018). However, early and mostly explorative research efforts were inconclusive. Scherer (1986) was the first to formulate specific predictions for the acoustic consequences of different emotional states, based on assumptions

about the underlying physiological and cognitive processes. His theoretical claims were empirically supported by Banse and Scherer (1996) and later in an extensive meta-analysis by Juslin and Laukka (2003). Essentially, happiness, fear and anger can be characterized by an increased fundamental frequency, amplitude and speech rate, whereas the opposite holds for sadness (Banse & Scherer, 1996; Juslin & Laukka, 2003; Thompson & Balkwill, 2006). Beyond this general pattern, however, enormous heterogeneity emerged across studies. On a closer look, this pattern seems rather disappointing, as the acoustic pattern seems to distinguish unspecific arousal rather than discrete emotional states (Brück et al., 2011). This is clearly at odds with the behavioral performance of listeners, who seem to have little problems with distinguishing emotional states, as well as arousal and valence of voices (Scherer, 1986, 2018). In the literature, this has been acknowledged as an apparent paradox between listeners’ ability to infer discrete emotions and the insufficient identification of vocal parameters that reliably differentiate them (Bachorowski & Owren, 2003).

How can this paradox be resolved? Answers may be found at both the conceptual and the methodological level. Conceptually, the assumption of a static set of acoustic cues to be diagnostic for different emotions may be too simplistic, as this view neglects one of the fundamental properties inherent to human perception and expression: flexibility. This flexibility is more adequately captured in Brunswik’s lens model (Brunswik, 1956). According to this model, vocal emotions are assumed to be encoded by a large number of vocal cues of which none is perfectly reliable in itself, but which combine in a probabilistic and partially redundant manner. As they are partly interchangeable, listeners can use these cues in a flexible way to infer the emotional quality (Laukka et al., 2016; Thompson & Balkwill, 2006; Thompson et al., 2010). In line with this idea, Spackman et al. (2009) identified very different styles of vocal expression across individuals, which were nevertheless equally well recognized. This interindividual variance is lost in studies which use only one speaker model, which may explain some inconsistent findings (Scherer, 1986).

At the methodological level, several aspects deserve consideration. First, undesired heterogeneity may be caused by an insufficient specification of emotional states and different protocols of vocal sampling (Bachorowski, 1999; Scherer, 1986). For example, Scherer (1986) argued that “hot” and “cold” anger would be associated with different acoustic patterns, which would appear as contradictory when treated as one emotional category. Concerning sampling protocols, a great debate has evolved around the question whether actor portrayals of posed emotions are valid or whether recordings of induced “real” emotions should be used only (Bachorowski & Owren, 2003; Banse & Scherer, 1996; Spackman et al., 2009), as they could differ systematically in their acoustic profiles (Scherer, 2013, 2018). However, for practical reasons, actor portrayals are still widely used and accepted (Scherer, 2018). Another issue concerns an insufficient coverage of relevant acoustic cues, which may have been limited simply by the technical possibilities in early works. Scherer (1986) specifically criticized the widespread neglect of voice cues related to timbre. One may assume that when fundamental frequency, amplitude and timing failed to differentiate emotional valence, important cues expressed by voice timbre may have been missed. With modern speech analysis tools, such hypotheses can now be addressed (Arias et al., 2021; Gobl, 2003).

Finally, and most importantly, the majority of research linking acoustics to emotions is only correlational and does not allow for causal inferences: Usually, acoustic cues were measured for a set of emotions and subsequently compared or used to predict listeners' responses using regression analyses (Juslin & Laukka, 2003; Scherer, 2018). However, a strong predictive power of an acoustic cue does not mean that listeners actually use it for emotional inferences. For example, smiling is strongly associated with increased F0, but can be reliably identified in unvoiced whispers as well, suggesting that listeners do not necessarily rely on these F0 cues (Tartter & Braun, 1994). Recently, this has motivated an explicit call to incorporate voice manipulation tools into the study of vocal emotion perception, to gain experimental control over acoustic features (Arias et al., 2021). Modern speech manipulation tools now provide ample possibilities to put these recommendations into practice. Although these new technologies come hand in hand with new challenges, which I address in more detail in section 2.3 and Chapter 4, they allow to manipulate separate parameters in the voice space with unprecedented rigor.

The aim of this thesis was to address some of the methodological limitations identified above. First, I used a novel **voice manipulation technique** called parameter-specific voice morphing, which enables resynthesis of voices that express an emotion via isolated vocal cues only (Kawahara & Skuk, 2018; Kawahara et al., 2008). In section 2.3, I will elaborate on this technology in more detail. Second, I focused specifically on the **role of fundamental frequency vs. timbre**, while holding timing and amplitude constant. By expanding our understanding of timbre in vocal emotional processing, I aim to fill a knowledge gap on a parameter that is underrepresented in the literature. Of importance, I probe the **relative** rather than absolute contribution of these cues. To this end, I explore the limits of the perceptual flexibility proposed in Brunswik's lens model, i.e. the degree to which unique emotional information is expressed in isolated vocal parameters, which cannot be compensated by other ones. Manipulation of isolated cues deconfounds any intercorrelation of parameters that may be inherent to natural emotions, and thus allows to study their unique contribution. For example, if emotion recognition declines when timbre is the only diagnostic cue and F0 is rendered uninformative, one would conclude that F0 carries unique emotional information, which cannot be compensated by timbre. By contrast, if emotion recognition with either timbre or F0 only would be comparable and above chance, this would mean that they both carry diagnostic and partly interchangeable information.

2.1.2. Electrophysiological correlates of vocal emotional processing

Listeners benefit from quickly grasping vocal emotions, because emotional prosody is inherently dynamic and can change from one instance to another (Paulmann & Kotz, 2018). This may explain why the processing of vocal emotions is underpinned by a time-critical neural architecture. With its high temporal resolution, electroencephalography (EEG) is a particularly well-suited measure to unravel the time-course of vocal emotional processing in the brain. According to a model proposed by Schirmer and Kotz (2006), vocal emotion processing constitutes of multiple steps that can be linked to different ERP (event-related potential) components.

The first step is an initial analysis of acoustic features of the sound, which is reflected in a modulation of the N100, a negative going wave peaking approximately 100 ms following stimulus onset. This component is affected by salient acoustic cues such as pitch and loudness of sounds (Paulmann & Kotz, 2018; Schirmer & Kotz, 2006). In a second step, these acoustic features are integrated to derive emotional significance. These emotional processes unravel as early as 200 ms past voice onset and have been linked to the P200 component (Paulmann & Kotz, 2008; Paulmann et al., 2013; Schirmer & Kotz, 2006; Schirmer et al., 2013). It has been debated whether these P200 effects truly reflect emotional processes or just sensory-driven activity, since the P200 is also reliably modulated by acoustic features (Paulmann et al., 2013; Schirmer et al., 2013). However, evidence that emotional processing indeed takes place at this point in time comes from the Mismatch Negativity (MMN). The MMN is a component that also peaks around 200 ms, in response to an unexpected stimulus change, i.e. to a deviate in a series of standard sounds (Schirmer, Striano & Friederici, 2005). Crucially, the MMN-effect is calculated based on responses to acoustically identical stimuli, which served as standards in one, but deviates in another condition. This MMN-effect was bigger for emotional compared to neutral vocalizations, suggesting an early attentional shift to the emotional significance of sounds (Schirmer & Escoffier, 2010). This detection of emotional meaning is followed by a third step including higher-order and potentially more effortful processes such as goal-directed processing, semantic integration and response preparation (Schirmer & Kotz, 2006). These have been linked to modulations of the P300, the N400 and the late positive potential (LPP, Hajcak & Foti, 2020; Paulmann & Kotz, 2018).

Note that most of the research on the time-course of vocal emotion processing focused on the contrast between neutral and emotional sound. Thus, there is little and inconsistent data informing us about how and when different emotional states can be distinguished by the brain (for a meta-analysis in the spatial domain, refer to Fusar-Poli et al., 2009). Some publications report that differentiation of emotional categories takes place around the P200 already (Frühholz & Schweinberger, 2021; Paulmann et al., 2013), others suggest a later point in time (Paulmann & Kotz, 2008).

While the basic time-course of vocal emotion processing is well-supported by empirical findings, two questions remain open: The first concerns the **specific role of acoustic features**. All current models on the neural structure underlying vocal emotion perception emphasize the monitoring and integration of emotional cues in real time (Frühholz et al., 2016; Grandjean, 2021; Schirmer & Kotz, 2006). However, although some ERP components have been shown to be reliably modulated by acoustic features, it is still unclear how this takes place for specific parameters in different emotions (Paulmann & Kotz, 2018). In EEG paradigms, a key challenge is the difficulty to disentangle acoustically driven vs. emotional processes, as acoustic features and emotional quality are confounded in natural voices (Paulmann et al., 2013). This may be achieved by using acoustically manipulated voices. For vocal emotions, to the best of my knowledge, such manipulation techniques have never been employed in the context of an EEG experiment. Therefore, in two experiments (Chapter 5 and Chapter 7), I explored the temporal processing of voices which expressed emotional quality through F0, timbre, or both.

The second question relates to **individual differences between listeners**. Several ERP findings suggested an effect of listener sex on the pre-attentive processing of vocal emotions. Females show a larger MMN to emotional than to neutral deviates (Schirmer, Striano & Friederici, 2005). Further, in a cross-modal priming study, females made earlier use of emotional prosody information for subsequent semantic word processing than males (Schirmer et al., 2002). However, this difference was no longer visible after directing explicit attention to the prosodic information (Schirmer, Kotz & Friederici, 2005), suggesting that task instructions and direction of attention modulate ERP responses to vocal emotions. Language-specific modulations on the time-course of vocal emotion perception were observed in a behavioral gating paradigm, with a reliable own-language advantage in groups of English and Hindi listeners (Jiang et al., 2015). In another behavioral study, age-related decline in vocal emotion perception was linked to pitch perception problems in older adults (R. L. C. Mitchell & Kingston, 2014), which may be reflected in early ERPs related to acoustic analysis of emotional voices. Differences in auditory processing styles and abilities, however, are rarely studied. The role of musical expertise, for example, is insufficiently understood, despite the close link between emotions in music and voices. While Strait et al. (2009) reported different patterns of brainstem activation in musicians compared to non-musicians in response to an unhappy infant's cry, findings on the cortical level are inconsistent and mostly failed to detect reliable group differences for vocal emotions (I. Martins et al., 2022; Pinheiro et al., 2015; Rigoulot et al., 2015). As a limitation, null findings in several studies could be attributed to small sample sizes ($N < 15$, Pinheiro et al., 2015; Rigoulot et al., 2015), illustrating the need for more systematic and well-powered studies on how musical expertise affects the time course of vocal emotion processing in the brain. The results of such a study will be reported in Chapter 7. In the next section, I will review the close relationship of emotions in music and voices that motivated the comparison of musicians and non-musicians in this dissertation.

2.2. Music and vocal expression

2.2.1. Emotions in voices and music – a tale of joint evolution

Voice and music are both powerful means of auditory expression, with a degree of voluntary control that is thought to be unique to humans (Juslin & Laukka, 2003). The intriguing parallels and interconnections between both channels that we observe today take us back to a consideration of the early roots of human communication (Mehr et al., 2019). The theory of evolution spawned the thought that music and vocal expression, including both emotional prosody and speech, share a common origin (Darwin, 1872).

One hypothesis is that music arose as a means to imitate the human voice (Mithen et al., 2006), proposing the voice as one of the oldest instruments of humankind. In line with this idea, researchers have highlighted the voice-like character of many musical timbres and argued that the human voice is among the most expressive of instruments (Akkermans et al., 2019; Juslin & Laukka, 2003). However, reducing music to its voice-like features is too simplistic, as music goes far beyond the acoustic possibilities of the human voice. There are many features of music that

have no counterpart in voices, such as harmonic progression (Juslin & Laukka, 2003). Therefore, a modified version of this hypothesis is that music and voice rather developed in parallel, but still root in the same basic form of vocal utterances used by pre-literate societies. Darwin called this basic form of communication a “prosodic protolanguage” (Darwin, 1871). Presumably, it consisted of prosodic non-verbal exclamations and was used for courtship, to promote parent-infant bonding, and the transmission of emotions (Fitch, 2013; Thompson et al., 2012). Emotional expression may therefore be one of the core features shared by music and voices, linked through their common origin. Accordingly, it has been suggested that music primarily developed as a means to harmonize emotions and create social cohesion in groups (Juslin & Laukka, 2003; Schäfer et al., 2013). Until today, emotional regulation is among the most frequently reported reasons why people engage in musical activities (Schäfer et al., 2013). Likewise, as for the vocal domain in today’s literate societies, emotional expressions continue to be more than just a byproduct of human speech and their importance for both speakers and listeners is widely acknowledged (Scherer, 2018; Schweinberger et al., 2014).

Although the true origin of music and vocal expression may not be conclusively resolved, there is a high degree of consensus about a joint evolution of vocal and musical emotions. Today, evidence for a close link between emotions in music and voices originates from three different lines of research: cross-cultural studies, exploration of acoustic patterns, and neuroscientific research highlighting overlapping neural networks.

Cross-cultural studies suggest that people can detect emotions in music and voices across cultures well above chance (Balkwill & Thompson, 1999; T. Fritz et al., 2009; Laukka et al., 2016; Thompson & Balkwill, 2006; Thompson et al., 2010). This cross-cultural recognition suggests some degree of innate representation of musical and vocal emotions shared by all humankind (Scheiner & Fischer, 2011). Nevertheless, there is still consistent evidence for an own-culture advantage, both in music and voices, suggesting that the expression and perception of auditory emotions undergo an enculturation process, shaped in early childhood (Hunter & Schellenberg, 2010; Laukka et al., 2016; Morrison & Demorest, 2009).

The **exploration of acoustic patterns** revealed that emotions are expressed in voices and music by similar acoustic codes (Hunter & Schellenberg, 2010; Juslin & Laukka, 2003). In fact, manipulations of tempo, amplitude and pitch lead to similar changes in perceived emotionality in voices and music (Ilie & Thompson, 2006). Parallels are also observed in recognition patterns, with anger and sadness usually being identified better than other emotions (Thompson & Balkwill, 2006; Thompson et al., 2010). However, despite these similarities, there are also marked differences: for emotional voices, the “composer” and the “performer” are usually the same person, whereas in music they may diverge, at least in Western music cultures. Thus, the expressed musical emotion results from a combination of features introduced by the composer (e.g. harmony, tonality, and instrumentation) and some degree of freedom by the performer (e.g. on timbre, loudness, and tempo, Schutz, 2017). Thus, despite their acoustic similarities, emotions in music and voices may result from different production mechanisms.

Mechanisms involving emotion perception, however, share many common features again. Neuroscientific research identified largely **overlapping neural networks** for the processing of emotional voices and music (Aubé et al., 2015; Escoffier et al., 2013; Fröhholz et al., 2014, 2016; Schirmer et al., 2012). The core network for the processing of auditory emotions includes the auditory cortex, the superior temporal sulcus, frontal areas, the amygdala, the insula and the cerebellum (Fröhholz et al., 2016). Taken together, these findings suggest a strong link between voices and music, with intriguing parallels regarding the expression and perception of emotions.

2.2.2. Links between musicality and non-musical skills

The capacity to make music requires a high degree of auditory, sensory and motor precision. It takes several years of training, which usually starts in childhood, to achieve a professional or solid amateur level. The high demands of musical activities have promoted the idea that musicians may also excel at other activities. Such an influence from the musical onto other domains is called transfer. In the literature, a distinction is made between close transfer, which occurs in domains closely related to music such as pitch perception, and far transfer, which relates to relatively distant domains, such as working memory (M. Martins et al., 2021). Note that the distinction between close and far transfer may be gradual rather than binary, and that both forms of transfer have been investigated thoroughly:

There is ample evidence that musicians are **auditory experts** (Kraus & Chandrasekaran, 2010). They are more sensitive to pitch, timbre, temporal patterns, intensity, and harmonic differences of musical sounds, and also excel at music-in-noise perception, auditory attention, and identification of musical emotions, compared with non-musicians (Bhatara et al., 2011; Kraus & Chandrasekaran, 2010; Lima & Castro, 2011). This auditory sensitivity seems to extend into the vocal domain, where robust **music-to-speech transfer** effects have attracted great scientific attention. Compared to non-musicians, musicians exhibit enhanced vowel and phoneme discrimination, tracking of language and metric structures of sentences, segmental processing of speech sounds, pitch processing in tonal languages, and speech-in-noise recognition performance (Elmer et al., 2018; Hallam, 2017; Schellenberg, 2016). Intriguingly, benefits were also observed in non-auditory language tasks, such as word memory, syntax and grammar processing, vocabulary, and reading skills (Chartrand et al., 2008; Elmer et al., 2018; Hallam, 2017). These findings suggest shared processes between music and language, which are not restricted to the auditory modality.

Transfer to non-verbal vocal signals, however, are less well understood (M. Martins et al., 2021). Musicality has been linked to superior voice timbre processing (Chartrand & Belin, 2006) and a benefit for vocal emotional processing (M. Martins et al., 2021). Empirical evidence for the latter, however, is quite heterogenous, and will be discussed in more detail below and in Chapter 3.

Benefits of musicality have also been extensively examined in **non-auditory cognitive domains**. Such benefits were observed for executive functions (i.e. response inhibition, selective attention), multimodal integration, short-term, long-term and working memory, intelligence,

academic achievement, and spatial abilities (Schellenberg, 2016). In terms of personality, musicians seem to display higher values of openness (Schellenberg, 2016). However, there is no clear evidence for benefits in socio-emotional abilities outside of the auditory domain (Farmer et al., 2020; Schellenberg, 2016).

Finally, neuroscientific research devoted much effort to the exploration of **brain differences** related to musicality. Musicians are a popular model of neuroplasticity, in order to study how years of dedicated training may shape the brain (Herholz & Zatorre, 2012; Kraus & Chandrasekaran, 2010; Pantev & Herholz, 2011). Musicality is associated with widespread functional and structural changes, such as larger grey matter volume, increased activity in auditory, somatosensory and motor areas, enhanced functional connectivity in auditory-motor networks, and stronger cortical responses to auditory stimuli such as music and speech (Chartrand et al., 2008; Hallam, 2017; Kraus & Chandrasekaran, 2010; Palomar-García et al., 2017; Pantev & Herholz, 2011; Strait et al., 2009). Differences between musicians and non-musicians can be found already in the brainstem, where musicians display a more robust and faithful representation of pitch, harmonic components, and timing information (Kraus & Chandrasekaran, 2010). Importantly, the increased response to auditory stimulation in musicians’ brains is not simply a “volume-knob” effect (Kraus & Chandrasekaran, 2010). Instead, brainstem activity suggests a selective enhancement of the most meaningful information, whereas irrelevant cues are suppressed. This was demonstrated by Strait et al. (2009), who observed increased responses to a spectrally complex portion of a sound, but reduced responses to the simpler part, compared with non-musicians.

Despite the extensive body of literature comparing musicians to non-musicians in a variety of auditory and non-auditory domains, insight into the causal role of musical training for the reported benefits is limited. It is widely acknowledged that musical skills emerge as a result of both aptitude (“nature”) and training (“nurture”), which are further assumed to interact in individuals. Twin studies suggest that musical abilities have a substantial genetic component (Schellenberg, 2016), but that does not deny the potential effects of training (A. D. Patel, 2011). Ultimately, this nature/nurture debate can only be resolved by longitudinal musical-training studies with randomized assignment and preferably an active control group. Unfortunately, due to their costly and time-consuming nature, these studies are rare. Hence, most of the existing evidence is cross-sectional. As remedy, other findings have been argued to shed light on the contribution of nature vs. nurture in transfer effects from music: correlations of a task benefit with years of musical lessons or the age at learning onset have been taken as evidence for a contribution of training (Hallam, 2017). Further, effects that were specific to the own instrument of an individual were assumed to reflect training-induced changes (Kraus & Chandrasekaran, 2010). Support for the impact of natural aptitude comes from studies using “naturally good musicians” (Correia et al., 2022), who display superior music perception abilities in the absence of any musical training. If these “natural musicians” perform equal to trained musicians in a task, then any observed difference between musicians and non-musicians cannot be driven entirely by musical training.

Considering such arguments together with the few existing longitudinal studies (Kraus et al., 2014), there is consensus that differences between musicians and non-musicians in brain responses, in the domain of speech, and in global auditory sensitivity reflect training induced changes to some degree (Elmer et al., 2018; Hallam, 2017; Kraus & Chandrasekaran, 2010), although this does not exclude potential differences in natural aptitude. Differences in non-auditory cognitive domains (e.g. intelligence), however, seem to be the result of pre-existing individual disposition (Schellenberg, 2016).

2.2.3. Musicality and vocal emotion perception

As described in the previous section, benefits of musicality have been researched thoroughly in the speech domain and non-musical cognitive abilities. Non-verbal vocal domains such as vocal emotion perception have received less attention. However, the strong link between music and vocal expression, especially with regard to emotion, makes a transfer from musical skills into the vocal emotional domain plausible. Indeed, previous research suggests that musical skills are linked to a benefit in vocal emotion perception (Lima & Castro, 2011). Further, on the lower end of the musicality spectrum, people with amusia (a specific impairment for the perception of music) seem to display a consistent disadvantage in vocal emotion perception (Lima et al., 2016; Thompson et al., 2012). However, findings are heterogenous and limited by methodological factors such as small sample sizes and confounding variables (Lima & Castro, 2011; M. Martins et al., 2021; Thompson et al., 2004).

Therefore, one objective of this dissertation was to provide a comprehensive overview over the current available literature on the link between musicality and vocal emotion perception (Chapter 3). In principle, the available evidence supports a link between musicality and vocal emotion perception abilities, but it has two important gaps, which I addressed in subsequent experiments (Chapters 6 and 7). The first is a limited understanding of the mechanisms underlying the musicality benefit for vocal emotions. While several studies emphasize the role of auditory sensitivity to the vocal features that express emotionality, it is not yet understood how auditory processing might differ between musicians and non-musicians. Specifically, it is unclear whether musicians excel in all aspects of auditory processing of vocal emotions, or whether they may be particularly tuned to specific acoustic features. This question is the main focus of Chapter 6, in which I used acoustically manipulated emotions to study the importance of different voice cues for the emotional judgements made by musicians and non-musicians. Second, the field lacks systematic neuroscientific experiments exploring differences in musicians' and non-musicians' brains that are related to vocal emotional processing. Emotional information in vocal expression evolves over time, and the EEG provides an excellent method to study time-critical brain responses. Therefore, in Chapter 7, I report on an EEG study that explored event-related potentials in response to acoustically manipulated emotions in musicians and non-musicians.

2.3. Voice morphing

Research on voice processing aims to provide valid insight into “natural”, i.e. “real-life” principles of human expression and perception. Paradoxically, this usually requires the utilization of highly controlled stimulus material to allow valid conclusions (although it may be noted that some researchers of voice identification propose an alternative perspective, cf. Lavan et al., 2019). How can such precisely controlled stimulus material be produced? One option is to synthesize artificial voices using computer algorithms, but the resulting utterances may not sound very human-like, which limits their ecological validity (Hajarolasvadi et al., 2020). Therefore, most scientists still resort to recordings of real human voices. Today, many online resources offer extensive databases controlled for speaker characteristics, speech material and recording environment (e.g. Burkhardt et al., 2005; Livingstone & Russo, 2018). While such databases offer high-quality stimulus materials for many purposes, they are still not sufficient for research questions which require the precise experimental control over the acoustic features of sounds. This can be achieved through further manipulation of the recordings, by means of **voice morphing** which balances the trade-off between experimental control and ecological validity by producing natural-sounding resynthesized voices with controlled acoustics. One tool that offers such functionality is **Tandem-STRAIGHT**² (Kawahara & Skuk, 2018; Kawahara et al., 2008). In what follows, I will outline the basic principles of voice morphing using Tandem-STRAIGHT, including its requirements and assumptions. Subsequently, I will elaborate on some of its applications in voice perception research, and vocal emotion perception in particular.

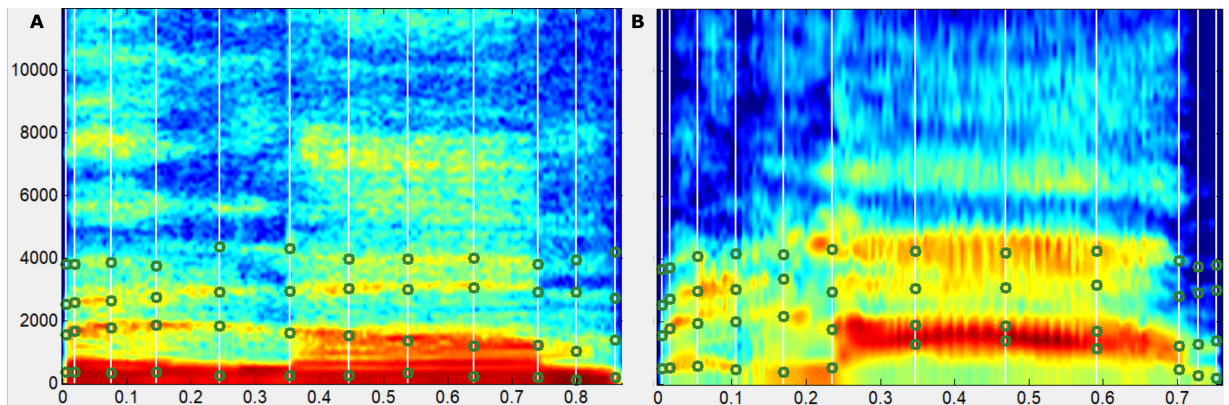
2.3.1. Voice morphing using Tandem-STRAIGHT – implementation, requirements, and assumptions

Tandem-STRAIGHT enables the creation of modified voices on the basis of original audio recordings, by altering some of their parameters (Kawahara & Skuk, 2018). In principle, voice morphing can be based on one, two, or many recordings. The most common form is the morphing on a trajectory between two voices, where morphs are resynthesized from the resulting continuum by using different parameter weights. Voice morphing always consists of three basic steps: First, the audio files have to be converted into a parametric representation (e.g. F0 contour, spectral features and timing). Then, these parameters are modified. Finally, the new parameters are used to resynthesize stimuli and convert them back to audio format. The first step, the conversion into a parameter representation, is the most challenging one. While the tool offers a quasi-automatic extraction of the fundamental frequency (F0), which usually requires only few manual adjustments, the division of the sound into its voiced and unvoiced proportions has to be entered by hand in a graphical user interface. Additionally, voice morphing requires manual mapping of time- and frequency anchors at key features of utterances, such as onset of a plosive, formant shifts, start and end of vowels, fricatives and nasals etc. (see Figure 2.1). The positions of these anchors must be congruent across all morphed stimuli, otherwise resulting morphs will

²STRAIGHT stands for **S**peech **T**ransformation and **R**epresentation using **A**daptive Interpolation of wei**G**H**T**ed spectrum (Kawahara et al., 1999)

be corrupted. Occasionally, it is necessary to cut small artifacts from the original recordings, such as little smacks or clicks. This has to be done with caution, but is explicitly recommended to enhance quality of the resulting morphs (Kawahara & Skuk, 2018). All these manual tasks are time-consuming, potentially subjective and error-prone. Therefore, the preparational work for voice morphing requires experience, some phonetic background knowledge, detailed documentation of all preprocessing steps, and a critical evaluation of stimulus quality.

Figure 2.1.: Time- and frequency anchors



Note. Visualization of the time- and frequency anchors of the pseudoword /belam/ uttered by the same female speaker with (A) neutral and (B) angry prosody. Time anchors are depicted as white lines and frequency anchors as grey dots. The x-axis shows the time scale in seconds, the y-axis shows the frequencies in Hz.

Once the parameter representation of voices is obtained, modification and resynthesis can be scripted and automatized. Tandem-STRAIGHT represents voices by means of five different parameters: *fundamental frequency* ($F0$), *spectrum level* (a representation of the spectral envelope), *aperiodicity* (a representation of aperiodic sound components), *spectral frequency* (the manually assigned frequency anchor positions), and *timing* (the manually assigned time anchor positions). In the literature, spectrum level, aperiodicity and formant frequency are often compiled together as timbre (Nussbaum, von Eiff et al., 2022; Skuk et al., 2015). This operationalization of timbre fits well with its formal definition as being “the difference between two voices of identical $F0$, intensity and temporal structure” (ANSI, 1973). These parameters can be modified in conjunction, called **full voice morphing**, or independently from each other, called **parameter-specific voice morphing**.

The rising number of studies making use of Tandem-STRAIGHT reflects the increasing popularity of this tool. However, it can only unfold its full potential if the assumptions and requirements stated by Kawahara and Skuk (2018) are met: First, generated voices without any changes to the parameter settings should sound virtually identical to the original audio input. In other words, the conversion of audio files into the parameter representation and their resynthesis should not introduce perceptual changes. Second, any combination of parameter weights should

not distort the sound quality of the voice. From the viewpoint of ecological validity, it is crucial that resynthesized voices appear natural in the sense that they “sound like an instance from a natural voice” (Kawahara & Skuk, 2018). This holds both for full as well as for parameter-specific voice morphs. In practice, however, this proves to be very challenging, especially when the parameter changes are extreme (Kawahara & Skuk, 2018). This can be particularly problematic for vocal emotions, which are sometimes characterized by rather extreme acoustic features, impeding both inter- and extrapolation between emotional categories (Grichkovtsova et al., 2012; Nussbaum, von Eiff et al., 2022). Therefore, in the next paragraph, I will not only summarize key findings based on voice morphing technology, but also critically reflect on the special case of vocal emotions.

2.3.2. Applications of (parameter-specific) voice morphing

Within the last decade, voice morphing has developed into a standard method for research on non-linguistic vocal processing. It has been used to study the perception of vocal sex (Burgering et al., 2020; Pernet & Belin, 2012; Schweinberger et al., 2008; Skuk & Schweinberger, 2014; Zäske et al., 2009), identity (Schelinski et al., 2017; Zäske et al., 2010), age (Zäske & Schweinberger, 2011; Zäske et al., 2013), emotion (Amorim et al., 2022; Bestelmeyer et al., 2010; Morningstar et al., 2021; Nussbaum, von Eiff et al., 2022; Pan et al., 2017; von Eiff et al., 2022), attractiveness (Bruckert et al., 2010), and sexual orientation (Kachel et al., 2018).

One important auditory paradigm enabled by voice morphing is perceptual adaptation. Adaptation refers to a perceptual bias to opposite features of a stimulus after repeated exposure. This perceptual bias manifests in a contrastive aftereffect: For example, after prolonged exposure to an angry voice (the adaptor), a subsequently presented ambiguous voice (the target) at an intermediate position within an angry-fearful continuum will more likely be classified as a fearful than an angry voice (Bestelmeyer et al., 2010, 2014; Nussbaum, von Eiff et al., 2022). Conversely, after exposure to a fearful adaptor, the very same ambiguous target voice will appear more likely as angry. Such adaptation aftereffects have not only been shown for vocal emotion, but also for sex, age, and identity (Bestelmeyer & Mühl, 2021; Bestelmeyer et al., 2010, 2014; Burgering et al., 2020; Nussbaum, von Eiff et al., 2022; Schweinberger et al., 2008; Skuk & Schweinberger, 2013; Zäske & Schweinberger, 2011; Zäske et al., 2009, 2010). The experimental design crucially depends on the creation of a morphing continuum between two voices (e.g. of different emotions or sex), since the endpoints of such a continuum are usually used as adaptors and the voice morphs at intermediate positions on an analogous continuum are used as targets. Other voice morphing applications go beyond interpolation of two voices only. Voice caricaturing, for example, is based on an extrapolation/exaggeration of parameters, making the unique acoustic features of a voice more distinctive. Bestelmeyer et al. (2010) and Whiting et al. (2020) showed that acoustic caricaturing of emotions increases their emotional intensity and can facilitate recognition. Another exciting, but technically more challenging application is voice averaging, which is based on the acoustic integration of many different voices (Bruckert et al., 2010; Fontaine et al., 2017; Kachel et al., 2018).

Although full voice morphing can nowadays be considered a well-established method, research exploring the potential of **parameter-specific voice morphing** is still sparse. Most studies were conducted on the perception of vocal sex. Several studies emphasized the role of both fundamental frequency and different timbre parameters in normal-hearing listeners (Pernet & Belin, 2012; Skuk & Schweinberger, 2014; Skuk et al., 2015). In contrast, cochlear implant (CI) users seem to base the judgement of vocal sex almost exclusively on F0 (Skuk et al., 2020). For age perception in CI users, however, timbre seems to be more informative than F0 (Skuk et al., 2020). In the context of cochlear implant research, this is in fact a surprising finding when considering previous claims that CIs are inefficient at conveying timbre cues, and one that inspired both extended discussion (Meister et al., 2020; Schweinberger et al., 2020) and follow-up studies (see below).

Research on vocal emotions that employed parameter-specific voice morphing is sparse, despite its potential to control acoustic features and, therefore, the possibility to open the door to causal inference (Arias et al., 2021). To the best of my knowledge, there are only two recent studies using parameter-specific morphs, that suggest a special role of timbre for vocal emotion adaptation and emotion perception in CI users, respectively (Nussbaum, von Eiff et al., 2022; von Eiff et al., 2022). What might have inhibited a wider use of this approach for vocal emotion so far are several methodological obstacles. For example, the number of possible combinations due to the different emotional categories can be intimidating. The biggest challenge, however, lies in the conservation of acoustic quality. A recombination of extreme acoustic features, as they likely occur in vocal emotions, can result in very distorted and “unnatural” voices (cf. Crookes et al., 2015; Schindler et al., 2017, for similar issues in the context of face perception). Indeed, several studies manipulating isolated vocal features comment on the perceived naturalness of their stimuli (Grichkovtsova et al., 2012; Skuk et al., 2015).

To date, it is insufficiently understood to which degree parameter-specific manipulations distort the quality of resynthesized emotions and how this limits their ecological validity. Therefore, one aim of this dissertation was to assess the potential of parameter-specific voice morphing for the study of vocal emotion perception.

2.4. Research questions

In this dissertation, I pursued three main questions:

(1) What is the contribution of different acoustic cues to vocal emotion perception?

As described in section 2.1, I will address some conceptual and methodological problems that likely explain the inconsistent findings regarding the role of acoustic cues for vocal emotions in previous literature. Specifically, the present research is distinguished by three key aspects: First, I focused on pitch contour (F0) and timbre only. The contribution of timbre to the formation of emotional judgements is far less understood than that of F0. By their direct comparison, I hope to shed some light on the role of this complex parameter. Second, I focused on the relative importance of F0 and Timbre, rather than on their absolute predictive value for emotional judgements. Absolute predictions are only informative if acoustic cues are assumed to be independent, which is not plausible for vocal emotions. Instead, and according to Brunswik's lens model, acoustic cues are likely to be somewhat redundant, and listeners may rely on them in a flexible manner. By rendering isolated acoustic cues uninformative, I test the limits of this flexibility and assess the unique contribution of F0 (or timbre) that cannot be compensated by timbre (or F0). For example, to the extent that emotion recognition performance drops when timbre is the only diagnostic cue while F0 is rendered uninformative, one would conclude that F0 carries unique emotional information that cannot be compensated by timbre. Third, I employed parameter-specific voice morphing to acoustically manipulate the voices. Most of the research assessing the importance of acoustic cues for vocal emotion perceptions is correlative in nature and does not allow causal inference. By testing the perceptual effects of a direct manipulation (as opposed to just a measurement) of acoustic cues in the experimental stimuli, the present work allows to assess a causal role of F0 and timbre for vocal emotion perception.

Further, the present work aims to shed some light on the brain mechanisms involved in emotional impression formation. Although current models highlight the integration of relevant acoustic cues in real time to inform emotional processing, very little is known about how this happens with different acoustic features in different emotions. Therefore, I employed an EEG paradigm to explore the parameter-specific modulations of early (N100, P200) and late (N400, LPP) ERP components.

(2) How does musicality affect the processing of emotional voice cues?

Previous literature suggests a link between musicality and vocal emotional processing, but evidence is very heterogeneous, and several questions remain unanswered. To address some of them, the present work had three objectives: The first was to provide a systematic review on the role of musicality for vocal emotion perception and, subsequently, to pursue empirical investigations of musicality effects while addressing some of the methodological limitations identified in the literature. Second, I assessed how musicians and non-musicians differed in their use of acoustic

cues to infer vocal emotions. I used acoustically manipulated stimuli, which allowed me to quantify the degree to which musicians might display a specific sensitivity to isolated acoustic cues. Third, I addressed the field's lack of systematic neuroscientific experiments by recording the EEG of musicians and non-musicians in response to these acoustically manipulated vocal emotions. Thus, I explored the parameter-specific modulation of event-related potentials and compared them between both groups.

(3) Is parameter-specific voice morphing a suitable tool to study the processing of vocal emotions?

With most research on acoustic cues being of correlational nature, voice manipulation software now enables causal inferences. Parameter-specific voice morphing is a tool with this promising functionality. However, in practice, it is technically still very challenging as it involves the re-combination of acoustic features from different voice recordings. Unfortunately, a re-combination of extreme acoustic values, as they can commonly occur in vocal emotions, could affect the quality of resynthesized voices by making them sound distorted and unnatural. Such unwanted side effects can pose a serious threat to ecological validity, which in turn would limit insight into the two research questions stated above. Therefore, testing the validity of parameter-specific voice morphs was a key concern of this work. To this end, I assessed the perceived naturalness (i.e. human-likeness) of the stimulus material and measured the degree to which reduced naturalness would interfere with emotion perception.

2.5. Research outline

The following chapters all address my research questions (1), (2), and (3), but not in chronological order. The Chapters 3 and 4 contain relevant preparatory work that subsequently motivated important design features of the empirical evidence reported in the Chapters 5-7, and were therefore placed first.

Chapter 3 addresses research question (2). It provides a systematic review on the link between musicality and vocal emotion perception. It aims to integrate the current available evidence for and against this link, potential moderating factors and current methodological limitations, which could foster future research. This chapter has been published as:

Nussbaum, C., & Schweinberger, S. R. (2021). Links Between Musicality and Vocal Emotion Perception. Emotion Review, 13(3), 211-224. <https://doi.org/10.1177/17540739211022803>

Chapter 4 is a behavioral study addressing question (3). It provides a critical evaluation of the parameter-specific voice morphing approach for vocal emotions. Specifically, perceived naturalness of the stimulus material was assessed. A key objective of this chapter was to investigate if the perception of emotion in these stimuli was substantially affected by a reduction of voice naturalness. A version of this chapter has been submitted as a manuscript to *Cognition and Emotion*, and is currently in revision:

Nussbaum, C., Pöhlmann, M., Kreysa, H., & Schweinberger, S. R. (2023). Perceived Naturalness of Emotional Voice Morphs [in revision]

Chapter 5 is an empirical study incorporating behavioral and electrophysiological data. It addresses question (1), exploring the role of F0 and timbre for vocal emotion perception and their related ERP-modulations. This chapter has been published as:

Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2022). Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates. Social Cognitive and Affective Neuroscience, 17(12), 1145-1154. <https://doi.org/10.1093/scan/nsac033>

Chapter 6 is a behavioral study addressing questions (1) and (2). It compares a group of (semi-) professional musicians to a group of non-musicians with regard to their vocal emotion perception performance. Of specific interest was how musicians make use of different acoustic cues to perceive emotions, and in what way such usage might differ from that in non-musicians. A version of this chapter has been submitted as a manuscript:

Nussbaum, C., Annett Schirmer & Schweinberger, S. R. (2023). Tuned to the Melody of Vocal Emotions [under review]

Chapter 7 is an EEG study addressing questions (1) and (2). It provides insights into the parameter-specific modulations of emotional voice morphs and explores differences between musicians and non-musicians. This chapter reports on EEG data in terms of an analysis of ERPs, for comparability of methods with the study reported in Chapter 5. Please note that, in view of this unique set of data, further EEG analyses (e.g., time-frequency analyses) are currently planned, but are beyond the scope of this dissertation.

Finally, in **Chapter 8**, a general discussion integrates findings from Chapters 3 to 7 in the context of my three questions. In that vein, I also reflect on potential limitations and develop an agenda for future research.

3. Links between musicality and vocal emotion perception

This chapter has been published as:

Nussbaum, C., & Schweinberger, S. R. (2021). Links Between Musicality and Vocal Emotion Perception. *Emotion Review*, 13(3), 211–224. Copyright © 2022 (SAGE Publications) DOI: <https://doi.org/10.1177/17540739211022803>

Abstract

Links between musicality and vocal emotion perception skills have only recently emerged as a focus of study. Here we review current evidence for or against such links. Based on a systematic literature search, we identified 33 studies that addressed either (a) vocal emotion perception in musicians and non-musicians, (b) vocal emotion perception in individuals with congenital amusia, (c) the role of individual differences (e.g., musical interests, psychoacoustic abilities), or (d) effects of musical training interventions on both the normal hearing population and cochlear implant users. Overall, the evidence supports a link between musicality and vocal emotion perception abilities. We discuss potential factors moderating the link between emotions and music, and possible directions for future research.

3.1. Introduction

Human social communication depends on the exchange and mutual representation of multiple social signals. Among these, vocally expressed emotions are fundamental to human interaction (Grandjean, 2021; Sauter, 2017; Scherer, 1986; Scherer et al., 2016; A. W. Young et al., 2020). While humans perceive emotions efficiently and often automatically, interindividual differences in vocal emotion perception skills have recently become a focus of scientific attention (Mill et al., 2009; Schirmer, Striano & Friederici, 2005). For humans, voices and music are both prominent means of auditory communication of emotions. Some researchers have emphasized the similarities in the acoustic features of certain emotions in the voice and in music (Juslin & Laukka, 2003; Scherer, 1995), whereas others reported similarities in the neural circuits involved in recognizing basic emotions from music as compared to voices (Aubé et al., 2015). Accordingly, differences in vocal emotion perception skills could be associated with different levels of musicality. Music forms a central part of human culture, and appreciation of music is interconnected with intense emotional experiences (Schäfer et al., 2013). However, there is huge variation in terms of musical

aptitude and musical training. Here, we integrate the currently available research that assessed possible links between musicality - defined as sensitivity and/or talent regarding music in terms of both aptitude and training effects - and vocal emotion perception.

This review adds vocal emotion perception to several previous integrative works that discussed the association of musical training with other nonmusical abilities. Compared to nonmusical peers, musicians show superior perception of pitch (referring to the perceived frequency of a sound) and timbre (referring to the perceived quality or “color” of a sound, which allows a listener to perceive that two sounds of the same loudness and pitch can be dissimilar), as well as superior temporal processing (Kraus & Chandrasekaran, 2010), and those advantages expand from the musical into the vocal domain (Chartrand et al., 2008). Beyond the auditory modality, musicians have better audiovisual and auditory-motor integration, working memory, spatial abilities, executive functioning, general intelligence, as well as speech and language skills (Elmer et al., 2018; Schellenberg, 2001, 2016). At the brain level, musicality is associated with widespread functional and structural differences such as larger grey matter volume and stronger connectivity between areas associated with auditory, motor, and visuospatial functions (Kraus & Chandrasekaran, 2010; Pantev & Herholz, 2011). Against this background of established links between musicality and comparatively distant competencies, the field remarkably lacks systematic integration of evidence concerning the related domain of voices, and vocal emotion perception in particular.

Given the aforementioned benefits of musicality, a link between musicality and vocal emotion perception seems plausible. We considered that individual musicality may be determined by a combination of genetic (“talent”) and environmental factors (“training”), the contributions of which can be difficult to determine. Similarly, underlying mechanisms for the link between musicality and vocal emotion could be determined by both nature and nurture factors. On the nature side, some people might have an innate capacity to perceive fine-grained acoustic structures of both musical and vocal sounds, alongside with an inner drive to engage in musical activities. Accordingly, musical and vocal emotional capacities would be linked through genetic factors. This view is in line with Darwin’s protolanguage hypothesis, claiming that music and speech both evolved from the same origin, a musical protolanguage comprised of rudimental vocalizations (Darwin, 1871; Thompson et al., 2012) in which the expression of vocal emotions was a key aspect of communication (Fitch, 2013). Indeed, expression of emotion in music and vocal channels seems to be based on similar acoustic cues, supporting the idea that emotional communication in both channels is intertwined (Juslin & Laukka, 2003). This gives rise to the possibility that capacities in both channels are driven by the same underlying genetic factors, and that there are innate forces that create transfer between musical and vocal capacities.

On the nurture side, it is typically assumed that musical training causes the differences that are observed in musicians and non-musicians. From this perspective, the acoustic similarity between music and vocal emotions might be a reason why extensive training in the musical domain can lead to an improvement in vocal emotion perception. A more elaborated nurture-based approach is offered by the OPERA hypothesis (A. D. Patel, 2011). Although it was originally developed

to explain music-to-speech transfer effects, it can also be considered in the context of vocal emotion perception. The OPERA hypothesis states that musical training benefits' transfer to other domains only occurs when five conditions are met: (1) overlap, (2) precision, (3) emotion, (4) repetition, and (5) attention. Overlap refers to shared neural networks between music and vocal emotion processing. Indeed, neuroimaging data suggest common neural networks for the processing of emotional sounds, including vocal, musical, and environmental sources (Escoffier et al., 2013; Fröhholz et al., 2014, 2016; Grandjean, 2021; Schirmer et al., 2012). Core structures include the auditory cortex, the superior temporal cortex, frontal regions, the insula, the amygdala, the basal ganglia, and the cerebellum (Fröhholz et al., 2016). Precision refers to the high auditory-motor demands that musical training places on these shared networks. The third condition, emotion, claims that the musical activity has to be perceived as rewarding. The subjective feeling of highly pleasurable experiences, such as "chills" or "shivers down-the-spine", is among the main reasons why humans engage in musical activities, and those experiences are associated with brain activity changes in regions involved in reward, emotion, and arousal - including the amygdala, the ventral striatum, midbrain structures, and orbitofrontal and ventromedial prefrontal areas (Blood & Zatorre, 2001; Stewart et al., 2006). Finally, the necessities of repetition and attention stress the point that training-induced benefits depend critically on how frequently and focused musical activity is pursued over time. It may take years of active musical engagement to observe stable differences in musicians' brains compared to non-musicians' (Kraus & White-Schwoch, 2017), which can be regarded as truly reflecting training-induced changes (Elbert et al., 1995; Kraus & Chandrasekaran, 2010; Pantev & Herholz, 2011).

As it stands, the link of musicality with vocal emotion perception has received relatively little attention and is poorly understood. This seems surprising given that adequate emotion perception is crucial for well-being and perceived quality of life (Phillips et al., 2010; Schorr et al., 2009). However, to the best of our knowledge, there has been no attempt so far to integrate the existing evidence in a systematic manner. In this review, we aim at closing this gap while including the full range of musical abilities: We survey findings from highly trained musicians but also from people with exceptionally poor musical abilities. We also assess studies that investigate the role of individual differences in terms of musical interests or psychoacoustic abilities. Finally, we include music intervention studies with normal hearing individuals as well as cochlear implant (CI) users, who have significant difficulty recognizing vocal emotions due to degraded auditory input (Jiam et al., 2017), which might be improved with music-based interventions (Paquette et al., 2018).

3.2. Systematic literature search

We conducted parallel literature searches on Web of Science, PubMed, and PsychInfo on March 19, 2020, using the search terms "(voice OR prosody) AND (emotion* OR affect*) AND (music* OR auditory expert* OR amusi* OR auditory training)." We restricted publication language to English and considered empirical studies only. In total, the initial search returned 1,723 articles (Web of Science: N = 755; PubMed: N = 405; PsychInfo: N = 563) to which we applied the following inclusion criteria: (a) vocal emotion perception was assessed as dependent variable, (b)

musicality was assessed, manipulated, or used as a defining criterion for a group-based comparison, and (c) responses were measured at the behavioral or brain level. We also screened the reference lists of the identified articles for relevant publications. This selection procedure resulted in a total of 33 articles, which we review in what follows (for a summary, please refer to Table 3.1). When screening the relevant literature, we noted that a substantial proportion (27 out of 33; 82%) of identified articles were published in 2011 or later, reflecting the current attention to this topic that contributed to motivating our present review.

3.3. Review of identified literature

After screening the identified literature, it became apparent that studies could be grouped according to different operationalizations of musicality, which focused either on individual differences that existed before the study was conducted or on experimental interventions seeking to create such differences via controlled treatment designs. Following this structure, we first review the evidence concerning individual differences, in three sections focusing on (a) differences between musicians and non-musicians, (b) data on individuals with congenital amusia, and (c) correlations of vocal emotion recognition with either musical interests or psychoacoustic abilities. Subsequently, (d) we discuss effects of musical training interventions on vocal emotion perception.

3.3.1. Differences in vocal emotion perception between musicians and non-musicians

Behavioral data Several studies found a musician effect on vocal emotion recognition in the adult population. The presumably first study comes from Nilsson and Sundberg (1985), where music students outperformed law students in judging whether a vocal sample was recorded during a depressive period of the speaker or not. These findings were replicated by Thompson et al. (2004). Musically trained participants performed better than untrained participants at categorizing emotional tone sequences extracted from vocal utterances. In a second experiment, however, which included additional sentences in familiar and unfamiliar languages, the effects were less straightforward: musicians' performance was better for sad, fearful, and neutral, but not for happy and angry prosody. As a caveat, musical training was associated with differences in cognitive abilities too, limiting conclusions from this study. This issue was addressed by Lima and Castro (2011), who compared highly trained musicians with non-musicians in two age groups (18–30 and 40–60 years). Musicians outperformed their nonmusical peers similarly across emotions and age groups, even when effects were controlled for cognitive differences. Further, similar patterns of misattributions and acoustic cue utilization were observed in both groups, suggesting that group differences were of a quantitative rather than a qualitative nature.

Table 3.1.: Summary of Empirical Articles on Musicality and Vocal Emotion Perception that met the Selection Criteria.

Reference	Sample (Age _{Mean})	Inclusion Criteria	Stimuli and Task	Emotions	Results
Musical Expertise – Behavioral Data		Musicians			
Başkent et al. (2018)	10 musicians (12 y); 11 non-musicians (12 y)	Training onset before age 7; > 5 y of musical training; regular training within the last 3 years	Original and degraded speech, melodic contour identification, vocal emotion identification, and speech understanding in noise	Joy, anger, sadness, and relief	Musician effect only on melodic contour identification in degraded speech condition
Dmitrieva et al. (2006)	48 musicians; 46 controls (three age groups: 7–10 y, 11–13 y, and 14–17 y)	Recruited from the Russian National Orchestra	Vocal emotion identification, presentation to one ear with white noise on the other	Joy, anger, and neutral	Musicians were better and faster at emotion recognition
Fuller et al. (2014)	25 musicians (23 y); 25 non-musicians (22 y)	Training onset before age 7; > 10 years of musical training; regular training within the last 3 years	Original and degraded (8-channel simulated) stimuli, tests on speech, vocal emotion, and melodic contour identification	Joy, anger, sadness, and relief	Musicians were better at emotion recognition in both original and degraded conditions
Lima and Castro (2011)	40 musicians; 40 non-musicians (two age groups: 18–30 y and 40–60 y)	Instrumentalists; > 8 years of musical training; started in childhood; current regular practice	Vocal emotion identification and intensity rating	Happiness, anger, fear, sadness, surprise, disgust, and neutral	Musicians were better at emotion recognition, controlling for cognitive differences
Nilsonne and Sundberg (1985)	62 musicians (21 y); 51 controls (law students, 24 y)	Music conservatory students	F0 contours extracted from utterances, presentation in pairs, choice which was made during depression	Depression versus recovery	Musicians made fewer errors
Parsons et al. (2014)	109 (29 y): parents/nonparents (25/26 musicians; 29/29 non-musicians, respectively)	> 4 years of formal music training	Pairs of infant cries, only differing in pitch, participants decided which of the two infants sounded more distressed	Distressed infant cries	Advantage for parents with musical training, increasing with years of musical training, no musical training effect for nonparents
Twaite (2016)	58 (22 y) musicians; 61 non-musicians (21 y)	Instrumentalists; > 8 years of musical training with onset before 12; currently > 2-hr weekly practice	Emotion perception in prosodic, lexical, facial, and musical channel	Happiness, anger, fear, sadness, surprise, disgust, and interest	Musicians were better at emotion recognition in the prosodic and musical channels

Table 3.1: *Continued*

Reference	Sample (Age _{Mean})	Inclusion Criteria	Stimuli and Task	Emotions	Results
Weijkamp and Sadakata (2017)	16 musicians (29 y); 16 non-musicians (22 y)	Instrumentalists; > 5 years of musical training; currently > 2.5-hr weekly practice	Audiovisual Stroop task, judging the emotion of either the face or the voice, plus unimodal classification tasks	Happiness, sadness, and neutral	No group differences in unimodal tasks, but musicians were better at the audiovisual tasks (for both faces and voices)
K. S. Young et al. (2012)	57 (27 y): depressed/healthy (13/15 musicians, 14/15 non-musicians)	> 4 years of formal music training	Pairs of infant cries, only differing in pitch, participants decided which of the two infants sounded more distressed	Distressed infant cries	Effect of musical training even during depression, correlation between task score and years of musical training
Thompson et al. (2004)	Exp. 1: nine musicians; 11 non-musicians (adults). Exp. 2: 28 musicians; 28 non-musicians (adults). Exp. 3: 43 (all 7 y)	> 8 years of formal music training; training onset during childhood	Exp. 1 and 2: VER in spoken sentences or melodic analogues. Exp. 3: VER after one year of keyboard, singing, drama, or nothing	Happiness, anger, fear, sadness, and neutral	Exp. 1 and 2: musicians performed better than non-musicians. Exp. 3: keyboard and drama better than nothing, no effect of singing
Musical expertise – Brain data		Musicians			
Nolden et al. (2017)	17 musicians (25 y); 20 non-musicians (25 y)	> 5 years of musical training; daily instrument practice	Music and vocal stimuli, task: pure tones detection, EEG recording, analysis of theta, alpha, beta, and gamma bands	Happiness, fear, sadness, and neutral	Effects of expertise on the alpha and theta bands (greater activation in musicians in both frequency bands)
Park et al. (2015)	12 musicians (20 y); 12 non-musicians (19 y)	Instrumentalists; received formal music training ($M_{\text{years}} = 13.8$, $SD = 2.6$)	fMRI scans with passive listening followed by a session with emotion classification	Happiness, fear, sadness, and neutral	Increased activation in musicians' frontal gyrus, posterior cingulate cortex, and the retrosplenial cortex for sad stimuli only
Pinheiro et al. (2015)	14 musicians (23 y); 14 non-musicians (23 y)	> 8 years of musical training; daily instrument practice	Emotion classification with intelligible semantic content (SCC) or unintelligible semantic content (PPC), EEG recording	Happiness, anger, and neutral	P50 more positive in controls than in musicians in SCC only; musicians better at recognition of anger in PCC only
Rigoulot et al. (2015)	15 musicians (24 y); 18 non-musicians (25 y)	> 5 years of musical training; daily instrument practice	Music and vocal stimuli, task: pure tones detection, EEG recording	Happiness, fear, sadness, and neutral	Differential electrophysiological response to vocal and music stimuli for musicians and non-musicians

Table 3.1: *Continued*

Reference	Sample (Age _{Mean})	Inclusion Criteria	Stimuli and Task	Emotions	Results
Strait et al. (2009)	30 adults (25 y): musicians (by onset age: 11; by years of training: 15); non-musicians (by onset age: 19; by years of training: 15)	Musicians by onset age: training onset before age 7. Musicians by years: > 10 years musical training	Exposure to infants' unhappy cries, recording of auditory brainstem responses during a complex and more periodic part of the stimulus	Infant cries	In musicians, subcortical activation is enhanced during the complex portion of the sound, and decreased during the more periodic portion
Congenital amusia		Amusia			
Cheung et al. (2020)	20 participants with amusia (22 y); 17 controls (22 y), all Cantonese speakers	Global MBEA score < 71%	Emotion prosody rating task, emotion judgment of written words task, valence judgment of written words task	Happiness, anger, fear, and sadness	Participants with amusia performed worse in the VER task than controls; no difference in tasks with written words
Lima et al. (2016)	13 participants with amusia (58 y); 11 controls (53 y)	Pitch-based MBEA scores more than 2 SD below population average (~72%)	Exp. 1: rating of emotions in voices and faces. Exp. 2: rating of spontaneous and posed laughs	Anger, fear, sadness, disgust, relief, amusement, and pleasure; laughs	Exp. 1: impairments in amusia for all stimulus types. Exp. 2: participants with amusia showed decreased sensitivity to authenticity of laughs
Lolli et al. (2015)	Exp. 1: 40 (aged 18–22), nine considered amusic. Exp. 2: 29 (aged 18–28), three considered amusic	Pitch threshold > 16 Hz in pitch discrimination task	Exp. 1: VER, low-pass filtered (500 Hz) and unfiltered speech. Exp. 2: high-pass filtered (4800 Hz) and unfiltered speech	Happiness, fear, sadness, irritation, tenderness, and neutral	Exp. 1: impairments in amusia in the low-pass filtered but not in the unfiltered condition. Exp. 2: no differences
Pralus et al. (2019)	18 participants with amusia (33 y); 18 controls (35 y)	Global MBEA score < 23/30 (76%) and/or a MBEA pitch score < 22/30 (73%)	Emotion categorization and intensity rating of emotional sentences and vowels	Joy, anger, fear, sadness, and neutral	Worse performance of participants with amusia on vowels but not sentences. Intensity ratings: no differences
Thompson et al. (2012)	12 participants with amusia (50 y); 12 controls (46 y)	MBEA scale subtest score < 22/30 (73%) on two consecutive occasions	VER, self-report questionnaire on emotional prosody perception in daily life	Happiness, sadness, fear, tenderness, irritation, and neutral	VER: participants with amusia worse than controls; report awareness of VER problems in daily life
Zhang et al. (2018)	19 participants with amusia (23 y); 19 controls (23 y), all Mandarin speakers	Pitch-based MBEA score < 65/90 (72%)	Vocal emotion recognition, speech and nonspeech (low-pass filtered at 500 Hz) conditions	Happiness, anger, fear, sadness, surprise, and neutral	Performance of participants with amusia worse than controls for all emotions and conditions

Table 3.1: *Continued*

Reference	Sample (Age _{Mean})	Inclusion Criteria	Stimuli and Task	Emotions	Results
Individual differences					
Dibben et al. (2018)	52 (32 y)	-	Rating of perceived emotion in music and speech, questionnaires on personality, emotional intelligence, and musical training	Valence and arousal	Ratings associated with emotional stability, agreeableness, musical training (but for musical stimuli only), and age
Globerson et al. (2013)	Exp. 1: 60 (25 y); Exp. 2: 37 (25 y)	-	Exp. 1 and 2: prosody recognition tasks (pragmatic and emotional) and different psychoacoustic tasks	Happiness, anger, fear, and sadness	Psychoacoustic thresholds explained 31% and 38% of affective and pragmatic prosody recognition performance
Trimmer and Cuddy (2008)	100 (22 y)	-	VER of speech utterances and melodic analogues, tests on intelligence, emotional intelligence, and musicality	Joy, anger, fear, sadness, and neutral	Emotional intelligence, not musical training, predicted vocal emotion recognition performance
Waaramaa and Leisiö (2013)	250: 50 per country (Finland: 48 y; Russia: 35 y; Estonia: 32 y; Sweden: 27 y; US: 23 y)	-	Vocal emotion recognition task with replay option, questionnaire on musical interests	Happiness, anger, fear, sadness, surprise, disgust, interest, and neutral	Emotion recognition performance above chance in all countries; musical interests tended to have a positive effect on VER
Musical intervention – Normal hearing listeners					
Bodner et al. (2012)	80 (29 y): 39 with social anxiety disorder; 41 healthy controls	-	Intervention: training in happiness recognition in music. Test: VER, pre- and postdesign	Happiness, anger, fear, sadness, and surprise	Music intervention improved recognition of vocal happiness in participants with social anxiety
Mualem and Lavidor (2015)	Exp. 1: 12 intervention group (24 y); 12 control group (25 y). Exp. 2: 23 musicians (26 y)	> 6 years of formal music training	Intervention: 4 x 30 min music sessions. Control: art session focused on emotion expression. Test: VER, pre- and postdesign	Happiness, anger, fear, sadness, and neutral	Exp. 1: intervention group better than control group in VER. Exp. 2: no difference between musicians and non-musicians
Nashkoff (2007)	42 intervention group; 39 controls	-	Intervention: Pitch discrimination practice (8 weeks). Test: VER, pre- and postdesign	Happiness, fear, sadness, and ambiguous	VER improved after intervention

Table 3.1: *Continued*

Reference	Sample (Age_{Mean})	Inclusion Criteria	Stimuli and Task	Emotions	Results
Musical intervention – Cochlear implant (CI) users					
Chari et al. (2020)	18 CI users (62 y), postlingually deafened	-	Intervention: auditory-motor, auditory-only, or no training (3 months). Test: speech, music, and VER tests, pre- and postdesign	Happiness, anger, fear, sadness, and neutral	Training effects only on melodic contour identification task
Fuller et al. (2018)	19 CI users (69 y), postlingually deafened	-	Intervention: music therapy, pitch/timbre group, or control (6 weeks). Test: speech, music, and VER tests, pre- and postdesign	Joy, anger, sadness, and relief	Improvement in VER in the music therapy group; improvement in melodic contour identification in the pitch-timbre group
Good et al. (2017)	18 CI users, children (10 y)	-	Intervention: music (piano) or art training (painting), 6 months. Test: music and VER tests, pre-, mid-, and postdesign	Happiness, anger, fear, and sadness	Music perception and emotional speech prosody perception improved in the music group compared to the art group
Petersen et al. (2012)	18 CI users (53 y); six NH controls (54 y)	-	Intervention: 6-month, one-to-one musical ear training or control (no task). Test: speech, music, and VER tests, pre-, mid-, and postdesign	Happiness and sadness	In the music group, earlier onset of improvement in emotional prosody perception; both CI groups still worse than NH controls
Waaramaa et al. (2018)	25 CI users (12 y); 18 NH controls (12 y)	-	Vocal emotion identification task, self-reported musical interests, and measurements of acoustic parameters of the stimuli	Anger, fear, excitement, and contentment	NH controls performed better than CI users; musical interests and voice quality parameters related to correct identification in both groups

Note. VER = vocal emotion recognition, MBEA = Montreal Battery of Evaluation of Amusia (Peretz et al., 2003), NH = Normal hearing, SCC = semantic content condition, PPC = pure prosody condition, y = years.

Evidence of a musician effect on vocal emotion perception in children and adolescents is less clear. Dmitrieva et al. (2006) tested vocal emotion perception by musicians and controls in children of three age groups (7–10, 11–13, and 14–17 years). Musicians outperformed their nonmusical peers, but this effect was mainly driven by the youngest group. This finding could indicate either that the limited musical experience in very young children allowed innate aptitude to be more visible or that musical experience might promote an earlier development of emotion sensitivity. Başkent et al. (2018) also studied adolescent musicians and non-musicians. Their work was motivated by Fuller et al. (2014), who conducted a similar design with adults. Both studies compared emotion recognition performance of unprocessed and degraded speech intended to reproduce the spectro-temporal degradation experienced by CI users (Başkent et al., 2018). Whereas Fuller et al. (2014) found a small musician advantage in both conditions, Başkent et al. (2018) did not, potentially due to limited test power with a much smaller sample compared to Fuller et al.’s (2014). Statistical comparison of both experiments revealed significant age effects, suggesting a maturation of the auditory system during emerging adulthood (Başkent et al., 2018).

At an even earlier stage of development, sensitivity to vocal prosody is of particular relevance in parent-infant interactions, where parents’ adequate behavior crucially depends on their capacity to infer the infants’ needs from the sound of their cries. Parsons et al. (2014) found that parents’ musicality could foster a positive parent-child interaction due to enhanced sensitivity to the infants’ emotional state. K. S. Young et al. (2012) used the same paradigm on musicians and non-musicians with and without depression to show that musicality can potentially protect against compromised auditory sensitivity towards the infant during a depression period.

Several studies raised the question whether emotional sensitivity in musicians is restricted to the auditory domain. Twaite (2016) compared performance of musicians and non-musicians for prosodic, lexical, facial, and musical emotions, and reported a musician advantage for the prosodic and the musical channels only, indicating that the musicians’ advantage was limited to the auditory modality. However, Weijkamp and Sadakata (2017) reported somewhat conflicting results. In this study, musicians performed better in an audiovisual task, which might point towards more efficient cross-modal integration.

Up to now, most studies comparing musicians and non-musicians support the hypothesis that musicians have an advantage in vocal emotion perception. This advantage seems to be of a quantitative rather than a qualitative nature (Lima & Castro, 2011; Twaite, 2016). However, the degree to which this advantage is moderated by factors such as innate musicality, the amount of musical training, age at training onset, or maturation of the auditory system all remain subjects for future research.

Brain data A small number of studies investigated differences between musicians and non-musicians with respect to the brain basis of vocal emotional processing. Existing work on brainstem potentials and auditory-evoked responses in the electroencephalogram (EEG) suggests

that effects of musicality can be observed in very early stages of vocal emotion processing. Strait et al. (2009) recorded brainstem potentials evoked by acoustically simple and complex portions of an infant's cry, and reported an intriguing interaction between musical expertise and stimulus complexity: Compared to controls, musicians showed reduced responses in the simple, but increased responses in the complex portion of the sound. The authors interpret these findings as indicating that (a) musical expertise results in fine neural tuning to acoustic features that are important to vocal communication, and (b) subcortical mechanisms contribute to vocal emotion perception. At the cortical level, Pinheiro et al. (2015) and Rigoulot et al. (2015) recorded event-related potentials (ERPs) and found that modulatory effects of musical expertise can be observed in early stages of cortical processing before 100 ms (P50, N100), as well as in later stages (P200). Nolden et al. (2017) reanalyzed the data of Rigoulot et al. (2015) with a focus on induced oscillatory activity, and found larger induced power for musicians in the theta (4–8 Hz) and the alpha bands (8–12 Hz).

Using functional magnetic resonance imaging (fMRI), Park et al. (2015) showed that musicians exhibited increased activation in frontal areas, the posterior cingulate cortex, and the retrosplenial cortex. However, these differences were observed for sad stimuli only. Accordingly, the authors hypothesized that sadness might be of “higher affective saliency” for musicians.

Taken together, the existing work on brainstem potentials and EEG suggests that modulatory effects of musicality can be observed in very early processing steps of vocal emotions, which are associated with a basic analysis of auditory cues and allocation of emotional significance (Schirmer & Kotz, 2006). Neuroimaging complements this by implicating brain regions associated with higher order functions, such as evaluative judgements or empathic engagement (Park et al., 2015). Overall, neuroscientific research on links between musicality and vocal emotion perception is still in its infancy, although it clearly has potential to shed light on the underlying mechanisms of musicians' enhanced ability to process vocal emotions.

3.3.2. Impairments of vocal emotion perception in individuals with congenital amusia

To gain a full understanding of the association between musicality and vocal emotion perception, it is worthwhile to consider the entire performance spectrum by including individuals with exceptionally poor musical abilities, such as in congenital amusia. Individuals with congenital amusia display a perceptual disorder specific to the musical domain, in the presence of normal hearing and otherwise intact cognition (Ayotte et al., 2002; Stewart et al., 2006). Congenital amusia is usually measured using the Montreal Battery of Evaluation of Amusia (MBEA; Peretz et al., 2003). Thompson et al. (2012) were the first to show poorer vocal prosody recognition in participants with amusia compared to controls, and further observed a certain degree of awareness of their perceptual limitations in daily life. Lolli et al. (2015) suggested that a core problem could be poor pitch (F0) perception: although participants with suspected amusia performed similar to controls on emotion perception from unfiltered or high-pass filtered (4800 Hz) utterances, they performed poorer for low-pass filtered (500 Hz) utterances, which presumably degraded timbre while preserving pitch information. Corroborating a selective deficit in pitch perception, Pralus

et al. (2019) found that controls and participants with amusia exhibited comparable emotion recognition for whole sentences, but participants with amusia performed worse for vowels. Of relevance, perceived emotional intensity was comparable in both groups for all stimuli, which was interpreted as preserved implicit processing of emotional prosody in amusia.

Lima et al. (2016) took a cross-modal approach: in two experiments, they tested adults with amusia and matched controls on their ability to identify emotions in different types of vocal stimuli and silent facial expressions. Participants with amusia were found to be impaired in the auditory and the visual domain, implying more universal emotion processing difficulties. Zhang et al. (2018) and Cheung et al. (2020) were interested in relationships between amusia and emotional prosody processing in tonal languages. Compared to controls, they reported poorer performance in participants with amusia, both with a Mandarin-speaking and a Cantonese-speaking background, disconfirming the hypothesis that tonal language acquisition might compensate for pitch processing deficits in participants with amusia. Taken together, published findings on amusia paint a fairly consistent picture, suggesting that musical impairments transfer to vocal emotion perception, and that impairments for vocal emotions may originate in poor pitch perception.

3.3.3. Correlation of vocal emotion perception with musical interests or psychoacoustic abilities

Complementing studies on extreme groups, other researchers measured normal interindividual variation in musicality in the general population, to link it to variability in vocal emotion perception. These studies result in conflicting findings. The most compelling evidence against such a link may have been provided by Trimmer and Cuddy (2008). Their correlational analysis of 100 participants revealed that musicality, as assessed via MBEA scores, was not associated with vocal emotion perception, a finding they replicated with another 92 participants. Trimmer and Cuddy (2008) concluded that emotion perception in music and the voice is not linked via auditory sensitivity but rather via a supramodal emotional processor. This finding conflicts with many results discussed before, and provoked large debates in the field. For instance, Lima and Castro (2011) argued that participants in the study had only 6.5 years of musical training on average, which might have been insufficient to observe a significant effect. However, Dibben et al. (2018) also failed to observe a relationship between musical interests and moment-to-moment reports of perceived emotion in longer (2–3 min) vocal excerpts. As a limitation, musical interests were assessed with a single dichotomous item in this study, which hardly captured fine-grained interindividual variation in musicality.

Other studies reported positive correlations. Globerson et al. (2013) did not assess full-scale musical ability, but psychoacoustic measures of sensitivity to pitch were found to predict vocal emotion perception performance. This highlights the importance of subtle pitch variations for emotional prosody perception, in line with the impairments found in amusia discussed in the previous section. Finally, Waaramaa and Leisiö (2013) investigated the link between musical interests and emotional prosody perception in a large-scale, cross-cultural study across five

different countries (Estonia, Finland, Russia, Sweden, US). Musical interests tended to have a positive effect on vocal emotion identification, but as in Dibben et al. (2018), this finding was based on very few self-reported items only. In summary, correlational studies on the link between musicality and vocal emotion perception have yielded conflicting results. This could be due to substantial differences in the assessment of musicality across studies, ranging from musical tests to short questionnaires. Indeed, the use of standardized and validated instruments for the assessment of musical interests is desirable, as this should promote better comparability across future studies, and contribute to resolving remaining controversies.

3.3.4. Effects of musical training interventions on vocal emotion perception

Apart from comparing individual differences in musicality, the effectiveness of musical interventions was the focus of several studies, which we review in this section. Note that designs with randomized assignments to intervention and control conditions are particularly valuable in the context of the nature/nurture debate, as they permit to de-confound training effects from self-selection effects when seeking musical education. Our literature survey indicated that intervention studies could be grouped into interventions for normal hearing listeners, on the one hand, and interventions for hearing-impaired individuals with cochlear implants, on the other hand.

Intervention-based studies on normal hearing individuals A few studies suggest the effectiveness of musical training interventions for normal hearing individuals. Thompson et al. (2004) randomly assigned six-year-old children to one year of training in keyboard, singing, drama, or no lesson. Post-intervention, the drama and keyboard groups outperformed the no-lesson group in vocal emotion perception. Perhaps surprisingly, this effect was not found in the singing group. Thompson et al. (2004) speculated that singing may have trained vocal production of pitch contours over time that conflicts with natural prosodic use of the voice. Nashkoff (2007) reported that simple pitch perception training alone can improve speech prosody decoding skills, but only for already highly trained musicians. Another attempt to show the effectiveness of musical interventions was made by Mualem and Lavidor (2015), who assigned participants either to music-based or visual-art-based interventions, which focused explicitly on expression of emotions in the respective domain. After only four sessions, an improvement in vocal emotion recognition performance was observed in the music compared to the art group. However, when both groups were compared to a group of highly trained musicians, no performance differences were found. This could suggest that the effectiveness of the intervention partially reflected “training to the test”, as the intervention explicitly focused on emotions. Finally, Bodner et al. (2012) reported that a music-based intervention improved recognition of happiness in patients with social anxiety disorder (SAD), who often display a persistent bias towards negative emotions. Although these findings need further verification, they suggest that perceptual biases in affective disorders may be attenuated by musical interventions. Together, the body of literature on interventions suggests musical training effects, but it is still sparse for the normal hearing population. Of interest, musical interventions have been studied more intensely in the field of hearing rehabilitation for cochlear implant users.

Intervention-based studies on cochlear implant users All studies reviewed in this section included vocal emotion perception as a part of larger test batteries to assess musical training effects on voice, speech, and music perception in cochlear implant (CI) users. Petersen et al. (2012) recruited CI users within 14 days after implantation. Half of them received a 6-month musical ear training. While distinct improvements were observed for musical perception, the pattern was less clear for vocal emotions: the intervention group showed an earlier onset of improvement but the endpoints were comparable. As a qualification, the freshly implanted CI users in this study were in speech therapy during the intervention, which could have interfered with the musical training. In contrast, Fuller et al. (2018) studied adult CI users with a minimum of one year postimplantation, who were randomly assigned to either (a) a pitch/timbre group that received receptive training, (b) a music therapy group with face-to-face sessions including active music production, or (c) a control group with nonmusical activities, over a period of 6 weeks. Crucially, vocal emotion recognition improved only in the music therapy group, emphasizing the importance of active musical engagement and/or social interaction for training success. Similarly, Chari et al. (2020) also studied adult CI users with at least one year of implant experience, and assigned them to auditory-motor, auditory-only, or no training, over a period of 3 months. However, there was no effect on vocal emotion perception even though the intervention period was about twice as long as in Fuller et al. (2018). Notably, both studies used very small sample sizes, with less than 10 participants per group. While findings are intriguing and potentially important, they call for further exploration and replication with more powerful designs, particularly when effect sizes and statistical power are not (yet) routinely reported.

Only one study investigated the role of musical training in children with CIs, aged 6 and 15 years (Good et al., 2017). Improvements in vocal emotion perception were found after 6 months of piano lessons compared to a visual art training. The authors concluded that musical training might be an effective supplement to auditory rehabilitation in children. In addition to intervention-based approaches, Waaramaa et al. (2018) showed that self-reported musical interests - especially a preference for dancing - predicted vocal emotion perception capacity in CI users. Taken together, findings emphasize an importance of active musical engagement, as compared to pure receptive training, in order to promote recovery of emotion perception after cochlear implantation.

3.4. Discussion

While the transfer of musicality to speech perception abilities is well documented, the transfer to emotion perception attracted substantial scientific interest only recently. Overall, while associations between musicality and vocal emotion perception ranged from strongly positive to absent, the majority of studies supported the idea that musicality is associated with better vocal emotion perception capacities. Both studies with highly trained musicians, on the one hand, and with individuals with amusia, on the other hand, suggest that musical capacities are positively associated with vocal emotion recognition. Correlational analyses with varying degrees

of musicality in a normal population revealed less consistent results, presumably partly due to their methodological heterogeneity. Musical intervention studies are still sparse but illustrate great potential to improve vocal emotion perception capacities both in the normal hearing population and in cochlear implant users. In the following lines, we will first discuss potential moderators by evaluating the effectiveness of active versus receptive musical training, and the role of different acoustic cues signaling emotionality in both music and the voice. We then will discuss how these studies inform us about the contribution of nature and nurture factors to this link.

3.4.1. Active engagement in musical activities versus receptive training

Several studies suggest that active engagement in a musical task is a crucial factor. They compared purely receptive training to auditory-motor training, and reported stable benefits of auditory-motor training in the vocal emotion domain (Chari et al., 2020; Fuller et al., 2018). There is high consensus in the neuroscientific literature that active engagement in music and the synchronized tuning of auditory, visual, somatosensory, and motor processes is a driving force to adaptive neuroplasticity (Kraus & Chandrasekaran, 2010; Kraus & White-Schwoch, 2017; Palomar-García et al., 2017). Specifically, it has been shown that sensorimotor musical training leads to more robust changes in the auditory cortex compared to pure receptive training (Lappe et al., 2008). This surely does not imply that purely receptive music training is ineffective (Bigand & Poulin-Charronnat, 2006), but motor engagement may add a boost to the auditory fine-tuning process during training. This could be of particular relevance for cochlear implant users (Lehmann & Paquette, 2015), who during rehabilitation face the challenge of massive postimplantation adaptation to the new auditory input. Here, auditory-motor interventions could be particularly efficient in fostering neuroplasticity in auditory areas, and in aiding hearing rehabilitation.

3.4.2. The role of different acoustic cues and supramodal processes

Previous literature suggests that musicians show superior processing of auditory cues (Elmer et al., 2018). The present review reveals that superior pitch processing capacities in people with high levels of musicality are particularly tightly associated with vocal emotion perception. On the one hand, pitch discrimination performance was correlated with emotion perception performance in musicians; on the other hand, there was strong agreement that impaired pitch processing was a key deficit in people with amusia, accounting for impairments in the domain of vocal emotions. However, this conclusion has its limitations since amusia was often defined only based on low scores on the pitch subsets of the MBEA (see Table 3.1). According to Juslin and Laukka (2003), pitch and timbre cues are highly relevant for vocal emotion perception, but timing parameters like speech rate were found to be equally important. The potential role of timing was largely neglected in all reviewed studies, despite its central role in music, therein often referred to as tempo and rhythm. Lagrois and Peretz (2019) showed that although pitch and rhythm deficits are often linked in people with amusia, they sometimes can appear as distinct disorders. In parallel, there is current evidence of different brain mechanisms processing pitch-related versus timing-related

structures in music (Sun et al., 2020). In the future, it would be very informative to investigate vocal emotion perception in people with specific impairments related to the temporal domain of music.

Alongside the notion that enhanced sensitivity to acoustic cues may lead to better emotion perception in people with a higher level of musicality, it was also suggested that there might be a domain-general supramodal process that mediates the link between musicality and emotional perception across domains (Schellenberg & Mankarious, 2012; Trimmer & Cuddy, 2008). Lima and Castro (2011) suggested that musical training might increase the level of “emotional granularity”, meaning a more fine-grained conceptualization and differentiation of emotions that, in turn, could aid emotional perception in other domains. However, although the involvement of supramodal processes seems plausible, the reviewed brain data suggest that modulatory effects of musicality can be observed in very early vocal emotion processing steps, which are associated with a basic analysis of auditory cues and detection of emotional saliency (Schirmer & Kotz, 2006). Hence, the link between musicality and vocal emotion perception seems to be based, at least partially, on a more sophisticated analysis of auditory cues.

3.4.3. Nature and nurture

Musicality in people emerges from a combination of genetic and environmental factors. Likewise, the observed link between musicality and vocally expressed emotions could be either explained by a dispositional sensitivity to the musical and the vocal channels or by a transfer from musical training effects into the vocal domain. Additionally, conditions of nature and nurture interact in individuals, making it difficult to estimate the degree of their respective contributions. Unfortunately, apart from few randomized treatment studies that potentially isolated training effects, all reviewed articles established correlational designs or studied preexisting groups, and thus cannot provide direct evidence of the relative contributions of nature or nurture conditions. Nevertheless, it is worthwhile to consider implications of certain findings for this debate.

Without exception, all the studies on people with amusia suggested that vocal emotion perception deficits can be associated with a congenital music perception impairment. In that sense, the link between musicality and vocal emotion perception seems to occur in the absence of training effects and might therefore be mediated by genetic factors. These may have evolved in parallel with acoustic similarities between vocal and musical emotions (Juslin & Laukka, 2003), and may be expressed in overlapping neural circuits involved in recognizing basic emotions in voices and music (Frühholz et al., 2016). As a qualification, Bigand and Poulin-Charronnat (2006) showed that a remarkable degree of auditory sophistication can be acquired through exposure to music only, without explicit training. Accordingly, it remains possible that these implicit musical learning processes could be limited in people with amusia if they avoid exposure to music because they enjoy it less. Thus, while the limited vocal emotion perception capacities observed in amusia could hint to a genetic predisposition, they could also result in part from selective exposure.

At the same time, a consistent set of findings in highly trained musicians suggests that explicit musical training does play a central role in the development of auditory and vocal perceptual skills. This points to an influence of environmental factors but there is always a possible confound with natural inclination, as people with better auditory skills may be more likely to start and pursue musical training (Pantev & Herholz, 2011). Accordingly, Dmitrieva et al. (2006) observed superior vocal emotion perception capacities in a very young group of musicians, who presumably had very little musical training yet but might have been selected for musical education based on their auditory sensitivity. Note that many authors who found the musician effect on vocal emotion perception argued that it is very unlikely that it is entirely based on predispositional differences (Lima & Castro, 2011; Strait et al., 2009; Thompson et al., 2004): On the one hand, the effect was still present when participants were matched in socioeducational variables, general intelligence, cognitive control, and personality traits (Lima & Castro, 2011). On the other hand, some studies found a correlation between emotion perception capacities and years of musical education, suggesting a clear impact of training duration (Parsons et al., 2014; Twaite, 2016; K. S. Young et al., 2012). However, this could also reflect a gene-environment interaction since people who have a dispositional aptitude might stick longer to the training. Further, vocal emotion perception was found to be related to age at training onset (Strait et al., 2009). Although it is often difficult to disentangle age at onset from years of musical training, this could suggest a sensitive period for the acquisition of some music-training-induced skills. Accordingly, many studies required musicians to have started training before the age of 7 (see Table 3.1).

Finally, a few intervention studies with randomized assignment to treatment and control groups aimed at isolating learning effects of musical training and succeeded to improve vocal emotion perception in a healthy population. Note that these interventions were qualitatively different from more “natural” settings of musical education where the focus lies on mastery of an instrument or the singing voice. They were shorter and often particularly focused on emotion expression in music (Bodner et al., 2012; Mualem & Lavidor, 2015), except for Thompson et al. (2004), who randomly assigned children to one year of keyboard or singing lessons, but found mixed results. Likewise, studies on cochlear implant users showed that musical training can improve vocal emotion perception in this particular group, but, again, those interventions had an entirely different purpose than for normal hearing participants: instead of fine-tuning a healthy auditory system, CI users have to learn how to restore perception from a severely degraded input. Music-based interventions may be particularly effective in groups with poor auditory resolution to improve sensitivity to auditory cues in the vocal domain (Fuller et al., 2018; Good et al., 2017). Overall, while it may be difficult to generalize the results of these intervention studies to settings of instrumental or vocal music lessons, they show that vocal emotion perception can be improved through musical training in some circumstances. Accordingly, it seems worthwhile to incorporate emotionally oriented teaching units in music lessons or treatment programs.

3.4.4. Identification of relevant topics for future research

The findings surveyed in this chapter highlight many relevant aspects that can guide future research on relationships between musicality and vocal emotion perception. We hope this chapter will inform systematic research programs with better powered designs and standardized research materials, and ultimately promote a refined understanding of the putative common mechanisms underlying musicality and vocal emotion perception. Unfortunately, neuroscientific research in this field is still sparse and unsystematic, and the heterogeneity of the previous studies illustrates the need for more systematic research on candidate subcortical and cortical mechanisms to mediate the link between musicality and vocal emotion perception (for a recent review on subcortical and cortical mechanisms of nonverbal voice perception, see Frühholz & Schweinberger, 2021). Important questions for neuroscientific research include how perceptual neuroplasticity is induced in musicians, how this relates to motor plasticity in musicians' brains (Elbert et al., 1995), what are the relative roles of training or ongoing maintenance (Merrett et al., 2013), and how each of these aspects relates to vocal emotion perception. Moreover, a particularly relevant comparison in the context of vocal emotion perception is the one between singers and instrumentalists. Most of the reviewed studies only included instrumentalists or did not report on that matter. Only Thompson et al. (2004) compared participants with piano and singing lessons, and their results suggested that singing lessons might even hinder vocal emotion perception, perhaps because the vocal patterns that are trained during singing lessons may conflict with natural vocal emotion expression. Another neglected but related field is the link between musicality and emotion production. It may be reasonable to assume that people who are highly trained in an emotionally expressive art have an advantage in vocal expression of emotion.

3.5. Conclusion and outlook

In this review, we systematically identified and discussed the current state of research on the link between musicality and vocal emotion perception. Overall, the available evidence suggests that musicality is indeed associated with better vocal emotion perception performance. Since adequate perception of vocal emotions is a fundamental prerequisite for everyday social interaction, these results also may add weight to the presumed importance of music and musical education for personal development and quality of life. Musical training can provide a promising supplemental intervention for people who struggle with vocal emotion perception, and while supporting evidence can now be considered strong in the case of cochlear implant users, future applied research seems promising in the context of other target groups as well (e.g., individuals with autism or with auditory impairments compromising social communication). Although data often do not allow for causal inferences, their combined consideration can provide useful information on the question of how different factors of nature and nurture contribute to related skills of emotion perception in the domains of voice and music.

4. Perceived naturalness of emotional voice morphs

Submitted as:

Nussbaum, C., Pöhlmann, M., Kreysa, H, and & Schweinberger, S. R. (2023). Perceived Naturalness of Emotional Voice Morphs [in revision]

Abstract

Research into voice perception benefits from manipulation software to gain experimental control over acoustic expression of social signals such as vocal emotions. Today, parameter-specific voice morphing allows a precise control of the emotional quality expressed by single vocal parameters, such as fundamental frequency (F0) and timbre. However, potential side effects, in particular reduced naturalness, could limit ecological validity of speech stimuli. To address this for the domain of emotion perception, we collected ratings of perceived naturalness and emotionality on voice morphs expressing different emotions either through F0 or Timbre only. In two experiments, we compared two different morphing approaches, using either neutral voices or emotional averages as emotionally non-informative reference stimuli. As expected, parameter-specific voice morphing reduced perceived naturalness. However, perceived naturalness of F0 and Timbre morphs were comparable with averaged emotions as reference, potentially making this approach more suitable for future research. Crucially, there was no relationship between ratings of emotionality and naturalness, suggesting that the perception of emotion was not substantially affected by a reduction of voice naturalness. We hold that while these findings advocate parameter-specific voice morphing as a suitable tool for research on vocal emotion perception, great care should be taken in producing ecologically valid stimuli.

4.1. Introduction

The human voice is a powerful transmitter of emotions, which are expressed through its acoustic properties (Scherer, 1986). The functional role of vocal parameters such as fundamental frequency contour (F0) and timbre in the expression and perception of different emotions has been extensively studied (Banse & Scherer, 1996; Juslin & Laukka, 2003), but findings are mostly based on correlational data and do not allow causal inferences (Arias et al., 2021). Recently however, technical and computational progress has led to the development of voice manipulation tools allowing experimental control over the acoustic properties of voices (Kawahara & Skuk, 2018).

In **parameter-specific voice morphing**, a parameter of voice A is combined with another parameter of voice B. For example, one can resynthesize a voice with a happy F0 contour together with the timbre information of a non-emotional voice, resulting in a voice which expresses happiness only via F0. While this technology offers exciting prospects in determining the acoustic correlates of socio-emotional signals in voices, it comes with a central caveat: these manipulations may lead to profound acoustic distortion, making them sound unnatural and less human-like. To date, it is unclear how an impression of naturalness in voices is formed, how it may be affected by voice manipulations, and how it interacts with the perception of vocal emotions. In two experiments exploring the perception of naturalness in parameter-specific voice morphs, we investigated these open questions. In what follows, we will first discuss the potentials and caveats of parameter-specific voice morphing. Then, we outline insights into voice naturalness across different research domains, motivating the design of our experiments.

4.1.1. The potentials and limits of parameter-specific voice morphing in vocal emotional research

Parameter-specific voice morphing is a useful tool to study how different acoustic cues facilitate the perception of vocal age, sex, and identity (Kawahara & Skuk, 2018; Skuk & Schweinberger, 2014; Skuk et al., 2020), as well as – recently – vocal emotion (Nussbaum, von Eiff et al., 2022). In most cases, the main focus has been on the functional role of F0 (perceived as voice pitch) and timbre (perceived as voice quality, and formally defined as “the difference between two voices of identical F0, intensity and temporal structure” ANSI, 1973). For emotional stimuli, the relative importance of F0 and timbre differs as a function of emotion category, but overall F0 seems to be more important for the perception of emotional quality, at least in the normal-hearing population (Nussbaum, Schirmer & Schweinberger, 2022). In individuals using cochlear implants, by contrast, von Eiff et al. (2022) observed a greater reliance on timbre cues. Further, timbre seems to play a predominant role in emotional adaptation (Nussbaum, von Eiff et al., 2022), similar to findings on vocal sex adaptation (Skuk et al., 2015). Interestingly, these findings are in contrast to Hubbard and Assmann (2013), who found F0 to be more important than timbre in sex and emotion adaptation, based on the absence of effects in a F0-removed condition. Both Skuk et al. (2015) and Nussbaum, von Eiff et al. (2022) argued that this discrepancy could be explained by a lack of naturalness in Hubbard and Assmanns (2013) F0-removed condition, which might have eliminated the adaptation effects. It therefore seems essential that parameter-specific voice morphing results in natural sounding stimuli, by which we understand them to constitute a **plausible outcome of the human speech production system**. In fact, many studies using voice manipulation explicitly comment on the naturalness of their stimulus material (Grichkovtsova et al., 2012; Nussbaum, Schirmer & Schweinberger, 2022; Skuk et al., 2015), although this is usually based on subjective listening impression only. Vocal emotions, however, are often characterized by acoustic extremes, and this could result in reduced naturalness to a degree that compromises stimulus validity, thus calling for an objective validation.

4.1.2. Perspectives on voice naturalness across different research domains

Due to different perspectives and motivations, **voice naturalness** is not uniformly defined across research contexts. First, in **speech-language pathology**, naturalness forms an important rehabilitation outcome in conditions such as stuttering, dysarthria, Parkinson’s disease, developmental communication disorders, and speech prostheses (Anand & Stepp, 2015; Coughlin-Woods et al., 2005; Eadie & Doyle, 2002; Klopfenstein et al., 2020; Mackey et al., 1997; Martin et al., 1984; Meltzner & Hillman, 2005; Yorkston et al., 1990, 1999). In these rehabilitative contexts, it is usually defined as a quality of voice that allows individuals to express their wants and needs efficiently, appropriately, and socially adequately (Klopfenstein et al., 2020). Note that this conceptualization has a subjective component with a strong dependency on the vocal expectations of listeners (Klopfenstein et al., 2020). Second, in **human-robot-interaction**, research is driven by the observation that robots and computers can be perceived as social actors (Nass et al., 1994), whose likability and human-likeness are important factors for user satisfaction and acceptance (Gong, 2008; McGinn & Torre, 2019; W. J. Mitchell et al., 2011; Schweinberger et al., 2020). To the extent that perceptions of naturalness correspond to those of human-likeness, one of the major challenges in the auditory domain is the creation of synthesized speech that sounds natural (Baird, Jørgensen et al., 2018; Baird, Parada-Cabaleiro et al., 2018; Mayo et al., 2011; Nusbaum et al., 1997; Yamasaki et al., 2017). Finally, a similar conceptualization can be found in studies addressing the **methodology of voice research**, which rely on the ecological validity of the stimulus materials used to study human voice perception (Alku et al., 1999; Burton & Blumstein, 1995; Kawahara & Skuk, 2018).

Despite the conceptual heterogeneity of these different perspectives, a surprisingly consistent picture emerges concerning the acoustic features that are associated with perceived naturalness in voices. There is ample evidence that **fundamental frequency variation** is linked to perceived naturalness (Anand & Stepp, 2015; Baird, Jørgensen et al., 2018; Ilves & Surakka, 2013; Vojtech et al., 2019). For example, when comparing different speech synthesis methods, Baird, Parada-Cabaleiro et al. (2018) found a relationship between perceived human-likeness (i.e. naturalness) and F0 variation showing that voices with higher F0 variation are generally rated as more natural than those with lesser variation. Likewise, Anand and Stepp (2015) found F0 variation and naturalness to be highly correlated in patients with Parkinson’s disease. Another important determinant of voice naturalness seems to be the **covariation of F0 and formant frequencies**, which was observed in recorded human speech (Assmann & Katz, 2000). In a subsequent experiment, frequency-shifted speech samples were rated as more natural when they followed this relationship, while utterances were judged to be less natural the more they deviated from it (Assmann et al., 2006). Further, synthetic voices which contain microvariations such as jitter and shimmer are perceived as more natural than those without (Yamasaki et al., 2017). Finally, several studies reported that a low speech rate was associated with a decline in perceived naturalness (Klopfenstein et al., 2020; Mackey et al., 1997; Vojtech et al., 2019; Yorkston et al., 1990).

Despite these initial insights into the acoustic determinants of voice naturalness, little is known about their interplay with vocal emotion perception. The few studies that have investigated the effects of voice naturalness on the processing of lexical emotional content (Ilves & Surakka, 2013; Ilves et al., 2011) have emphasized the importance of vocal naturalness to support transportation of emotional messages. However, to the best of our knowledge, the interaction of voice naturalness with emotional prosody has never been explicitly assessed; a gap we aim to fill with the following two experiments using emotional pseudowords.

4.1.3. Aims of the present studies

In the present studies, we investigated the perceived naturalness and emotionality of parameter-specific voice morphs containing emotional information in either F0 or timbre only, while the other parameter is held at an emotionally non-informative level. In the F0-condition, the emotional F0-contour is combined with the timbre of the non-emotional reference stimulus, and vice versa in the Timbre-condition. This procedure inevitably results in a mismatch between fundamental frequency and formant frequencies, which has been reported to be an important acoustic feature related to voice naturalness (Assmann et al., 2006). Accordingly, we predicted that the F0 and the Timbre conditions would be perceived as less natural compared to a Full condition comprising both parameters. We further considered F0 variance as an important factor of perceived naturalness, based on the literature discussed above. When creating the parameter-specific voice morphs using neutral voices as non-emotional reference, we noted that a neutral voice quality in these recordings was expressed by a monotonous voice with limited F0 variance. This could have a particularly detrimental effect on the Timbre-morphs, where the emotional timbre is combined with the monotonous F0 of the neutral voices. We therefore employed a second morphing approach, using an emotional average as reference, exhibiting more F0 variance. Both neutral voices and averaged emotions have in common that they are assumed to be non-informative with respect to the given emotional quality. We predicted that the naturalness of the Timbre condition could be improved by using the emotional average as reference. In Experiment 1, all stimuli were rated regarding perceived naturalness and emotionality. In Experiment 2, participants chose the more natural-sounding option of two voices that differed only with regard to the morphing reference, to allow a direct comparison of these approaches.

4.2. Experiment 1

4.2.1. Method

Stimuli – Original Audio Recordings The original audio recordings from a database of vocal actor portrayals were provided by Sascha Frühholz, similar to the ones used in Frühholz et al. (2015). For voice morphing, we used three pseudowords (/belam/, /molen/, /loman/) expressing happiness, pleasure, fear, sadness, and produced in a neutral voice by four male and four female speakers.

Stimuli – Voice Morphing Using the Tandem-STRAIGHT software (Kawahara et al., 2008, 2013), we created morphing trajectories between each emotion and a reference stimulus of the same speaker and pseudoword, generating resynthesized vocal samples on these trajectories via weighted interpolation of the originals. Of importance, Tandem-STRAIGHT allows independent interpolation of five different parameters: (1) F0-contour, (2) timing, (3) spectrum-level, (4) aperiodicity, and (5) spectral frequency; the latter three parameters constitute timbre (for a more detailed description see Kawahara & Skuk, 2018).

For the purposes of this study, three different morph types (Morph Types) were created (see Figure 4.1): **Full-Morphs** were stimuli with all Tandem-STRAIGHT parameters taken from the emotional stimulus (corresponding to 100% from the emotion and 0% from reference), except for the timing parameter, which was always taken from the reference (corresponding to 0% emotion and 100% reference). **F0-Morphs** were stimuli with the F0-contour taken from the emotional version, but timbre and timing taken from the reference. **Timbre-Morphs** were stimuli with all timbre parameters taken from the emotional version, but F0 and timing from the reference. Note that the timing was kept constant across all conditions to allow a pure comparison of F0 vs. timbre. As **reference stimuli**, we used two different options (Reference types): we either used the neutral expression or an emotional average of the four emotions. Accordingly, we assumed that both reference types would be uninformative with respect to the expression of one of the four emotions, even if they did not necessarily sound fully non-emotional.

In addition to the Morph Types, the reference stimuli were included in the rating study. In total, this resulted in 8 (speakers) x 3 (pseudowords) x 4 (emotions) x 3 (morphing conditions) x 2 reference types + 48 reference (8 speakers x 3 pseudowords x 2 reference types) = 624 stimuli. Using PRAAT (Boersma, 2018), we normalized all stimuli to a root-mean-square of 70 dB SPL (duration $M = 751\text{ ms}$, $Min = 411\text{ ms}$, $Max = 968\text{ ms}$, $SD = 138\text{ ms}$). A summary of the acoustic properties of the resulting stimuli can be found in Table 4.1, and in more detail in Tables A.1 and A.2. Stimulus examples can be found on <https://osf.io/jzn63/>.

Data collection and participants

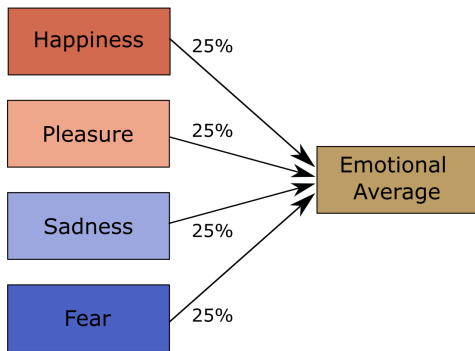
Data were collected online via PsyToolkit (Stoet, 2010, 2017) from February to April 2021. Participants were required to use a computer with a physical keyboard and headphones. As browser, we recommended Google Chrome, and excluded Safari for technical reasons. Participants had to be between 18 and 40 years old, speak German as their native language, have normal hearing abilities and ensure a quiet environment for the duration of the study. All participants had to provide informed consent before completing the experiment, and data were collected completely anonymized. To avoid fatigue, each participant rated only stimuli of one of the pseudowords. Average duration of the experiment was about 30 minutes. Participants who completed the experiment were compensated with course credit. The experiment was in line with the ethical guidelines of the German Society of Psychology (DGPs) and covered by an approval from the ethics committee of the Friedrich Schiller University Jena (Reg.-Rr. FSV 19/045).

Figure 4.1.: Illustration of parameter-specific voice morphs based on two different references

(1) Neutral Reference

Reference	Emotion		Full Morph	Timbre Morph	F0 Morph	Reference Stimuli
Neutral	Happiness	Timbre F0 Timing				
Neutral	Pleasure	Timbre F0 Timing				
Neutral	Sadness	Timbre F0 Timing				
Neutral	Fear	Timbre F0 Timing				

(2) Voice Averaging Process



(3) Average Reference

Reference	Emotion		Full Morph	Timbre Morph	F0 Morph	Reference Stimuli
Emotional Average	Happiness	Timbre F0 Timing				
Emotional Average	Pleasure	Timbre F0 Timing				
Emotional Average	Sadness	Timbre F0 Timing				
Emotional Average	Fear	Timbre F0 Timing				

Note. (1) Morphing matrix for stimuli with an actor portrayal of “neutral” as reference. (2) Schematic depiction of the voice averaging process. (3) Morphing matrix for stimuli with averaged voices as reference.

Table 4.1.: Acoustic properties of the stimulus material used in this study

MType	Ref	F0 _{Mean}	F0 _{SD}	F0 _{Glide}	FormDisp	HNR
<i>Female</i>						
Full	AVG	260	42	-39	1082	20
Full	NEU	258	41	-34	1083	20
F0	AVG	260	42	-39	1095	21
F0	NEU	258	41	-34	1054	19
Tbr	AVG	247	25	-37	1077	20
Tbr	NEU	197	11	1	1075	20
<i>Male</i>						
Full	AVG	173	36	-43	1045	16
Full	NEU	173	36	-47	1041	16
F0	AVG	173	36	-43	1045	16
F0	NEU	173	36	-47	972	15
Tbr	AVG	158	21	-43	1037	16
Tbr	NEU	110	4	0	1041	15

Note. All acoustical parameters were adapted from McAleer et al. (2014) and extracted using PRAAT software (Boersma, 2018) and the F0 contour information from the Tandem-STRAIGHT object in Matlab (MATLAB, 2020). F0 Glide = F0End - F0Start; Formant Dispersion (FormDisp) = ratio between consecutive formant means (from F1 to F4, maximum formant frequency set to 5.5 kHz, window length 0.025 s); HNR (harmonics-to-noise ratio) was extracted with the cross-correlation method (mean value; time step = 0.01 s; min pitch = 75 Hz; silence threshold = 0.1, periods per window = 1.0). Full = full morphs, F0 = F0 morphs, Tbr = Timbre morphs, AVG = average reference, NEU = neutral reference.

Prior to data collection, we conducted a power-analysis using the R-package “Superpower” (Lakens & Caldwell, 2019) with a medium effect size of $f = .23$, an alpha level of .05 and a power of .80 for the interaction of Morph Type and Reference Type on the naturalness ratings, resulting in a required sample size of 16. Since participants rated only stimuli from one pseudoword each, we decided to collect 16 participants per pseudoword, resulting in a total required sample size of 48. This would allow detection of a small effect if data could be collapsed across pseudowords ($f = .13$). The online experiment was accessed by approximately 100 participants, of whom 59 contributed complete data. Of these, eight datasets (13.6%) had to be removed (four participants reported that the sounds were not played properly, three admitted in the post-experimental questionnaire that they had responded randomly, one had a native language other than German). Thus, the final sample consisted of 51 participants (40 females, 11 males, aged 19 to 31 years [$M = 21.49$; $Mdn = 21$; $SD = 2.65$], with 16/18/17 per pseudoword). Three participants reported a minor hearing problem such as occasional tinnitus, but since they reported that they were not limited in their hearing, they were included in the analysis.

Design

Prior to the two rating tasks, participants entered demographic information such as age and sex.

Ratings of perceived naturalness Participants were assigned to one of the three pseudoword-conditions randomly and instructed to rate the naturalness of each voice they heard. They were informed that “natural” in the context of this study meant that “the voices sound human/natural and do not sound distorted or robotic in any way” [German original: “dass sich die Stimmen tatsächlich menschlich/natürlich anhören und nicht auf irgendeine Art verzerrt oder robotisch klingen”]. The participants entered their ratings via keyboard on a 6-point Likert scale with the endpoints 1 = very inauthentic/robotic and 6 = very human. After 8 practice trials with different stimuli, all 208 voice stimuli were presented in randomized order in two blocks of 104 trials each, and participants could take a short break in between. Each trial started with a green fixation cross and after 300 ms the rating stimulus was played. Then, a screen with the 6-point scale was presented and participants had to enter a response within 5000 ms after stimulus offset. If no answer was given in that time, participants were prompted to respond faster by a slide with red lettering for 500 ms. Otherwise, only a black screen was shown, before the next trial started.

Ratings of perceived emotionality After completion of the naturalness ratings, the same stimuli were rated for emotionality on a rating scale from 1 = very negative to 6 = very positive. Note that on this rating scale, 1-3 corresponds to negative and 4-6 to positive valence with different intensity levels. The procedure was identical to the naturalness ratings, except that the voice was played 500 ms (instead of 300 ms) after presentation of the green fixation cross, due to a programming error. Stimuli were presented in a different randomized order.

Post-experimental questions After the experiment, participants were asked whether all sounds were played and whether they had understood the instructions. Furthermore, they could comment on the task and indicate whether they had developed a certain strategy.

Data processing and analysis

Trials of omission ($< .01\%$) were removed. Data were analyzed using R Version 4.1.0 (R Core Team, 2020). Analyses of Variance (ANOVAs) and correlational analyses were performed on averaged rating data, whereas cumulative link mixed models (calculated with the “ordinal”-Package in R, R. H. B. Christensen, 2015) were used to model ratings of single trials. Please note that we interpret our findings based on effect sizes rather than significance values only, in line with recent recommendations (Cumming, 2014; C. O. Fritz et al., 2012). Due to the novelty of the design, our approximate power analysis resulted in a somewhat overpowered design, making even small effects ($d < 0.4$) appear significant, which we nevertheless treated as negligible. Preprocessed data, analysis scripts, stimulus examples and supplemental materials can be found in the associated OSF repository (<https://osf.io/jzn63/>).

4.2.2. Results

Perceived naturalness

Naturalness ratings were averaged across speakers and the reference stimuli (average and neutral) were excluded for the first analysis. Mean ratings were analyzed with a mixed-effects

3×4×3×2 ANOVA with the between-subject factor pseudoword (/belam/, /molen/, /loman/) and the within-subject factors Emotion (happiness, pleasure, fear, and sadness), Morph Type (Full, F0, and Timbre) and Reference Type (NEU and AVG). A summary of all significant main effects and interactions is displayed in Table 4.2 ¹.

Table 4.2.: Results of the 3 x 4 x 3 x 2 mixed-effects ANOVAs on mean ratings of Naturalness and Emotionality

	Naturalness					Emotionality		
	df1	df2	F	p	η_p^2 [95%-CI]	F	p	η_p^2 [95%-CI]
Pseudoword	2	48	4.63	.014	.16 [.01, .34]		-	
Morph Type	2	96	257.78	<.001	.84 [.79, .88]	79.22	<.001	.62 [.51, .71]
Emotion	3	144	54.41	<.001	.53 [.43, .62]	174.69	<.001	.78 [.73, .83]
Pw x Emo	6	144	3.36	.007	.12 [.02, .20]		-	
MType x Ref	2	96	104.05	<.001	.68 [.58, .75]	9.50	<.001	.17 [.05, .29]
MType x Emo	6	288	120.01	<.001	.71 [.66, .75]	121.54	<.001	.72 [.67, .76]
Ref x Emo	3	144	10.42	<.001	.18 [.07, .28]	19.24	<.001	.29 [.17, .40]
Pw x MType x Emo	12	288	2.92	.001	.11 [.02, .15]		-	
Pw x MType x Ref	4	96		-		3.11	.024	.12 [.00, .22]

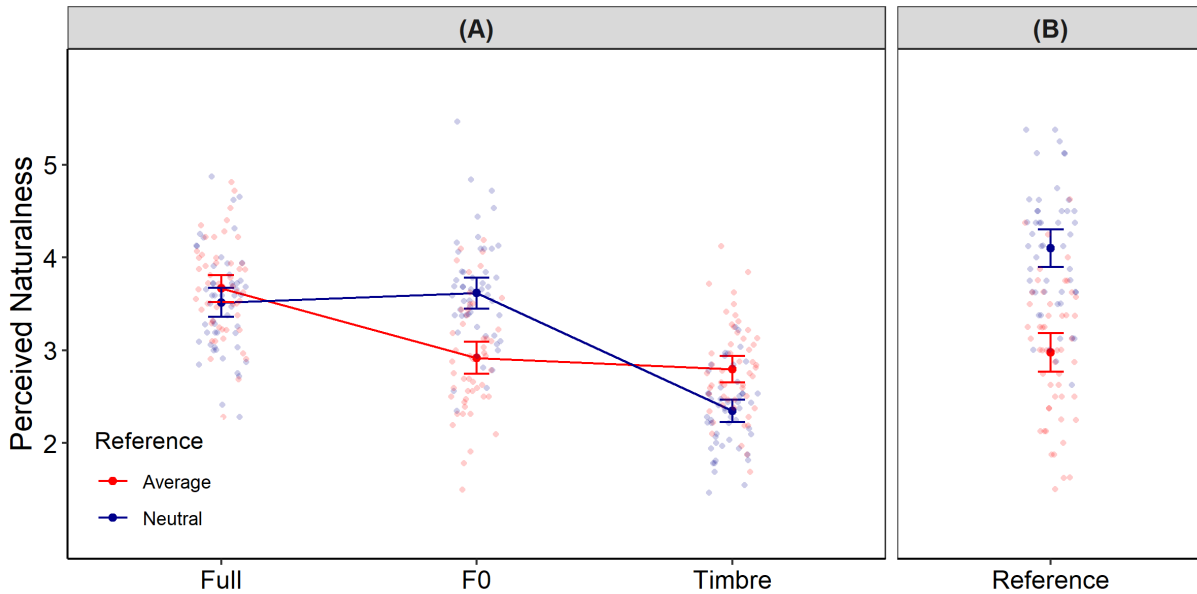
Note. PW = Pseudoword, MType = Morph Type, Ref = Reference Type, Emo = Emotion

Post-hoc tests on the main effect of **Pseudoword** revealed that the pseudoword /molen/ was perceived as less natural than the other two ($|ts(32.69)| \geq 2.46, ps \leq .019$), which did not differ ($t(30) = 0.05, p = .962; M = 3.27 \pm 0.08, M = 3.28 \pm 0.10, M = 2.90 \pm 0.11$, for /belam/, /loman/, and /molen/, respectively). There was a prominent main effect of **Morph Type**, but crucially, this was qualified by an interaction of **Morph Type x Reference Type** (Figure 4.2, A). A comparison of the different Morph Types separately for the two Reference Types revealed the following pattern: With the neutral reference, Timbre-Morphs were rated as substantially less natural than the other two ($|ts(50)| \geq 15.23, ps \leq .001, ds \geq 2.15 [1.65, 2.65]$), whereas Full and F0 differed only marginally ($t(50) = 1.95, p = .057, d = 0.28 [-0.01, 0.56]$). Note that in the Timbre-Morphs, the F0 information is contributed by the reference stimuli, in this case neutral. With the average refence, both Timbre- and F0-Morphs were rated as more unnatural than the Full-Morphs ($|ts(50)| \geq 12.85, ps \leq .001, ds \geq 1.82 [1.36, 2.26]$). However, the difference between them was very small ($t(50) = 2.43, p = .019, d = 0.34 [0.06, 0.63]$). For the same interaction, a comparison of the different Reference Types within each Morph Type revealed that in F0-morphs, stimuli with neutral as reference were perceived as more natural ($t(50) = 9.23, p < .001, d = 1.31 [0.93, 1.69]$), whereas in Timbre- and Full-morphs, stimuli with averaged emotions as reference were perceived as more natural ($|ts(50)| \geq 4.32, ps \leq .001, ds \geq 0.61 [0.31, 0.91]$).

¹Note that the number of participants is slightly unequal for each pseudoword (16/18/17). Therefore, we ran a second analysis where we randomly excluded three participants to have equal group size, resulting in an identical pattern of effects.

The prominent interaction of **Morph Type x Emotion** suggested that while in all Emotions Timbre was rated as more unnatural than the other two, this effect was most pronounced for happiness. For the detailed statistical report including the other main effects and interactions, please refer to [<https://osf.io/jzn63/>]. Finally, a planned comparison between the two reference conditions revealed that the neutral one was rated as substantially more natural than the averaged emotions ($t(50) = 7.53, p < .001, d = 1.06 [0.71, 1.41]$, refer to Figure 4.2, B).

Figure 4.2.: Interaction of Morph Type and Reference Type on perceived Naturalness

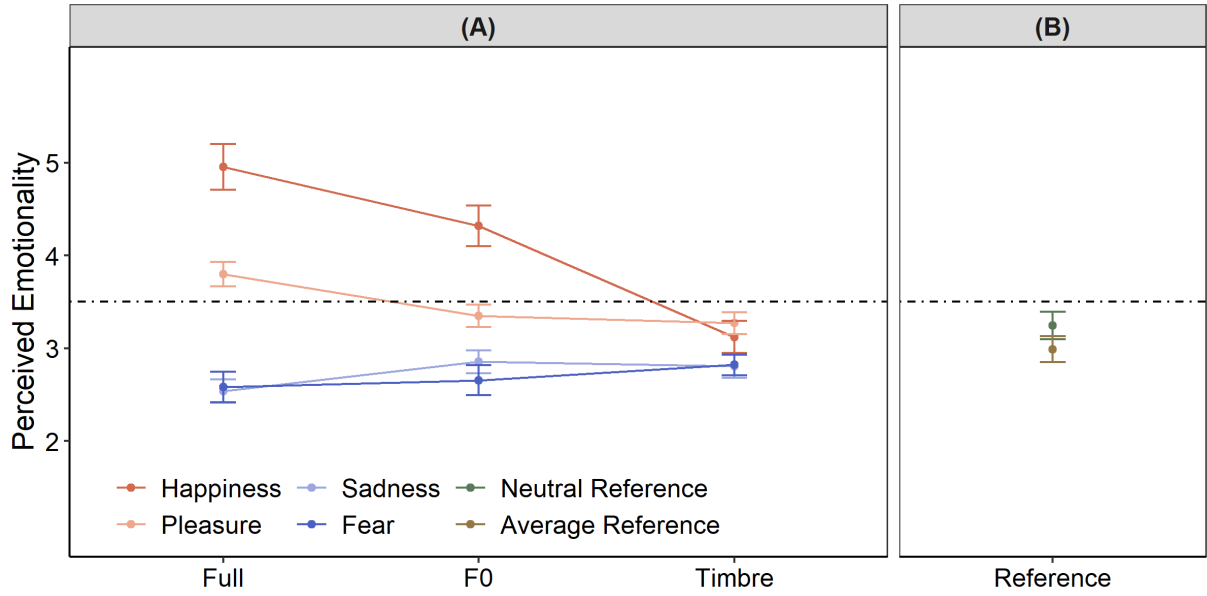


Note. Whiskers represent 95%-confidence intervals. Dots represent individual participants' data.

Perceived emotionality

Similar to the naturalness ratings, mean ratings of emotionality were analyzed with a mixed-effects $3 \times 4 \times 3 \times 2$ ANOVA with the between-subject factor pseudoword (/belam/, /molen/, /loman/) and the within-subject factors Emotion (happiness, pleasure, fear, and sadness), Morph Type (Full, F0, and Timbre) and Reference Type (NEU and AVG), refer to Table 4.2. The main effects of **Morph Type** and **Emotion** were qualified by a prominent interaction (Figure 4.3, A). While in Full-Morphs, emotionality ratings were widespread, this range was reduced in F0-Morphs and almost absent in the timbre condition. Thus, emotions could be much better discriminated in the F0 compared to the Timbre condition, suggestion that F0 is more effective in signaling emotional quality, although not as informative as Full Morphs. Crucially, this pattern was not further qualified by the Reference Type, suggesting it does not depend on the morphing reference used. For the detailed statistical report including the other main effects and interactions, see [<https://osf.io/jzn63/>]. A planned comparison between the two reference conditions revealed that the averaged emotions were rated a bit more negative than the neutral ones, but this effect was small ($t(50) = 2.72, p = .009, d = 0.38 [0.10, 0.67]$, refer to Figure 4.3, B).

Figure 4.3.: Interaction of Morph Type and Emotion on perceived Emotionality



Note. The dashed line marks the midpoint of the rating scale at 3.5. Whiskers represent 95%-confidence intervals.

Relationship between perceived naturalness and emotionality

To assess whether the perception of naturalness and emotionality was linked, we averaged both ratings across participants to correlate mean ratings of each stimulus. We found no relationship, $r(624) = -.043, p = .279$, suggesting that we observe both natural and unnatural stimuli across all levels of emotional quality (Figure 4.4, A). However, since emotional valence and intensity are combined in our emotionality rating, we ran a second analysis in which we were interested in the link between naturalness and emotional intensity. To this end, we recoded our rating scores such that responses of 1 or 6 corresponded to high intensity ($= 3$), 2 or 5 to medium intensity ($= 2$), and 3 or 4 to low intensity ($= 1$). However, there was no correlational relationship either, $r(624) < .001, p = .988$ (Figure 4.4, B). Thus, we concluded that perceived naturalness and perceived emotional quality/intensity were not related in our stimuli.

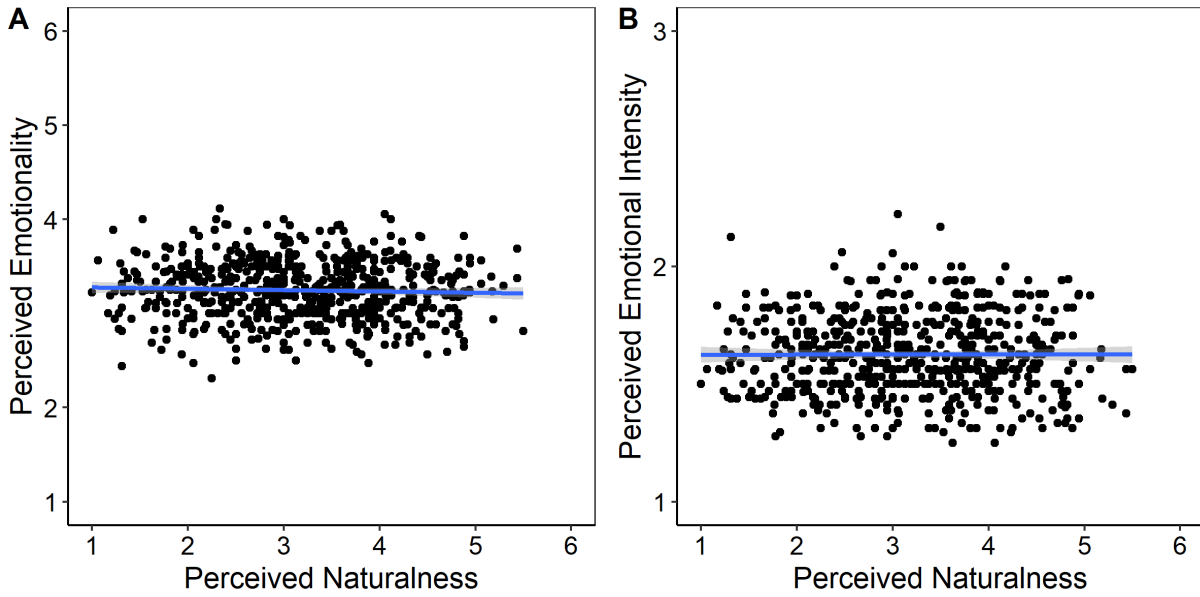
Link between ratings and acoustic properties of the stimuli

For modelling the influence of acoustic properties on the perception of naturalness and emotionality in the stimuli (including neutral and average reference stimuli), the standardized predictor variables $F0_{\text{Mean}}$, $F0_{\text{SD}}$, $F0_{\text{Glide}}$, formant dispersion (FormDisp) and harmonics-to-noise ratio (HNR) were chosen to calculate two cumulative link mixed models with the syntax

$$\begin{aligned} \text{Rating} \sim & F0_{\text{Mean}} + F0_{\text{SD}} + F0_{\text{Glide}} + \text{FormDisp} + \text{HNR} \\ & + (1|\text{Participant}) + (1|\text{SpID}) \end{aligned} \quad (4.1)$$

on ratings of naturalness and emotionality separately. The results are summarized in Table 4.3.

Figure 4.4.: Relationship between mean ratings of perceived naturalness and emotionality (A), or emotional intensity (B)



Note. Data points represent mean ratings of individual stimuli averaged across participants. The blue line illustrates the linear regression, the shaded grey area around it the standard error.

In short, all included parameters seemed to play a significant role for both ratings, except for Formant Dispersion in the context of emotionality ratings. For both ratings, the biggest effect was observed for F0 variability.

Table 4.3.: Results of the regression analyses using cumulative link mixed models

	Naturalness				Emotionality			
	β	SE	z	p	β	SE	z	p
F0 _{Mean}	-0.551	0.044	-12.59	<.001	-0.608	0.047	-12.88	<.001
F0 _{SD}	0.596	0.035	17.21	<.001	1.182	0.040	29.77	<.001
F0 _{Glide}	-0.228	0.021	-10.85	<.001	-0.275	0.022	-12.38	<.001
FormDisp	-0.070	0.027	-2.64	.008	0.012	0.027	0.44	.663
HNR	0.387	0.031	12.68	<.001	-0.232	0.031	-7.44	<.001

4.2.3. Short summary

In Experiment 1, we showed that perceived naturalness was affected by the choice of the morphing reference, whereas perception of emotionality was not. In fact, we did not find evidence for a relationship between perception of naturalness and emotionality. In a regression analysis, most of the vocal parameters we took into consideration predicted ratings of both naturalness and emotionality, with the biggest effect observed for F0 variability.

4.3. Experiment 2

In a second Experiment, we aimed to replicate and expand the findings of Experiment 1 with a different paradigm. In a two-alternative forced-choice task, participants listened to pairs of corresponding stimuli (same speaker, emotion, pseudoword, and morph type) which only differed in the morphing reference. Their task was to decide which sample sounded more natural. This provided a direct comparison of morphing approaches.

4.3.1. Method

Stimuli

Stimuli were identical to Experiment 1.

Data collection and participants

Data were collected online via PsyToolkit (Stoet, 2010, 2017) from May to July 2021, with the same general conditions and inclusion criteria as in Experiment 1. Average duration of the experiment was about 35 minutes. The online experiment was accessed by approximately 65 participants of whom 34 contributed complete data. Of these, six datasets (17.4%) had to be removed (two participants reported that the sounds were not played properly, three exceeded the age range of 18-40, one had >5% trials of omission). Thus, the final sample consisted of 28 participants (14 females, 14 males, aged 18 to 30 years [$M = 22.39$; $Mdn = 22$; $SD = 2.75$]).

Design

Each trial started with a fixation cross for 500 ms. Afterwards, a black screen with two loudspeaker symbols labelled “1” and “2” appeared. Then the first sound was played, visually highlighted with the first sound symbol turning green. After an inter-stimulus interval of 750 ms, the second sound was played, with the other sound symbol turning green. The participants decided via keypress ($f = 1$, $j = 2$) which voice sounded more natural, in a time window of 5000 ms after the second stimulus offset. Within trials, the two stimuli were of the same speaker, emotion, pseudoword, and morph type, and differed only in reference type (AVG/NEU). Trials with the neutral and average reference stimuli were included as well. Whether AVG or NEU was presented first was randomized. After 6 practice trials with different stimuli, all 312 voice pairs were presented in randomized order in four blocks of 52 trials each, and participants could take short breaks between blocks.

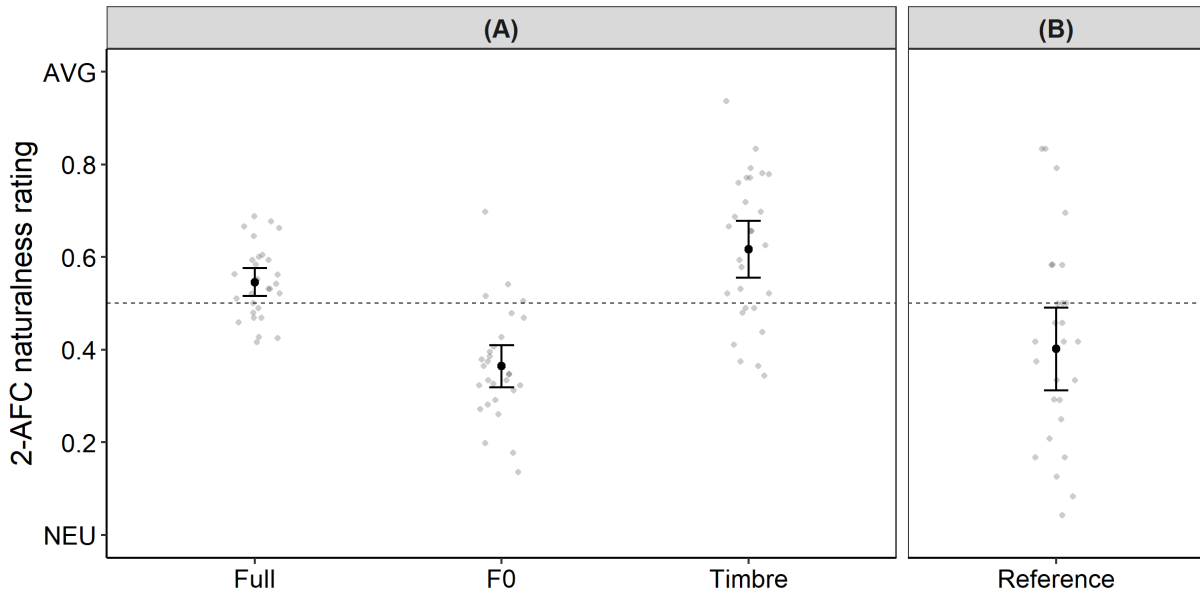
Data processing and analysis

Trials of omission (< .01%) were removed. Data were transformed to display the response tendency as the proportion of “average sounds more natural”-responses. The rest of the data analysis pipeline was comparable to Experiment 1.

4.3.2. Results

Responses were averaged across speakers and pseudoword and trials with reference stimuli (average/ neutral) were excluded for the first analysis. A 3×4 repeated-measures ANOVA on the response tendency revealed main effects of both factors **Morph Type**, $F(2, 54) = 45.34, p < .001, \eta^2 = .63$ [0.46 0.73]; and **Emotion**, $F(3, 81) = 11.39, p < .001, \eta^2 = .30$ [0.13, 0.43]. Post-hoc analyses revealed that in Full and Timbre morphs, average-referenced stimuli were perceived as more natural, whereas in F0 morphs the neutral option was chosen more often ($|ts(27)| \geq 2.51, ps \leq .019, |ds| \geq 0.48$ [0.08, 0.88], see Figure 4.5, A). This pattern was further supported by a planned comparison against 0.5 as the point without a response tendency ($|ts(27)| \geq 3.09, ps \leq .004, |ds| \geq 0.60$ [0.18, 1.00]). In the trials comparing average and neutral stimuli directly, neutral stimuli were chosen more often (test against 0.5: $t(27) = -2.28, p = .031, |d| = 0.44$ [0.04, 0.83], see Figure 4.5, B). This represents a full replication of the pattern found in Experiment 1 (refer to Figure 4.2). The main effect of Emotion was mainly driven by happiness, which was perceived as more natural with average reference, and sadness, which was perceived more natural with neutral reference ($M_{Happiness} = .55 \pm 0.02; M_{Pleasure} = .51 \pm 0.02; M_{Fear} = .50 \pm 0.02; M_{Sadness} = .47 \pm 0.02$; detailed analysis on <https://osf.io/jzn63/>).

Figure 4.5.: Response tendency towards the more natural reference category as a function of Morph Type



Note. The dashed lined represents the 0.5 point with no response tendency. Whiskers represent 95%-confidence intervals. Grey dots represent individual participants' data. 2-AFC = two alternative-forced choice task.

4.3.3. Short summary

Experiment 2 employed a two alternative-forced choice task to provide a conceptual replication of Experiment 1 regarding the perception of naturalness as a function of morphing reference. For

Full and Timbre morphs, average-referenced emotional voices were perceived as more natural than neutral-referenced emotional voices, whereas the opposite was found for F0 morphs.

4.4. Discussion

The present experiment explored a number of important determinants of the perception of naturalness and emotionality in voices. Specifically, we investigated how the impression of naturalness in voices is formed, how it can be affected by different voice manipulations (especially those related to parameter-specific voice morphing), and how it can interact with the perception of vocal emotions. In line with our hypotheses, we observed that voice manipulation affected perceived naturalness, presumably due to an inherent mismatch between fundamental frequency contour and timbre features. Perceived naturalness was also strongly affected by fundamental frequency variation: On the one hand, naturalness could be tremendously improved in the Timbre condition by using the average emotion as reference, which expressed much more F0 variation than the neutral one. On the other hand, a regression analysis revealed F0 variation to be an important predictor of both naturalness and emotionality ratings. Most importantly, we found no evidence that emotionality ratings were affected by a lack of stimulus naturalness, suggesting that stimuli like the ones used here are valid for vocal emotional research. In what follows, we discuss how these findings relate to (a) the role of naturalness in emotion perception, (b) the possible existence of an uncanny valley for voices, and (c) the potentials and limits of emotional voice morphing.

4.4.1. The role of naturalness in the perception of emotion and other social signals

Although the communication of emotion is limited to humans and living creatures, emotional processing per se is not. In 1944, Heider and Simmel (1944) presented a short film with geometrical figures moving on the screen and asked participants to describe it. Intriguingly, most of the participants provided a description of animated beings with personalities, backstories, and emotions. One figure was consistently perceived as aggressive and angry, whereas another was perceived as frightened. This shows that humans attribute human traits and emotions to non-living objects. In fact, our brain displays a strong tendency to pick up and process emotions, even in highly artificial settings (Hortensius et al., 2018; Spatola & Wudarczyk, 2021). This property is deliberately employed for improving communication with non-human actors, such as robots (Crompton & Bethel, 2016). Thus, emotional processing may not depend on naturalness or human-likeness. Our data fit into this line of argumentation, by showing that the processing of emotionality was remarkably unrelated to the perceived naturalness of voices. It is noteworthy that in the facial domain, Calder et al. (2000) observed a comparable pattern using emotional caricatures: With increasing caricaturing level, faces were rated as more emotionally intense, despite being perceived as less natural. Based on these findings, one could assume that emotional processing can suppress any disruptive effects of unnaturalness or artificial circumstances.

However, both theoretical considerations and conflicting empirical evidence suggest that this might not be entirely true: Models of both face and voice perception suggest that voices and faces

are “special” to the brain, in the sense that they recruit neural resources which are not recruited by other types of stimuli (Belin et al., 2011; A. W. Young & Bruce, 2011). With stimulus material deviating profoundly in human-likeness, recruitment of these networks might be disrupted. Evidence from the domain of face recognition that computer-generated faces do not fully tap face expertise (Crookes et al., 2015) could indicate that the same might hold for emotional processing. Indeed, in studies using electroencephalography (EEG), stimulus naturalness and emotionality interact at the neural level for both faces and voices (Schindler et al., 2017; Schirmer & Gunter, 2017). Further, the human voice is perceived as more expressive and likeable than an expressive synthetic one (Cabral et al., 2017; Ilves & Surakka, 2013). Finally, adaptation paradigms using sine tones or F0-removed stimuli as adaptors fail to elicit reliable adaptation aftereffects in voices (Hubbard & Assmann, 2013; Schweinberger et al., 2008), presumably due to their unnatural/non-human quality.

Taken together, these findings imply that naturalness of the stimulus materials does play a role in emotional processing. Yet, the circumstances under which somewhat unnatural stimuli still allow a direct generalization to perception of real human voices remain unclear. As it stands, emotional processing can to some degree disregard unnatural features but is likely not completely detached from them. Thus, it remains the responsibility of researchers to give this matter explicit consideration for specific voice stimulus sets. For the voice stimuli used in the present study, naturalness does not seem to play a crucial role for emotional processing.

4.4.2. Is there an uncanny valley for voices?

A question related to the interplay of naturalness and emotion is the existence of an uncanny valley for voices. The uncanny valley, originally proposed by Mori in 1970 (Mori et al., 2012), has been described as a sudden drop in likability of robots that almost approach, but do not entirely reach a human-like appearance. This almost human-like quality is assumed to evoke a sudden feeling of eeriness, although its empirical evidence remains inconsistent (Kätsyri et al., 2015). So far, this effect has been observed for static and dynamic visual depictions of robots, as well as for a mismatch of human-likeness between the auditory and the visual channel (W. J. Mitchell et al., 2011; Schweinberger et al., 2020). However, the presence of an uncanny valley for voices alone remains elusive. For the present investigation this could be highly relevant, since emotional voice morphs, which are resynthesized from human voices, could fall into an “almost human-like” gap which results in the uncanny valley phenomenon. However, so far, there is no evidence for an uncanny valley in voices. Previous studies only found a linear relationship between human-likeness and likability (Baird, Parada-Cabaleiro et al., 2018) and in the present data we observed no patterns that would suggest anything else.

4.4.3. Emotional voice morphing – a tool of unlimited possibilities?

In the past, research linking voice acoustics to socio-emotional signals was predominantly based on correlational inference. This has improved through the development of voice manipulation tools, such as voice morphing (Kawahara & Skuk, 2018). While offering exciting research prospects, the degrees of freedom allowed by this method are both tempting and intimidating, especially when

morphing vocal emotional utterances: It is possible to morph between two emotions of choice (Nussbaum, von Eiff et al., 2022; von Eiff et al., 2022), or to morph one emotion with respect to a reference, which in turn can be non-emotional (i.e. neutral) or emotionally ambiguous (i.e. average). If an emotional average is used, consideration should be given to the emotions that enter into this average: An average comprised of the six basic emotions (Ekman, 1992) would sound different from the one used in the present experiments, composed of two negative and two positive emotions. Further, voice averaging itself constitutes a special form of voice morphing which is still in its infancy and technically very challenging (Kawahara & Skuk, 2018). The more voices enter an average, the more prone it is to stimulus artifacts such as reduced aperiodicities and higher harmonics-to-noise ratios (Bruckert et al., 2010). This could make the average sound less natural than original human recordings, a pattern we observed in both Experiments, when our averages were compared to neutral voices (cf. Figure 4.2 and 4.5, B). Further, one can not only interpolate between voices, but also extrapolate and thus create emotional caricatures (Whiting et al., 2020). Finally, morphing allows a parameter-specific manipulation of the voice, as for F0 and timbre in the present study. While undeniably powerful, all these options carry a potential to affect empirical findings to a substantial degree, making them hard to compare across studies – an important caveat when designing and interpreting voice morphing studies.

For faces, Calder et al. (2000) demonstrated that perception of emotional caricatures was comparable when they were created with respect to a neutral, an averaged or a different emotional face. For voices, the present data also confirm that the perception of emotion was not substantially affected by the choice of morphing reference. Still, many of our methodological choices are likely to have impacted on our results: First, our design was limited to four emotions balanced in valence and we included only these to create the emotional average. Second, we specifically focused on the contrast of F0 and timbre as vocal parameters. We found that F0 played a larger role than timbre in emotion discrimination, in line with previous research in the normal-hearing population (Nussbaum, Schirmer & Schweinberger, 2022). However, we would not claim that this would necessarily generalize to a different set of emotions. Third, we showed that even though different voice morphing approaches did not affect emotional ratings, they affected perceived naturalness.

In both experiments, F0 morphs were perceived as more natural with neutral reference, but Timbre and Full morphs were perceived as more natural with average reference. The effect of the average reference on the Timbre morphs was predicted, because of its increased F0 variation (Baird, Parada-Cabaleiro et al., 2018; Vojtech et al., 2019). More importantly however, perceived naturalness between F0 and Timbre was comparable in the average-referenced condition only, thus excluding differences in naturalness as a potential confound when comparing the two. This clearly advocates the average-referenced approach as more suitable for research contrasting these two parameters, since naturalness and emotional processing may interact at the neural level, as discussed above. Altogether, the present investigation demonstrates both the potentials and the pitfalls of emotional voice morphing and encourages an explicit consideration of its methodological subtleties.

4.4.4. Directions for future research

The present investigation only provides a starting point in understanding the role of naturalness in the context of voice perception and emotional voice processing. For example, without further investigation, we can only speculate how stimulus naturalness might affect perception of emotions other than the ones that were studied in our experiments. Further, while several studies comment on the acoustic quality of their stimulus material (Grichkovtsova et al., 2012; Nussbaum, von Eiff et al., 2022; Skuk et al., 2015), objective research efforts to validate stimulus material with respect to such aspects remain sparse. In this context, it is important to note that perceived naturalness may not be a function of physical stimulus properties alone but can also be affected by perceptual exposure and adaptation. For instance, it is well-known that a sufficient degree of adaptation to highly unnatural (e.g., spatially expanded or compressed) faces can make subsequent faces of the same distortion appear far more natural (Kloth et al., 2017; Webster & MacLin, 1999). Future research will have to elucidate the psychological and neuronal mechanisms by which perceptual experience with morphed stimuli (by experimental participants, but potentially also by researchers who are in daily contact with such stimuli) may affect perceptions of naturalness. More generally, with the present experiments, we hope to inspire more research on naturalness and its impact on the processing of different social signals in the vocal and facial domain. This could offer insight into the processing of both human and non-human signals, making valuable contributions to psychological models of person perception as well as human-robot interaction.

4.5. Summary and conclusion

In two experiments, we explored the impact of parameter-specific voice morphing on the perception of naturalness and emotionality. We compared Full, F0 and Timbre morphs of emotions based on two different morphing references, neutral and average. In line with our hypotheses, we found that parameter-specific voice morphing affected perceived naturalness. In F0 morphs, stimuli with neutral as reference were perceived as more natural, while Timbre and Full morph stimuli were perceived as more natural with averaged emotions as reference. Crucially, naturalness of F0 and Timbre morphs was comparable only in the average-reference condition, making this form of reference more suitable for future research. Finally, we found no relationship between ratings of emotionality and naturalness. This suggests that perceived emotionality was not extensively affected by a lack of stimulus naturalness and that parameter-specific voice morphing is thus a suitable tool for vocal emotional research.

5. Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates

This chapter has been published as:

Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2022). Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates. *Social Cognitive and Affective Neuroscience*, 17 (12), 1145–1154. Copyright © 2022 (Oxford University Press) DOI: <https://doi.org/10.1093/scan/nsac033>

Abstract

Our ability to infer a speaker's emotional state depends on the processing of acoustic parameters such as fundamental frequency (F0) and timbre. Yet, how these parameters are processed and integrated to inform emotion perception remains largely unknown. Here we pursued this issue using a novel parameter-specific voice morphing technique to create stimuli with emotion modulations in only F0 or only timbre. We used these stimuli together with fully modulated vocal stimuli in an event-related potential (ERP) study in which participants listened to and identified stimulus emotion. ERPs (P200 and N400) and behavioral data converged in showing that both F0 and timbre support emotion processing but do so differently for different emotions: Whereas F0 was most relevant for responses to happy, fearful and sad voices, timbre was most relevant for responses to voices expressing pleasure. Together, these findings offer original insights into the relative significance of different acoustic parameters for early neuronal representations of speaker emotion and show that such representations are predictive of subsequent evaluative judgments.

5.1. Introduction

It is well established that listeners readily infer a speaker's emotional state based on the speaker's voice acoustics (Banse & Scherer, 1996; Juslin & Laukka, 2003). Yet, after over 30 years of research, and in some contrast to the accuracy with which listeners infer vocal emotions, the identification of emotion-specific acoustic profiles has been only partially successful (Banse & Scherer, 1996; Brück et al., 2011; Juslin & Laukka, 2003). Specifically, it remains uncertain how different vocal cues such as fundamental frequency and timbre are processed in the listener's brain to inform emotional inferences (Frühholz & Schweinberger, 2021; Frühholz et al., 2016). Here, we review past efforts and identify important conceptual and methodological challenges (Gobl, 2003;

A. D. Patel, 2011; Scherer, 1986). We address these challenges by complementing earlier work with a parameter-specific voice morphing approach that specifically manipulates individual vocal cues. We focus on fundamental frequency contour and timbre to understand the mechanisms by which they influence neural integration and subsequent behavioral responses in vocal emotions.

5.1.1. The role of different acoustic parameters in vocal emotion perception

That listeners can infer emotions from voices with remarkable accuracy has prompted the assumption that different emotions are characterized by distinct patterns of acoustic parameters (Banse & Scherer, 1996; Juslin & Laukka, 2003; Paulmann & Kotz, 2018). To date, the literature has focused on four groups of parameters including (i) fundamental frequency contour (F0), (ii) amplitude, (iii) timbre and (iv) temporal aspects. Indeed, all these parameters have been found to be important in signaling emotional quality (Juslin & Laukka, 2003). However, despite enormous efforts, a potential mapping of vocal parameters to specific emotions remains elusive. For instance, anger, fear and happiness have all been linked to a high F0 mean and variability, a large amplitude and a fast rate of articulation, whereas the opposite was found for sadness (Banse & Scherer, 1996; Brück et al., 2011; Juslin & Laukka, 2003; Lausen & Hammerschmidt, 2020; Lima & Castro, 2011). These findings seem to reflect that vocal parameters signal unspecific arousal rather than more differentiated emotional states and thus fail to account for listener performance (Brück et al., 2011). Here, we consider this apparent paradox, suggesting that methodological challenges inherent in the study of natural speech may preclude insights into the functional significance of different acoustic parameters. In what follows, we will outline these challenges focusing on difficulties associated with the interpretation of correlational data, the selection of relevant parameters and the partial redundancy of vocal cues.

Past research typically measured a set of acoustic parameters and used the obtained measures to study differences between emotional categories or to predict listener responses (Banse & Scherer, 1996; Juslin & Laukka, 2003; Lima & Castro, 2011). However, this approach is intrinsically correlational and does not allow for causal inference. Therefore, Arias et al. (2021) explicitly called for voice manipulation techniques to gain control over the acoustic properties expressing vocal emotions. An experimental elimination of the natural covariation between specific auditory parameters and emotion quality could prove particularly beneficial in research on event-related potentials (ERPs), where dissociating sensory from emotional responses poses a major challenge (Paulmann et al., 2013; Schirmer et al., 2013). Parameter-specific voice morphing has been recently established as a suitable tool to study how different acoustic cues facilitate the perception of speaker age, sex, and identity (Kawahara & Skuk, 2018; Skuk et al., 2015, 2020). Applications in the domain of vocal emotion perception are still sparse but offer great potential (Nussbaum, von Eiff et al., 2022; von Eiff et al., 2022).

When choosing the vocal parameters under study, the majority of research focused on measuring F0, a perceptually dominant parameter, which is relatively easy to measure. However, it has been widely acknowledged that other parameters, in particular timbre, may be equally important but have been rarely considered (Banse & Scherer, 1996; Gobl, 2003; S. Patel et al., 2011).

Defined as “the difference between two voices of identical F0, intensity and temporal structure” (ANSI, 1973), timbre reflects a complex combination of several parameters, including formant frequency and bandwidth, high spectral energy and spectral noise (Juslin & Laukka, 2003; Lima & Castro, 2011). Timbre perception is likely based on an integration of all its features (Piazza et al., 2018), and previous works that studied timbre suggest a central role of this parameter in voice processing (Gobl, 2003; Nussbaum, von Eiff et al., 2022; Skuk et al., 2015; Tursunov et al., 2019). In particular, Grichkovtsova et al. (2012) found that both timbre and prosodic contour carry unique information for different emotions.

Finally, the idea that universal acoustic patterns signal discrete emotions discounts a central aspect of our perceptual system: flexibility. In fact, Spackman et al. (2009) showed that marked vocal and expressive differences between speakers have little impact on listeners’ ability to infer emotions, suggesting that listeners flexibly adapt their inferential processes to a speaker’s overall vocal profile. Conceptually, this flexibility is captured in Brunswik’s lens model (Brunswik, 1956), in which acoustic cues are understood as probabilistic and partly redundant. Crucially, decoders are thought to rely on these cues in a partly interchangeable manner (Juslin & Laukka, 2003). Thus, simply comparing different acoustic parameters with respect to their significance or predictive value for emotional judgments can be very misleading if their contribution is implicitly assumed to be non-redundant. Instead, this can be made explicit by exploring to which degree a particular vocal parameter carries unique information that cannot be transported by other parameters. Notably, this may be achieved by creating voices expressing emotions through only one parameter while other parameters are held at a non-informative neutral level.

5.1.2. Electrophysiological correlates of vocal emotion perception

Although distinct neural networks involved in the processing of different acoustic parameters have been discussed for voice and speech perception, e.g. a lateralization of pitch and timing information (Belin et al., 2011; Poeppel, 2001), this has rarely been linked to emotional processing. Likewise, while current models on the neural processing of vocal emotions emphasize the importance of monitoring and integrating relevant acoustic cues in real time (Frühholz et al., 2016), it is not yet understood how this takes place for specific vocal parameters in different emotions. To this point, research using electroencephalography (EEG) highlights different processing stages that unfold dynamically across time (Paulmann & Kotz, 2018; Schirmer & Kotz, 2006). The initial analysis of acoustic features presumably already modulates the N100 component, whereas subsequent emotional salience has been linked to later processes at around 200 ms following stimulus onset as indexed, for example, by the P200 (Paulmann & Kotz, 2008; Paulmann et al., 2013; Pell et al., 2015; Schirmer & Gunter, 2017; Schirmer, Kotz & Friederici, 2005). Finally, top-down and goal-directed vocal analyses seem to involve mechanisms associated with the N400 or the late positive component (Paulmann & Kotz, 2018). All these ERP components, especially the N100 and the P200, are sensitive to changes in vocal parameters such as pitch and loudness, but to date, it is unclear how these acoustics are integrated specifically to derive emotional meaning (Paulmann & Kotz, 2018).

5.1.3. Aims of the present study

Although the importance of individual acoustic parameters for emotion perception is widely recognized, these parameters have been rarely pursued experimentally and, to the best of our knowledge, not in the context of functional neuroimaging. The present study sought to address this gap and to answer the following two questions: (1) What are the unique contributions of F0 vs timbre to the perception of specific vocal emotions and (2) how does the neural processing of these parameters unfold in time? To this end, we used parameter-specific voice morphing to create F0-only and timbre-only morphs, which contained emotional information in only one of these parameters. Additionally, we created Full morphs, which encompassed emotional information from both F0 and timbre. Participants listened to all stimuli in random order and were asked to classify speaker emotion, while their EEG was being recorded.

For the emotion classification performance, we predicted that compared to a condition with full emotional information, accuracy in both parameter-specific conditions would be inferior since both F0 and timbre carry unique information important for successful emotional decoding. However, we speculated that the relative importance of F0 vs timbre would differ as a function of emotion. With respect to the EEG, we were particularly interested in evidence regarding the temporal pattern of F0 vs timbre processing. In an exploratory cluster-based permutation analysis, we examined a time range from 0 to 500 ms following voice onset to detect potential modulations in both earlier (N100/P200) and later (N400) ERP components, speculating that such modulations could be relevant in predicting parameter-specific behavioral responses.

5.2. Method

5.2.1. Listeners

Based on prior behavioral data (Grichkovtsova et al., 2012), we conducted a power analysis using the R-package “Superpower” (Lakens & Caldwell, 2019) with a medium effect size of $f = 0.13$, an alpha level of 0.05 and a power of 0.80 for the interaction of Emotion and Morph Type on recognition accuracy, resulting in a required sample size of 36. We collected data from 44 healthy native German speakers with no hearing impairments, as confirmed by a short audio test (Cotral-Labor-GmbH, 2013). All participants were students at the Friedrich Schiller University of Jena. Sessions lasted about 2.5h. Participation was compensated with course credit or 8.50€/h. The experiment was approved by the ethics committee of the Friedrich Schiller University of Jena. The data from five participants had to be excluded (three had >3% of missing trials and two had <80% correct in the word naming task). The final sample consisted of 39 participants [27 females and 12 males, aged 18–29 years ($M = 22.41$; $Mdn = 22$; $SD = 2.92$), 2 left-handed].

5.2.2. Stimuli

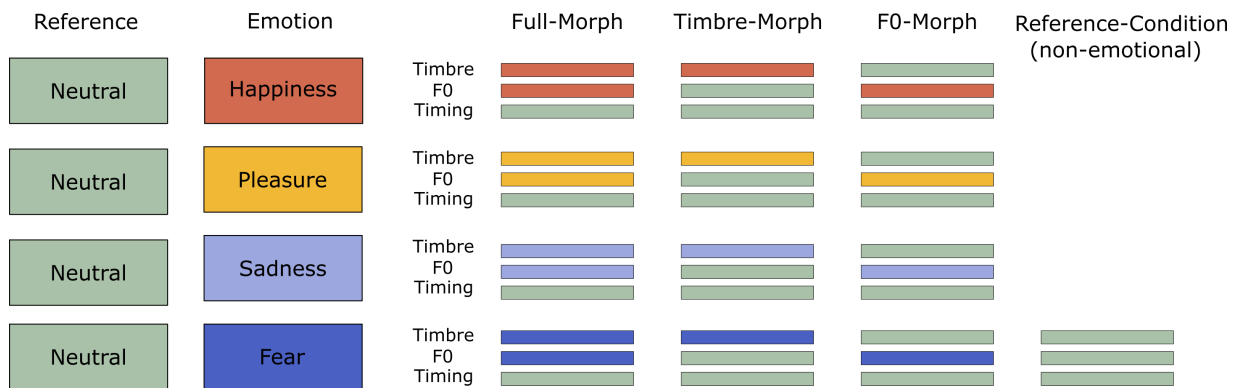
Original audio recordings We selected original audio recordings from a database of vocal actor portrayals provided by Sascha Frühholz from the Department of Cognitive and Affective

Neuroscience of the University of Zurich, similar to the ones used in Fröhholz et al. (2015). For the present study, we used three pseudowords (/molen/, /loman/ and /belam/) with expressions of happiness, pleasure, fear, sadness and neutral. We opted for two positive and two negative emotions for various reasons, including that previous studies often focused on happiness as the only positive emotion and that comparing only one positive and one negative emotion would have enabled only valence-based (i.e. positive vs negative) insights. Stimuli were validated after applying the voice morphing procedure in an independent rating study with 20 raters, including more emotions and morph levels. Based on these ratings, we selected two positive and two negative emotions with different degrees of intensity [happiness vs pleasure: $t(19) = -9.57$, $p < .001$, with $M_s = 3.40 \pm 0.06$ and $M_s = 2.88 \pm 0.07$; fear vs sadness: $t(19) = 6.58$, $p < .001$, $M_s = 3.01 \pm 0.06$ and $M_s = 2.78 \pm 0.07$; on a rating scale ranging from 1 to 4]. For the complete documentation of the rating study, refer to the supplemental material in Appendix C.

Voice morphing Using the Tandem-STRAIGHT software (Kawahara et al., 2008, 2013), we created morphing trajectories between each emotion and the neutral expression of the same speaker and pseudoword. After manual mapping of time and frequency anchors at key features of a given utterance pair (e.g. onset and offset of vowels), vocal samples on an emotion/neutral continuum were synthesized via weighted interpolation of the originals; for a more detailed description see Kawahara and Skuk (2018). Crucially, Tandem-STRAIGHT allows independent interpolation of five different parameters: (i) F0 contour, (ii) timing, (iii) spectrum level, (iv) aperiodicity and (v) spectral frequency; the latter three are summarized as timbre.

Three types of morphed stimuli were created (Figure 5.1). **Full-Morphs** were stimuli with all Tandem-STRAIGHT parameters taken from the emotional version (corresponding to 100% from the emotion and 0% from neutral), with the exception of the timing parameter, which was taken from the neutral version (corresponding to 0% emotion and 100% neutral).

Figure 5.1.: Schematic illustration of the different parameter-specific voice morphs



Note. Parameters encompassing emotional information were morphed using 100% from the emotional utterances and 0% from the neutral one, and parameters encompassing neutral information vice versa, respectively.

F0-Morphs were stimuli with the F0 contour taken from the emotional version, but timbre and timing taken from the neutral version. **Timbre-Morphs** were stimuli with all timbre parameters taken from the emotional version, but F0 and timing from the neutral version. In addition, all original neutral stimuli were included as an extra non-emotional reference category. Note that the timing was kept constant across all conditions to allow a pure comparison of F0 vs timbre. In total, this resulted in $8 \text{ (speakers)} \times 3 \text{ (pseudowords)} \times 4 \text{ (emotions)} \times 3 \text{ (morphing conditions)} + 24 \text{ neutral (8 speakers} \times 3 \text{ pseudowords)} = 312 \text{ stimuli}$. For analysis purposes, we collapsed data across speakers and pseudowords.

Using Praat (Boersma, 2018), we normalized all stimuli to a root mean square of 70 dB sound pressure level (duration $M = 670 \text{ ms}$, $\min = 411 \text{ ms}$, $\max = 878 \text{ ms}$). Please refer to Tables A.1 and A.2 as well as <https://osf.io/sybrd/> for a detailed summary of acoustic parameters, some examples of the sound files and a rating study validating the stimuli.

5.2.3. Design

Experimental setup and EEG recording After providing informed consent and completing a short audio test (Cotral-Labor-GmbH, 2013), participants were prepared for the EEG-recording and subsequently started the emotion classification experiment using E-Prime 3.0 (Psychology Software Tools, Inc., 2016). The EEG was recorded using a 64-channel BioSemi Active II system (BioSemi, Amsterdam, Netherlands) with electrodes being attached with a cap on the 10–20 system (for EEG channel locations refer to Figure B.1). This system works with a “zero-ref” setup with a common mode sense/driven right leg circuit instead of ground and reference electrodes (for further information, see <https://www.biosemi.com/faq/cms&drl.htm>). The horizontal electrooculogram (EOG) was recorded from two electrodes at the outer canthi of both eyes, and the vertical EOG was monitored with a pair of electrodes attached above and below the right eye. All signals were recorded with direct current (120 Hz low-pass filter) and sampled at a rate of 512 Hz. During the EEG recording, participants were seated in a dimly lit, electrically shielded and sound-attenuated cabin (400-A-CT-Special, Industrial Acoustics™, Niederkrüchten, Germany) with their heads on a chin rest to ensure a constant distance of 90 cm to the computer screen. The sound stimuli were presented via in-ear headphones (Bose®MIE2 mobile headset).

Experimental task The participants’ task was to classify the stimulus emotion as happiness, pleasure, fear or sadness. There was no neutral response option to avoid that participants would choose neutral whenever they were unsure about their response. Assignment of response keys and response hands to emotion categories was counterbalanced across participants, using four different key mappings (see Table A.3).

Each trial started with a white fixation cross centered on a black screen. After $1000 \pm 100 \text{ ms}$, the cross changed into green and a vocal stimulus started playing. Behavioral responses were recorded from voice onset until 3000 ms after voice offset. As soon as a response was given, the fixation cross changed to gray, signaling the logging of the response. The cross remained on screen until the end of the response window. In case of no response (omission error), the final

trial slide (500 ms) was a feedback screen prompting participants to respond faster; otherwise, the screen turned back. Then the next trial started.

Because emotion judgments are subjective, judgment accuracy may not be ideal to gauge a participant's conscientiousness. Therefore, we added a second task on 10% of the trials. Here, participants were prompted to identify the last pseudoword by pressing one of four response options (/molen/, /namil/, /loman/ and /belam/). Please note that we added the /namil/-response option to have a label for each of the four keys on screen. In fact, we only used three different pseudowords, so /namil/ was never the correct response. A participant's data entered data analysis only if word identification accuracy was 80% or more. The experiment started with 10 practice trials presenting stimuli not used for the actual task. Subsequently, all 312 experimental stimuli were presented once in random order and then again in a different random order, resulting in 624 trials. Individual self-paced breaks were encouraged between blocks of 78 trials. The total duration of the experiment was about 50–60 min. After the experiment, participants completed a set of questionnaires that entered an exploratory analysis (details in Tables A.4 and A.5).

5.2.4. Data processing and analysis

Trials with omitted or preemptive responses (<200 ms) were excluded from the analysis of behavioral data. Mean accuracy and confusion data were analyzed using R version 4.0.2 (R Core Team, 2020). All trials entered EEG data analysis, which was done using EEGLAB (Delorme & Makeig, 2004) in Matlab R2020a (MATLAB, 2020). Raw EEG recordings were downsampled to 250 Hz and re-referenced to the average reference. Then the data were low-pass filtered at 30 Hz, high-pass filtered at 0.1 Hz (both filters -6 dB/octave, zero-phase shift) and epoched using a time interval of -200 to 1000 ms relative to voice onset. Epochs were then visually scanned for noisy channels and other unsystematic artifacts, such as drifts or muscle movements. The cleaned data were 1 Hz high-pass filtered and subjected to an independent component analysis. The resulting component structure was applied to the preprocessed data with the 30 to 0.1 Hz filter settings. Components reflecting typical artifacts (e.g. eye movements, eye blinks or ECG activity) were removed before back-projecting information from component space into EEG channel space. Next, the data were baseline corrected with a window of -200 to 0 ms relative to stimulus onset, and channels that had been removed earlier due to noise were interpolated using a spherical spline procedure (one channel in two participants and two channels in two participants). The resulting data were again scanned visually and residual artifacts and epochs were removed. Remaining epochs were submitted to a current source density (CSD) transformation using the CSD toolbox in EEGLAB (Kayser, 2009; Kayser & Tenke, 2006). This transformation returns essentially reference-free data which optimize the segregation of spatially overlapping sources (Kayser & Tenke, 2015). An analysis with the original average-referenced data replicates the results reported here and can be found in the aforementioned OSF repository. ERPs were derived by averaging epochs for each condition and participant. A minimum of 40 trials and an average of 47.48 trials per condition (out of a possible maximum of 48) and participant entered statistical analysis.

In order to assess the effects of F0 and timbre on the ERPs, we calculated difference waves by subtracting from the Full condition either F0 or timbre conditions, for each emotion separately. This resulted in two difference waves per emotion ($\text{Diff}_{\text{Full-F0}}$ and $\text{Diff}_{\text{Full-Timbre}}$) and was done to enable a more meaningful visual examination of the data and of how the removal of only one parameter affected the ERP when compared with the full condition. Please note that a comparison between $\text{Diff}_{\text{Full-F0}}$ and $\text{Diff}_{\text{Full-Timbre}}$ is mathematically equivalent to a simple comparison of F0 and timbre conditions. To explore the divergence between $\text{Diff}_{\text{Full-F0}}$ and $\text{Diff}_{\text{Full-Timbre}}$ for both topography and time course of ERP deflections, we performed a cluster-based permutation test on all 64 electrodes using the FieldTrip toolbox (Maris & Oostenveld, 2007; Oostenveld et al., 2011). The latency range was set from 0 to 500 ms, which offsets before the participants' mean behavioral response ($M_{\text{RT}} = 1489\text{ms}$, with 99% of trials between 697 and 2911ms). The analysis was done separately for each emotion using the Monte Carlo method with 1000 permutations and minimum cluster size of two channels. Based on the obtained cluster results, we then selected a frontocentral region of interest (ROI) including nine channels [F1, Fz, F2, FC1, FCz, FC2, C1, Cz, and C2] in latency ranges of the P200 [150, 250] and an N400-like negativity [300, 400] for further visualization and exploration. The behavioral and preprocessed EEG data together with respective analysis scripts are accessible on <https://osf.io/sybrd/>.

Note that averages included trials with both correct and incorrect emotion identifications, while previous studies used correct trials only (Schirmer et al., 2013). In the current dataset, the rate of misclassifications was fairly high, and an exclusion of these trials would have resulted in a substantial reduction of signal-to-noise ratio and statistical power. However, to ensure that our results were not biased by the inclusion of incorrect trials, we also repeated analyses based on correct trials only. The results replicated the pattern based on all trials, except that the difference in the N400-like negativity was slightly reduced for fearful stimuli. For a detailed report of effects sizes in different subsets of trials, please refer to Figure B.2.

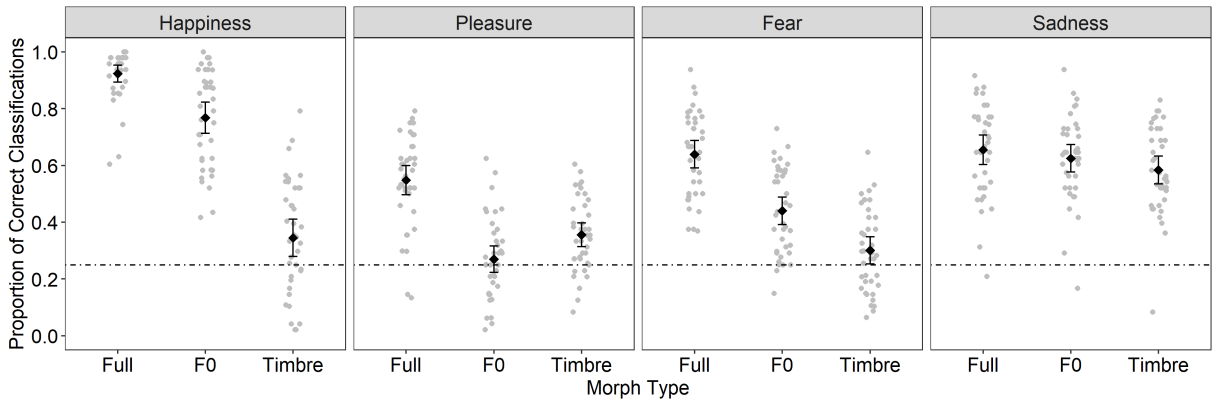
5.3. Results

5.3.1. Behavioral data – proportion of correct classifications

The mean proportion of correct responses was averaged separately for Emotion, Morph Type and participants. As there was no response option for “neutral”, neutral stimuli were excluded from analysis. An initial 4×3 analysis of variance with the within-subject factors Emotion and Morph Type revealed main effects of **Emotion**; $F(3, 114) = 45.42, p < .001, \omega^2 = 0.53 [0.40, 0.62]$, $\epsilon_{\text{HF}} = 0.983$; and **Morph Type**; $F(2, 76) = 295.67, p < .001, \omega^2 = 0.88 [0.83, 0.91]$, $\epsilon_{\text{HF}} = 0.896$; which were further qualified by an interaction; $F(6, 228) = 57.80, p < .001, \omega^2 = 0.59 [0.52, 0.64]$, $\epsilon_{\text{HF}} = 0.753$ (Figure 5.2). Post hoc comparisons of the different Morph Types for each Emotion revealed the following pattern: For all emotions, performance in the Full condition was better than in the F0 and timbre conditions, $|ts(38)| \geq 4.41, ps \leq .001$, Cohen's $d > 0.72 [0.36, 1.07]$. The only exception was the F0-sadness condition which differed from the Full-sadness condition only marginally, $t(38) = 1.88, p = .067, d = 0.31 [-0.02, 0.63]$. Importantly, the relative contributions of F0 and timbre differed. Specifically, comparing F0 vs timbre revealed a larger impact of

F0 on recognizing happiness, $t(38) = 10.48, p < .001, d = 1.70$ [1.20, 2.19]; fear, $t(38) = 5.98, p < .001, d = 0.97$ [0.58, 1.35]; and sadness, $t(38) = 2.06, p = .046, d = 0.33$ [0.01, 0.66]. In contrast, a larger impact of timbre was seen for pleasure, $t(38) = -3.28, p = .002, d = -0.53$ [-0.19, -0.87]. In addition to the proportion of correct responses, we calculated confusion data for each Emotion per Morph Type, this time including the neutral stimuli. The response matrices are displayed in Figure 5.3. The full statistical analysis is provided on <https://osf.io/sybrd/>.

Figure 5.2.: Mean proportion of correct responses per Emotion and Morph Type



Note. Whiskers represent 95%-CIs. Gray dots represent individual participants' data. The dotted line represents guessing rate at 0.25.

Figure 5.3.: Confusion matrices for each Emotion separately for the three Morph Types

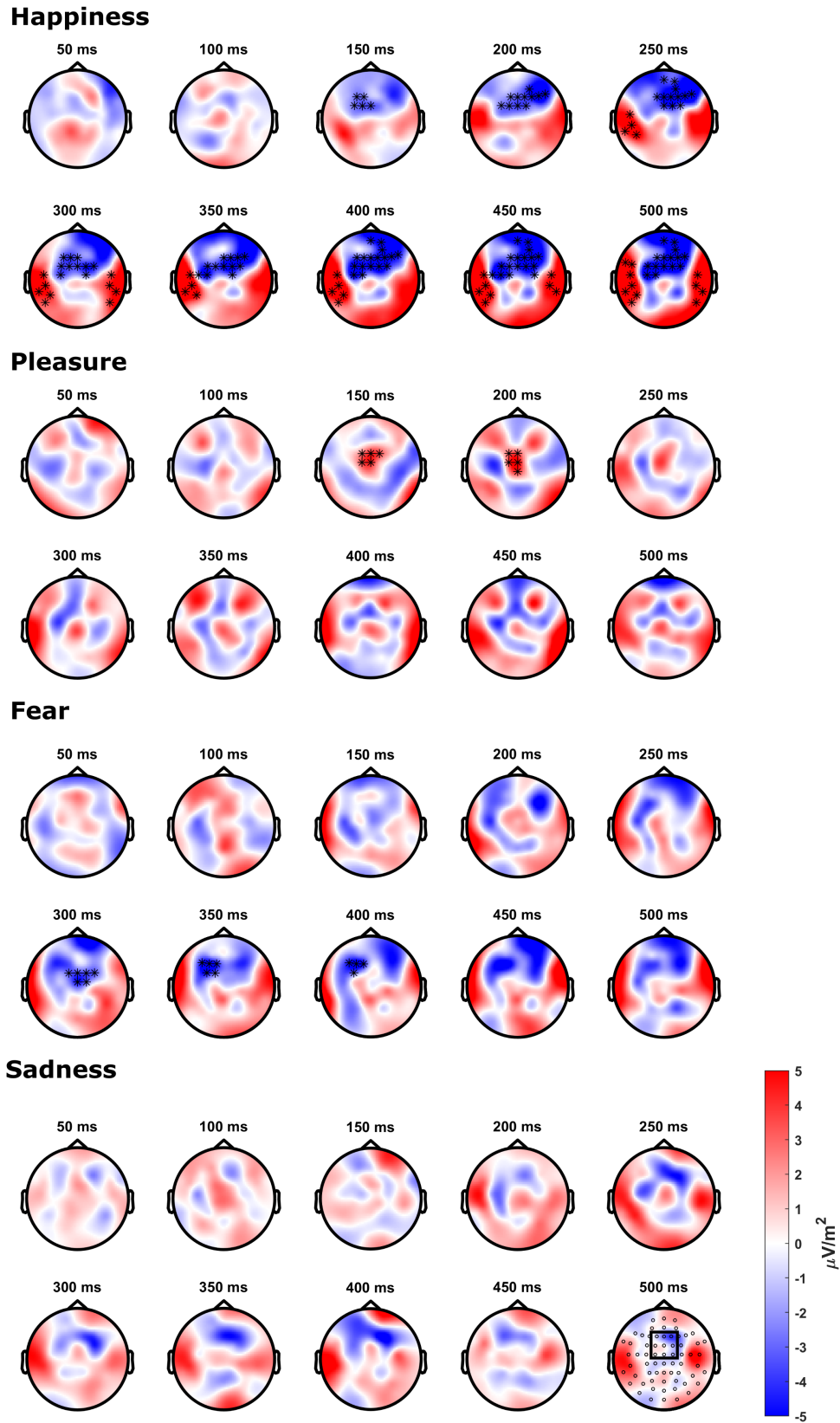
Classification Proportion in %	Full					F0					Timbre					
	Sad	2	15	21	66	56	Sad	9	33	37	63	Sad	27	34	43	58
	Fea	3	8	64	20	11	Fea	8	10	44	15	Fea	23	16	30	17
	Ple	3	55	4	9	14	Ple	6	27	3	9	Ple	15	36	16	17
	Hap	92	23	11	5	19	Hap	77	31	16	13	Hap	34	15	11	8
		Hap	Ple	Fea	Sad	Neu		Hap	Ple	Fea	Sad		Hap	Ple	Fea	Sad
Emotion																

Note. Numbers represent the proportion of classification responses per Emotion and Morph Type. Hap=happiness, Ple=pleasure, Fea=fear, Sad=sadness, Neu=neutral.

5.3.2. ERP data

Nonparametric cluster-based permutation test Cluster-based permutation tests were run to compare the Full minus F0 and Full minus Tbr difference waves separately for each emotion in a time window from 0 to 500 ms. The results are visualized in Figure 5.4. For happiness, the cluster-based permutation test revealed a significant difference between the F0 and the Timbre condition

Figure 5.4.: Scalp topographies of the contrast between the difference waves $\text{Diff}_{\text{Full-F}_0}$ and $\text{Diff}_{\text{Full-Timbre}}$ for each emotion separately from 50 to 500 ms



Note. Clusters of significant differences are indicated by the black asterisks. The black rectangle in the bottom right scalp shows the electrodes included into the ROI-based analysis. Color scheme developed by Adam Auton (2021).

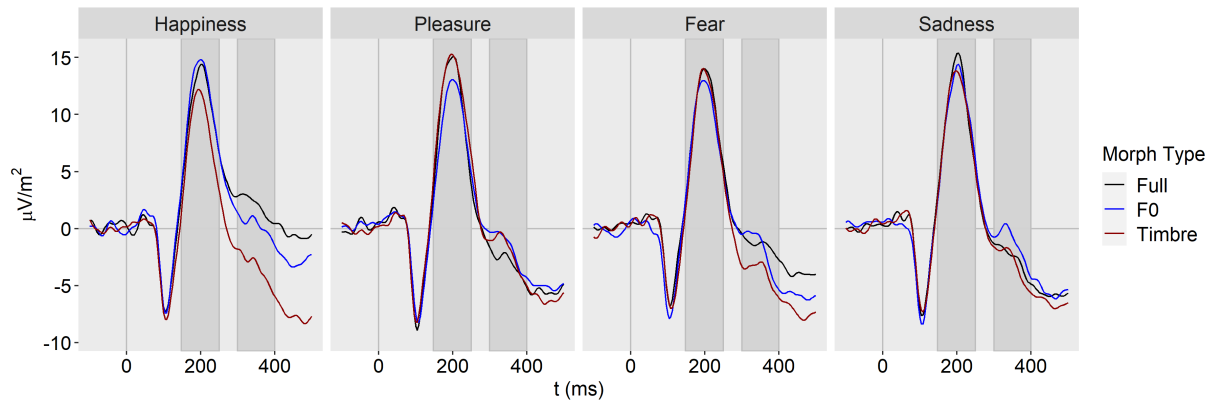
($p < .05$), in a pronounced frontocentral cluster between 130 ms and the end of the analyzed time range at 500 ms. Additionally, two bilateral temporal clusters appeared at an onset latency of around 230 ms. For pleasure, a frontocentral cluster was observed in the time range of 150–200 ms and for fear in a later time range of 300–400 ms, which seemed lateralized to the left. For sadness, no clusters of significant differences were observed. Please note that the spatial-temporal pattern of these clusters has to be interpreted with caution, since cluster-based permutation tests do not allow a definite conclusion about where an effect begins and ends in space and time, but only indicate that there is a difference within a given spatiotemporal window (Maris & Oostenveld, 2007).

Analysis of the frontocentral ROI To explore the frontocentral cluster in more detail, ERP-data were averaged across an ROI of nine channels [F1, Fz, F2, FC1, FCz, FC2, C1, Cz, and C2] (Figure 5.5). The difference between F0 and timbre was quantified by comparing mean amplitudes in the time windows of the P200 [150, 250] and the N400-like negativity [300, 400]. To compare the contrasts across emotions, we quantified them in terms of effect sizes (Cohen’s d). Since the ROI was preselected based on significant clusters, we refrained from further null-hypothesis significance testing.

P200. The contrast of F0 vs timbre revealed a strong effect for happiness, $d = 0.70$ [0.34, 1.05] and an effect in the opposite direction for pleasure, $d = -0.53$ [−0.19, −0.86], whereas effects for fear and sadness were negligibly small, with $d = -0.05$ [−0.36, 0.27] and $d = 0.03$ [−0.28, 0.34], respectively.

N400-like negativity. In the time window of the N400-like negativity, the strong effect in happiness persisted, $d = 0.87$ [0.50, 1.24], while the effect in pleasure ceased, $d = 0.10$ [−0.21, 0.41]. For fear and sadness, medium effects were observed, with $d = 0.41$ [0.08, 0.73] and $d = 0.38$ [0.06, 0.70], respectively.

Figure 5.5.: ERPs separately for Emotion and Morph Type, averaged across nine channels



Note. Averages are collapsed across [F1, Fz, F2, FC1, FCz, FC2, C1, Cz, and C2]. Gray shaded areas illustrate the time window of the P200 [150, 250] and the N400-like negativity [300, 400].

Parameter-effects on ERP amplitude predict parameter effects on behavior. To model the relationship between behavior and ERPs, we calculated performance and amplitude differences between F0 and timbre for corresponding stimuli and averaged them across the two stimulus presentations.

A cumulative link mixed model (calculated with the “ordinal” Package in R, R. H. B. Christensen, 2015) with the syntax

$$\begin{aligned} \text{Accuracy}_{\text{F0-Timbre}} \sim & \text{Emotion} + P200_{\text{F0-Timbre}} + N400_{\text{F0-Timbre}} + \\ & (\text{Emotion} + P200_{\text{F0-Timbre}} + N400_{\text{F0-Timbre}} | \text{Participant}) \end{aligned} \quad (5.1)$$

revealed that parameter differences in the amplitude of the N400-like negativity predicted the relative predominance of F0 over timbre in emotion recognition ($\beta = 0.004 \pm 0.002, z = 2.042, p = .041$). Thus, the bigger the F0 vs timbre amplitude difference in the N400-like negativity, the bigger was the performance difference between F0 and timbre. In additional exploratory analyses, we split the N400-like negativity into an early [300–350] and later [350–400] interval and observed that the predictive power was driven by the later interval ($\beta = 0.004 \pm 0.002, z = 2.336, p = .019$), but not the early one ($\beta = 0.002 \pm 0.002, z = 1.277, p = .202$). The P200-effect was non-significant (P200: $\beta = -0.002 \pm 0.002, z = -1.004, p = .315$).

5.4. Discussion

This study explored the relative contributions of timbre and F0 to the perception of vocal emotions and pursued the temporal course underpinning emerging vocal representations. Task performance and the ERPs underlined the importance of both parameters, while revealing their unique processing contributions as a function of emotion. The following paragraphs outline these contributions and present a discussion of how they inform extant models of vocal emotion perception.

5.4.1. The unique contribution of F0 and timbre in vocal emotion processing

While much research has pursued the functional significance of F0, considerably less attention has been directed to timbre (Banse & Scherer, 1996; Juslin & Laukka, 2003). Yet, based on the recurring finding that F0 correlates with perceived arousal (Brück et al., 2011), timbre was suggested to signal valence. This view was supported by machine-based classification approaches and behavioral data from nonverbal vocalizations (Anikin, 2020; Tursunov et al., 2019). The present data disagree with this perspective. A functional link between F0 and arousal should have led to more confusions across valence in the F0-only condition. In other words, participants should have mixed up high arousal emotions with other high arousal emotions (i.e. happiness and fear) and low arousal emotions with other low arousal emotions (i.e. pleasure and sadness; refer to the rating data in Appendix C). Likewise, a functional link between timbre and valence should have led to more confusions across arousal in the timbre-only conditions. Mix-ups should have happened primarily within rather than across positive (i.e. happiness and pleasure) and

negative emotions (i.e. fear and sadness). Neither pattern was observed in the present confusion data (Figure 5.3). Instead, all emotions tended to be confused most often with sadness.

Other proposals exist that better match the available evidence. For example, Gobl (2003) speculated that F0 expresses stronger emotions, while timbre may more effectively signal milder affective states. While the present data cannot directly speak to this, they accommodate such functionality. F0 effects were most pronounced for happiness and fear, which were rated high in intensity (for details, refer to the rating data in Appendix C). For emotions of lower intensity, such as sadness and pleasure, F0 effects were either reduced or absent. Similarly, Grichkovtsova et al. (2012) found prosody contour (including F0) to be more important for the recognition of happiness, whereas timbre seemed more important for sadness. Although our findings slightly diverge, they align with the fact that timbre seemed more relevant for weaker emotions.

Nevertheless, we reason that a framework linking F0 and timbre to rigid functional meanings is overly simplistic. Such a framework fails to account for the variability and flexibility in producing and perceiving vocal emotions. Very different styles of emotional expression can result in comparable recognition performance (Gobl, 2003; Spackman et al., 2009), underlining the perceivers' ability to adjust reliance on different vocal parameters when extracting emotional meaning. Another important aspect is the potential interaction of vocal parameters. Timbre and F0 naturally co-vary (Arias et al., 2021). Thus, when studied in isolation, one does not only eliminate the impact of the controlled vocal parameter but also their joint contribution. On the one hand, this would be particularly detrimental if important changes in one vocal parameter depend on coherent changes in the other (Grichkovtsova et al., 2012). On the other hand, one could also assume that one parameter is particularly important for emotional signaling while the other is naturally less informative. If so, the importance of timbre in the present pleasure stimuli could be partly due to the natural lack of information in F0 contour (Anikin, 2020).

5.4.2. Electrophysiological correlates of F0 vs timbre processing

How are vocal parameters analyzed and integrated in the brain to extract the emotional salience of voices? Although much debated, this process is still poorly understood (Paulmann & Kotz, 2018). We sought to shed light on this question by explicitly comparing the divergence of the two parameter-specific conditions from the Full emotion condition to study the relative importance of F0 vs timbre. We found that happy voices elicited a smaller P200 amplitude in the timbre relative to the F0 condition, whereas vocal pleasure elicited an opposite effect, in line with the observed performance data. For the N400-like negativity, parameter-specific effects were observed for happiness, sadness and fear, with larger amplitudes for timbre relative to F0, again in line with the behavioral results. Of importance is that the N400 amplitude difference between timbre and F0 positively predicted the associated performance difference in the behavioral data.

These findings add to our understanding of the functional significance of the P200 and the N400. With emotional quality and acoustic cues being confounded in natural stimuli, it has been

difficult to ascertain whether these components reflect emotional processing or are subject merely to basic acoustic influences (Paulmann et al., 2013; Schirmer & Gunter, 2017). In the present study, we employed stimuli with controlled acoustics and the intriguing resemblance between the behavioral and ERP results implies that emotional rather than acoustic processes shaped the P200 and the N400. Together, these findings agree with conclusions drawn from acoustically uncontrolled studies (Paulmann & Kotz, 2008; Schirmer et al., 2013) and corroborate existing models of vocal emotional processing (Frühholz et al., 2016; Schirmer & Kotz, 2006). Moreover, the finding that amplitude differences in the N400 (but not the P200) predicted overt emotion identification suggests that this process was fairly independent from early automatic responses and shaped instead by later more controlled processes such as conceptual processing of emotional meaning (Paulmann & Kotz, 2018). Note that for this study we adopted an exploratory approach and identified components based strictly on their timing and polarity. Moreover, regarding the N400, we wish to clarify that although this component was originally described in the context of lexical integration and semantic incongruity (Kutas & Hillyard, 1980), it has since been pursued more broadly including, for example, in the context of perceptual and semantic picture priming (Barrett & Rugg, 1989; Barrett et al., 1988), face processing (Wiese et al., 2017) and emotional processing (Paulmann & Pell, 2010). Thus, somewhat different N400 components, varying with regard to timing and scalp topography, have been documented and linked to a range of processes. For a more detailed discussion of this, please refer to Kutas and Federmeier (2011).

The observed ERP modulations suggest an emotion-specific time course in the neural processing of voices, with an earlier onset of emergent representations for happiness and pleasure when compared with sadness and fear. Similar effects have been reported for static faces (Schindler & Bublatzky, 2020). However, in contrast to static faces, the acoustics in voices evolve over time and may unfold their emotional information simply as a function of when and how a given cue becomes available. Thus, to what extent the latency differences we observed in this present study reflect relative differences in the ease or accessibility of positive and negative emotions or are tied strictly to acoustic stimulus constraints awaits further research.

5.4.3. Directions for future research

The present study presents a novel approach to the long-standing question of how the brain represents a speaker's emotional state. While it offers important new insights, it also generates a number of important questions. One such question concerns potential considerations associated with voice morphing. Although this technique results in stimulus materials of high quality, it also inevitably leads to parameter combinations that are unlikely to occur in natural voices, potentially making morphed stimuli sound less natural or human-like (Grichkovtsova et al., 2012; Skuk et al., 2015). Note that this concern is not specific to parameter-specific voice morphing but is equally prevalent in experiments using parameter-specific facial morphs (Sormaz et al., 2016). The extent to which both facial and vocal naturalness can be perceived and might influence emotion processing deserves further research. Another question concerns whether and how a listeners' goals might shape parameter-specific processes. For example, it would be interesting to

investigate under which circumstances the present effects replicate. Would they be still observable if participants were not instructed to explicitly identify the emotions? Based on the present findings, one would expect the N400 to be more malleable to task effects than the P200. Finally, an interesting direction for future research would be to pursue individual differences. For example, Schneider, Sluming, Roberts, Scherg et al. (2005) distinguished “fundamental pitch listeners” and “spectral listeners” with profound structural and functional differences in the auditory cortex. Likewise, there may be “F0 listeners” and “timbre listeners” who rely to different degrees on these parameters in vocal emotions.

5.5. Summary and conclusion

The present study demonstrated that the relative contributions of timbre and F0 to vocal emotion processing vary as a function of emotional category, with F0 being more important for happy, fearful and sad expressions and timbre being more important for pleasure. Furthermore, the relative importance of vocal cues for behavioral performance was mirrored in the ERPs at time points overlapping with the P200 and the N400. Indeed, N400 effects significantly predicted overt judgments delineating an important link between parameter-specific neural and behavioral processes. Thus, future research may leverage on parameter-specific voice morphing as a useful tool when studying how the human brain translates voice acoustics into emotional meaning.

6. Musicality - tuned to the melody of vocal emotions

Submitted as:

Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2023). Tuned to the Melody of Vocal Emotions [under review]

Abstract

Musicians outperform non-musicians in vocal emotion perception, likely because of an increased sensitivity to acoustic cues, such as fundamental frequency (F0) and timbre. Yet, how musicians make use of these acoustic cues to perceive emotions, and in what way such usage might differ from that in non-musicians, remains uncertain. To address these points, we created vocal stimuli that conveyed happiness, fear, pleasure, or sadness, either in all acoustic cues, or selectively in either F0 or timbre only. We then compared vocal emotion perception performance between two groups of professional/semi-professional musicians ($N = 39$) and non-musicians ($N = 38$), all socialized in Western music culture. Compared to non-musicians, musicians classified vocal emotions more accurately. This advantage was seen in the full and F0-modulated conditions but was absent in the timbre-modulated condition. Accordingly, musicians excel at perceiving the melody (F0), but not the timbre of vocal emotions. Further, F0 seemed more important than timbre for the recognition of all emotional categories. Additional exploratory analyses revealed a link between dynamic F0 perception in music and voices that was independent of musical training. Together, these findings suggest that musicians are particularly tuned to the melody of vocal emotions, and that this may in part be due to a natural predisposition to exploit melodic patterns.

6.1. Introduction

High levels of musicality are linked to advantages in non-musical domains such as speech perception and overall cognitive functioning (Elmer et al., 2018; Schellenberg, 2001, 2016). However, while several decades of systematic research have established robust evidence for transfer effects from musical abilities to relatively distant domains such as language skills (Elmer et al., 2018; Hallam, 2017), transfer effects to more closely related domains such as vocal emotion perception are less well established (M. Martins et al., 2021; Nussbaum & Schweinberger, 2021). Moreover, although accumulating evidence suggests a vocal emotion perception advantage in musicians

compared to non-musicians, the underlying mechanisms remain poorly understood. An important debate concerns the locus of transfer. While high-level supramodal processes such as emotional integration and decision making presumably play a role (Lima & Castro, 2011; Trimmer & Cuddy, 2008), the available evidence more consistently points to low-level acoustic sensitivity towards musical and vocal cues mediating the advantage in highly trained musicians (Correia et al., 2022). However, it remains unclear how musicians use different vocal cues to infer vocal emotions, and how this might differ from non-musicians. In the present study, we addressed this issue by investigating the degree to which musicians differ from non-musicians in their use of vocal cues that signal vocal emotion. To this end, we manipulated voices to constrain emotional information to specific acoustic cues, which we then presented in an emotion perception task. Thus, we examined how these cues, in isolation and in combination, inform vocal emotion perception in musicians and non-musicians.

6.1.1. What are the acoustic features of emotions and what is shared between music and voice?

Both music and voices convey emotions. In fact, emotional processing measures have identified remarkable overlap between these domains. Psychological overlap has been demonstrated by priming research, as emotional voice and music primes similarly modulate the semantic processing of subsequent positive and negative target words (Schirmer et al., 2002; Steinbeis & Koelsch, 2011). On a neural level, emotional processing of musical and vocal sounds recruits shared networks (Aubé et al., 2015; Escoffier et al., 2013; Frühholz et al., 2016). A reasonable explanation for these processing parallels highlights acoustic commonalities between musical and vocal emotions: In both domains, emotions are characterized by similar patterns of acoustic cues such as fundamental frequency (F0), amplitude, timing, or timbre (Juslin & Laukka, 2003; Scherer, 1995). F0 refers to a sound's lowest harmonic constituent, which we perceive as pitch. In both voices and music, time-varying pitch contour may be more simply described as melody. Timbre refers to a sound's quality independent of F0, timing and amplitude. It enables listeners to distinguish, for example, a trumpet from a violin, or one voice from another even when F0, tempo and loudness are identical. Amplitude and timing relate to the loudness and temporal unfolding of sounds, respectively. Importantly, research suggests that the manner in which acoustic cues combine in the context of emotions is shared between music and voice. Anger, for example, is often characterized by a high pitch, a rough timbre, a large amplitude, and fast speech rate, whereas the opposite holds for sadness (Banse & Scherer, 1996).

Research suggests that the different acoustic cues may play different roles in the perception of distinct emotions, albeit their exact roles remain contentious. In early emotional voice perception studies, F0 has been considered the perceptually dominant cue (Banse & Scherer, 1996; Juslin & Laukka, 2003). Recent work, however, suggests that timbre can play a central role in voice processing, and vocal emotion perception in particular (Nussbaum, Schirmer & Schweinberger, 2022; Nussbaum, von Eiff et al., 2022; Piazza et al., 2018; von Eiff et al., 2022). In fact, some data imply that both F0 and timbre carry unique information for different emotions (Anikin, 2020; Grichkovtsova et al., 2012). For example, Nussbaum, Schirmer and Schweinberger (2022)

found F0 to be perceptually dominant for the recognition of happiness, fear and sadness, whereas timbre seemed more important for the recognition of pleasure (as reported in Chapter 5). In music, pitch cues, timing and instrumentation have been highlighted as main tools for composers to convey emotional meaning (Juslin & Laukka, 2003; Schutz, 2017). However, the great variety of music styles, instrumentation-dependent acoustic possibilities or constraints, and performers' degrees of freedom make it hard to draw universal conclusions (Schutz, 2017).

6.1.2. How does musicality benefit vocal emotion perception?

Although methodological heterogeneity and limited test power are challenges to existing studies (M. Martins et al., 2021; Nussbaum & Schweinberger, 2021; Thompson et al., 2004), musicality appears to benefit vocal emotion perception (see Chapter 3). To explain this benefit, some authors evoked the concept of auditory sensitivity. When compared with non-musicians, musicians are better at perceiving the pitch, timbre and temporal aspects of musical sounds (Kraus & Chandrasekaran, 2010), and it has been argued that this extends to vocal sounds (Chartrand & Belin, 2006; Correia et al., 2022). Yet, exactly how acoustic processing differs between musicians and non-musicians remains elusive. One possibility is that compared to non-musicians, musicians use all acoustic cues more efficiently (e.g., faster, to a greater extent) leading to a general improvement of vocal emotion perception. Alternatively, musicality may affect the perception of individual acoustic cues and improve performance in a cue-specific way.

Some authors favor a cue-specific benefit and propose a special role of pitch contour (F0). For example, Globerson et al. (2013) identified time-varying pitch perception as a predictor for vocal emotion perception performance. Similarly, pitch is implicated by evidence from participants with amusia, a selective deficit for the processing of musical sounds, despite normal hearing and cognitive abilities (Ayotte et al., 2002; Stewart et al., 2006). In people with amusia, a consistent disadvantage for vocal emotion perception has been reported and linked to problems in pitch discrimination (Lima et al., 2016; Lolli et al., 2015; Pralus et al., 2019; Thompson et al., 2004). However, individuals with amusia represent the tail-end of the musicality spectrum. Their performance does not readily lend itself to inferences about what is special in highly trained musicians. Further, in most studies, amusia was defined based on pitch perception problems only, neglecting the potential influence of other vocal cues (Lagrois & Peretz, 2019). In general, a research focus on pitch may have precluded the potential role of other cues such as timbre, which has recently been shown to play a significant role in vocal emotional processing (Nussbaum, Schirmer & Schweinberger, 2022; Nussbaum, von Eiff et al., 2022). Indeed, research that linked response patterns to different acoustic cues suggest general rather than cue-specific differences between musicians and non-musicians (Lima & Castro, 2011). Thus, a systematic investigation of how musicians and non-musicians use different vocal cues for emotion perception is pending.

Besides auditory sensitivity, high-level supramodal processes have been raised as relevant for explaining performance differences between musicians and non-musicians (Trimmer & Cuddy, 2008). In that vein, musicality has been linked to skills like empathy, emotional differentiation, mind reading and decision making, all of which could foster emotional processing (Clark et al.,

2015; Lima & Castro, 2011; Trimmer & Cuddy, 2008). However, a benefit of musicality for emotional processing seems contained within the auditory modality, as it has not been observed for facial or lexical stimuli (Correia et al., 2022; Farmer et al., 2020; Twaite, 2016; Weijkamp & Sadakata, 2017). Further, a comparison of brain responses to vocal emotions between musicians and non-musicians suggests differences at early stages associated with acoustic analysis (Pinheiro et al., 2015; Rigoulot et al., 2015; Strait et al., 2009). Finally, Correia et al. (2022) found that the link between music training and vocal emotion perception was fully mediated by auditory perception skills. Taken together, these findings suggest that the link between musicality and vocal emotion perception is largely acoustic-bound.

6.1.3. Methodological challenges and aims of the present study

As mentioned above, some evidence is in line with the proposal that pitch sensitivity explains the superior performance of musicians in vocal emotion perception. However, the neglect of non-pitch cues such as timbre, as well as a reliance on individuals with amusia, makes this evidence inconclusive. Additionally, most of the reported evidence is purely correlational in nature, and therefore fails to establish a causal link between acoustic cues and emotion perception performance. This situation has recently contributed to an explicit call for more use of voice manipulation tools (Arias et al., 2021). The present study sought to tackle these issues by using the approach of parameter-specific voice morphing. This tool allows, among other things, a resynthesis of vocal stimuli such that they express emotional information through pitch contour or timbre cues only, while rendering the respective other cue uninformative (Kawahara & Skuk, 2018; Kawahara et al., 2008). Hence, this approach enables an experimental assessment of the relative importance of pitch (F0) and timbre for emotional judgements in musicians and non-musicians.

In the present study, we pursued two objectives: The first is a replication of the musicians' advantage for vocal emotional judgements, by recruiting a well-powered sample with (semi-) professional musicians and non-musicians. Second, we assessed how musicians and non-musicians differed in their use of acoustic cues to infer vocal emotions, focusing on the relative importance of F0 vs. timbre. Considering prior work, we expected that musicians would outperform non-musicians in a condition with full emotion modulation and when F0 only was informative of the respective emotion. Given the scarcity of data examining timbre, we were also interested in whether this acoustic cue would be equally or less affected by musicality.

6.2. Method

6.2.1. Participants

In line with previous research comparing vocal emotion perception between musicians and non-musicians (Lima & Castro, 2011), we aimed at a sample size of 40 participants per group. A power analysis using the R-package "Superpower" (Lakens & Caldwell, 2019) revealed that this sample size would allow the detection of a medium effect ($f = 0.25$) for an interaction between group (musicians, non-musicians) and the stimulus morphing condition (Full, F0, Timbre) with

0.8 power. Data collection took place from June 2021 to May 2022. All participants were fluent German speakers, aged between 18 and 50 years, and provided informed consent before completing the experiment. Data were collected pseudonymized. Participants were compensated with 25€ or with course credit. The experiment was in line with the ethical guidelines of the German Society of Psychology (DGPs) and approved by the local ethics committee of the Friedrich Schiller University Jena (Reg.-Nr. FSV 19/045).

Musicians We recorded data from 41 professional and semi-professional musicians. The data from two musicians had to be excluded because they omitted >5% trials in the emotion classification task. Thus, data from 39 musicians entered analysis (19 male, 20 female, aged 20 to 42 years [$M = 29.6$; $SD = 5.64$]). Mean onset age of musical training was 7 years ($SD = 2.53$, 4 – 17 years). Twenty-four participants were professional musicians with a music-related academic degree, all others had a non-academic music qualification (i.e. they worked as musicians or won a music competition; for more details see Table A.6). Thirty-five participants had studied their instrument for over 10 years, three between 6-9 years and one between 4-5 years.

Non-musicians Our recruitment criteria specified that non-musicians had not learned an instrument and did not engage in any musical activities like choir singing during childhood. We recorded data from 40 non-musicians, of which two exceeded the >5% omission criterion. Thus, we analyzed data from 38 non-musicians (18 male, 20 female, aged 19 to 48 years [$M = 30.5$; $SD = 6.54$]). Despite specifying inclusion/exclusion criteria during recruitment, 11 participants later reported having pursued learning an instrument or singing for a short period of time (two reported 2 and three reported 4-5 years of formal musical training; mean age at onset was 16 [$SD = 10.44$, range = 6 – 30 years]; for details see Table A.6). These participants were retained for data analysis.

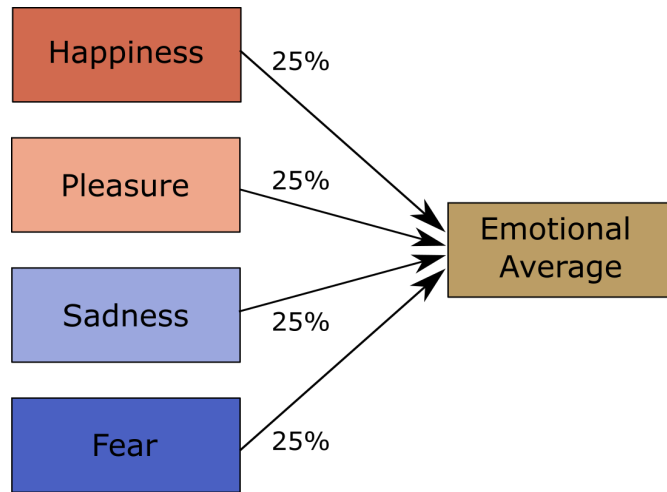
6.2.2. Stimuli

Original audio recordings We selected original audio recordings from a database of vocal actor portrayals provided by Sascha Frühholz, similar to the ones used in Frühholz et al. (2015). For the present study, we used three pseudowords (/molen/, /loman/, /belam/) uttered by eight speakers (four male, four female) with expressions of happiness, pleasure, fear, and sadness.

Voice averaging Using the Tandem-STRAIGHT software (Kawahara et al., 2008, 2013), we created emotional averages from the four emotions used in the study (see Figure 6.1) for each speaker and pseudoword. These averages, although not neutral, were uninformative and unbiased with respect to the four emotions of interest. We opted for average rather than neutral stimuli because a previous study showed that averages are more suitable for the subsequent generation of voice morphs ensuring that such morphs do not differ systematically in perceived naturalness (Nussbaum et al., 2023, in revision, reported in Chapter 4).

Parameter-specific voice morphing To synthesize parameter-specific emotional voice morphs, we created morphing trajectories between each emotion and the emotional average of the same

Figure 6.1.: Schematic depiction of the voice averaging process












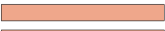
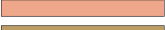





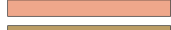












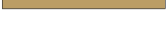


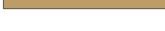


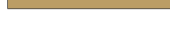
speaker and pseudoword. After manual mapping of time- and frequency anchors at key features of a given utterance pair (e.g., on- and offset of vowels), vocal samples on an emotion/average-continuum were synthesized via weighted interpolation of the originals; for a more detailed description see Kawahara and Skuk (2018). Crucially, Tandem-STRAIGHT allows independent interpolation of five different parameters: (1) F0-contour, (2) timing, (3) spectrum-level, (4) aperiodicity, and (5) spectral frequency; the latter three are summarized as timbre.

We created three types of morphed stimuli (see Figure 6.2). **Full-Morphs** were stimuli with all Tandem-STRAIGHT parameters taken from the emotional version (corresponding to 100% from the emotion and 0% from average), with the exception of the timing parameter, which was taken from the average (corresponding to 0% emotion and 100% average). **F0-Morphs** were stimuli with the F0-contour taken from the emotional version, but timbre and timing taken from the average. **Timbre-Morphs** were stimuli with all timbre parameters taken from the emotional version, but F0 and timing from the average. In addition, all average stimuli were included as a further ambiguous reference category. Note that the timing was kept constant across all conditions to allow a pure comparison of F0 vs. timbre. In total, this resulted in $8 \text{ (speakers)} \times 3 \text{ (pseudowords)} \times 4 \text{ (emotions)} \times 3 \text{ (morphing conditions)} + 24 \text{ average (8 speakers} \times 3 \text{ pseudowords)} = 312$ stimuli. For analysis purposes, we collapsed data across speakers and pseudowords. Using PRAAT (Boersma, 2018), we normalized all stimuli to a root-mean-square of 70 dB SPL (duration $M = 780$ ms, range 620 to 967 ms, $SD = 98$ ms). Please refer to <https://osf.io/5tczs/> for a summary of acoustic parameters and some examples of the sound files.

6.2.3. Design

The study consisted of two sessions: all participants first completed an online session outside the laboratory and were subsequently invited to an EEG session in the laboratory. Here, we only report the results of the online study. The results of the EEG session are reported in Chapter 7.

Figure 6.2.: Morphing matrix for stimuli with averaged voices as reference

Reference	Emotion		Full Morph	F0 Morph	Timbre Morph
Emotional Average	Happiness	F0 Timbre Timing	  	  	  
Emotional Average	Pleasure	F0 Timbre Timing	  	  	  
Emotional Average	Sadness	F0 Timbre Timing	  	  	  
Emotional Average	Fear	F0 Timbre Timing	  	  	  

Data were collected online via PsyToolkit (Stoet, 2010, 2017). Participants were required to use a computer with a physical keyboard and headphones, and were asked to ensure a quiet environment for the duration of the study. As browser, we recommended Google Chrome, and excluded Safari for technical reasons. In the beginning, participants entered demographic information, including age, sex, native language, profession, and potential hearing impairments such as tinnitus. Next, participants had the opportunity to adjust their sound settings to a comfortable sound pressure level.

Emotion classification experiment The participants' task was to classify vocal emotions as happiness, pleasure, fear, or sadness. Each trial started with a green fixation cross presented for 500 ms. Subsequently, a loudspeaker symbol appeared, and the sound was played. After voice offset, a response screen showed the emotion labels and participants could enter their response within a 5000 ms time window starting from voice offset. Participants responded with their left and right index and middle fingers. The mapping of response keys to emotion categories was randomly assigned for each participant, out of four possible key mappings. Emotions of the same valence were always assigned to the same hand and emotions with similar intensity (fear – happiness and sadness – pleasure) were always assigned to the corresponding fingers of both hands (details in Tables A.9 and A.10). In case of no response (omission error), the final trial slide (500 ms) provided a feedback prompting participants to respond faster; otherwise, the screen turned black. Then the next trial started.

At the beginning of the experiment, participants completed eight practice trials with stimuli not used during the actual task. Subsequently, all 312 experimental stimuli were presented once in randomized order across six blocks of 52 trials each. Between blocks, participants could take self-paced breaks. The total duration of the experiment was about 25 minutes.

Profile of Music Perception Skills (PROMS) To measure music perception skills beyond self-reports, we adopted the modular version of the Profile of Music Perception Skills (Law & Zentner, 2012; Zentner & Strauss, 2017). We selected the four subtests “Melody”, “Pitch”, “Timbre”, and “Rhythm”, which we considered most informative for the present research. For each subtest, participants completed 18 items, preceded by one practice trial. During each trial, participants heard a reference stimulus twice followed by a target stimulus. Then, they indicated whether reference and target were the same or different. Although this was a binary decision, the test employs a 5-point Likert scale with the labels “definitely same”, “maybe same”, “don’t know”, “maybe different”, and “definitely different”, which we also adopted here. Participants completed the test in about 20 minutes. One participant encountered technical problems in the “Melody” subtest, which was therefore repeated several months later to be included in data analysis.

Questionnaires After the PROMS, participants completed several questionnaires: the German Version of the Autism Quotient Questionnaire, AQ, (Baron-Cohen et al., 2001; Freitag et al., 2007), a 30-item Personality Inventory measuring the Big Five domains (Rammstedt et al., 2018), the Goldsmiths Musical Sophistication Index, Gold-MSI, (Müllensiefen et al., 2014) to assess the participants’ degree of self-reported musical skills, additional questions concerning music experience and musical engagement, their socioeconomic background, and the 20-item version of the Positive-Affect-Negative-Affect-Scale, PANAS (Breyer & Bluemke, 2016; Watson et al., 1988). Mean duration of the whole online experiment was about 75 minutes.

6.2.4. Data analysis

Data were analyzed using R Version 4.1.0 (R Core Team, 2020). Response omissions (~1%) were treated as errors and participants with more than 5% of such omissions excluded from data analysis. Analyses of Variance (ANOVAs) and correlational analyses were performed on data averaged across speaker and pseudoword. Post-hoc tests were Benjamini-Hochberg corrected where appropriate (Benjamini & Hochberg, 1995). Preprocessed data, analysis scripts and supplemental materials can be found in the associated OSF repository (<https://osf.io/5tczs/>).

Concerning the PROMS, we computed a measure that we thought reflected a combination of classification accuracy and certainty. We coded responses from 0 to 1 in 0.25 steps starting with the “definitely” correct option down to the “definitely” incorrect option (thus, “don’t know” was always coded with 0.5) and subtracted 0.5 from the final measure. Thus, a positive score indicates that participants were more correct/confident, whereas a negative score indicates more incorrect/uncertain ratings. We then averaged performance across trials for each subtest. Originally, the test authors recommend a d-prime measure which weighs hits and false alarms for response certainty. The results for such a d-prime measure converge with our own scoring reported here (see <https://osf.io/5tczs/>).

6.3. Results

6.3.1. Demographic, musicality, and personality characteristics of participants

Musicians and non-musicians did not differ in the socioeconomic status assessed via educational level, $X^2(2, N = 77) = 5.21, p = .074$, highest academic degree, $X^2(8, N = 77) = 6.40, p = .603$, and household income, $X^2(4, N = 77) = 5.66, p = .226$ (details in Table A.7). Further, they were comparable in age as well as positive and negative affect (see Table 6.1 for a summary of participant characteristics assessed via self-report and music performance in the PROMS). For the Big Five, slightly higher levels of openness and neuroticism were observed in musicians compared to non-musicians. With respect to autistic traits, musicians and non-musicians did not differ in their overall score. However, there were differences in the two subscales proposed by Hoekstra et al. (2008): Musicians scored higher than non-musicians on the Attention to Detail subscale, but lower on the Social Communication subscale. Splitting the Social Communication subscale into the four subscales originally proposed by Baron-Cohen et al. (2001), group differences were due to self-reported Social Skills and, although to a lesser degree, to Imagination rather than to Communication or Attention Switching. In the Gold-MSI, musicians scored considerably higher than non-musicians on all subfactors as well as the general musicality score. Further, musicians outperformed non-musicians in all four subtests of the PROMS.

6.3.2. Emotion classification performance

Proportion of correct classifications The mean proportion of correct responses was submitted to an ANOVA with Emotion (happiness, pleasure, fear, and sadness) and Morph Type (Full, F0, and Timbre) as repeated measures factors and Group (musicians, non-musicians) as a between subject factor. Reference stimuli (emotional averages) were excluded from this analysis. In addition to examining the proportion of correct responses, we also examined unbiased hit rates H_u as outcome measure, as proposed by Wagner (1993). As both approaches yielded identical results with only one exception (reported below), we report the simpler accuracy data here.

Our results included main effects of **Group** ($F(1, 75) = 5.937, p = .017, \omega^2 = .06 [0.00, 0.19]$), **Emotion** ($F(3, 225) = 74.18, p < .001, \omega^2 = .49 [0.40, 0.56]$) and **Morph Type** ($F(2, 150) = 905.25, p < .001, \omega^2 = .92 [0.90, 0.94], \epsilon_{HF} = .902$). These were qualified by an interaction of **Group x Morph Type** ($F(2, 150) = 6.10, p = .005, \omega^2 = .06 [0.00, 0.14], \epsilon_{HF} = .902$) as well as an interaction of **Emotion x Morph Type** ($F(6, 450) = 26.44, p < .001, \omega^2 = .25 [0.18, 0.31], \epsilon_{HF} = .904$). The three-way interaction did not reach significance ($F(6, 450) = 0.67, p = .663$).

Post-hoc tests revealed that musicians outperformed non-musicians in Full- and F0- morph conditions, whereas there was no difference in the Timbre-morph condition (Full: $|t(69.15)| = 3.35, p = .001, d = 0.81 [0.31, 1.29]$; F0: $|t(67.97)| = 2.31, p = .023, d = 0.56 [0.07, 1.04]$; Timbre: $|t(74.95)| = 0.30, p = .769, d = 0.07 [-0.38, 0.52]$, see Figure 6.3).

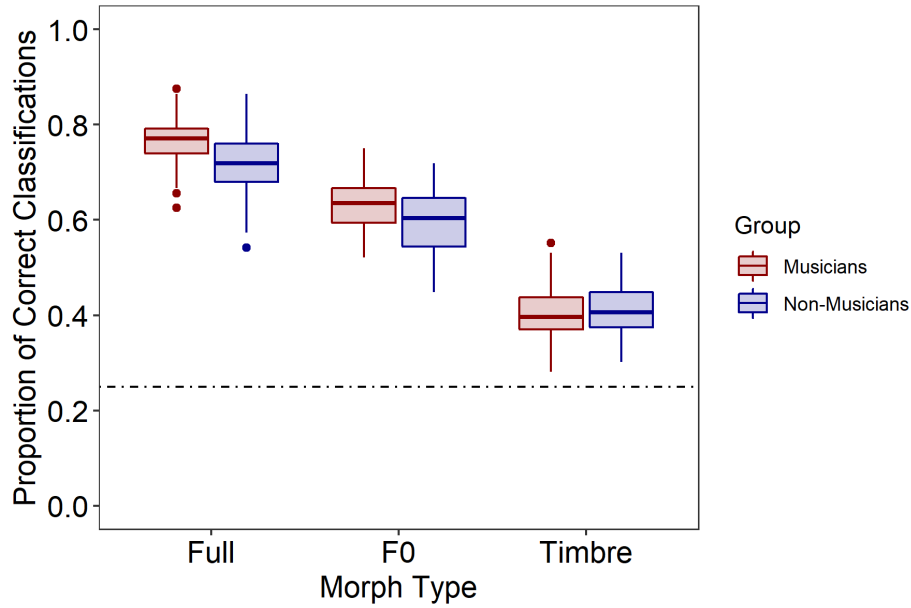
Table 6.1.: Characteristics of participants - Demography, personality, and musicality

	Musicians	Non-Musicians					
	M (SD)	M (SD)	t	df ^a	p	Cohens d	
Age	29.7 (5.60)	30.5 (6.5)	-0.63	72.82	.528	-0.15 [-0.61, 0.31]	
<i>PANAS</i>							
positive Affect	3.33 (0.66)	3.10 (0.67)	1.51	74.83	.136	0.35 [-0.11, 0.80]	
negative Affect	1.68 (0.47)	1.49 (0.69)	1.39	65.37	.170	0.34 [-0.15, 0.83]	
<i>Big Five</i>							
Openness	4.11 (0.50)	3.81 (0.80)	1.99	61.77	.050	0.51 [0.00, 1.01]	*
Conscientiousness	3.49 (0.72)	3.76 (0.72)	-1.63	74.96	.108	-0.38 [-0.83, 0.08]	
Extraversion	3.48 (0.66)	3.38 (0.79)	0.61	72.31	.543	0.14 [-0.32, 0.60]	
Agreeableness	3.91 (0.57)	3.75 (0.66)	1.20	72.93	.236	0.28 [-0.18, 0.74]	
Neuroticism	2.94 (0.66)	2.58 (0.82)	2.10	70.77	.039	0.50 [0.02, 0.97]	*
<i>AQ</i>							
Total	15.64 (5.03)	17.58 (6.41)	-1.47	70.15	.145	-0.35 [-0.82, 0.12]	
Attention to Detail	5.46 (2.05)	4.32 (2.01)	2.47	74.99	.016	0.57 [0.11, 1.03]	*
Social	10.18 (4.72)	13.26 (6.51)	2.38	67.38	.020	-0.58 [-1.06, -0.09]	*
Social Skills	1.44 (1.68)	2.61 (2.63)	-2.32	62.75	.024	-0.59 [-1.09, -0.08]	*
Communication	1.87 (1.63)	2.39 (1.73)	-1.37	74.39	.176	-0.32 [-0.77, 0.14]	
Imagination	2.13 (1.51)	2.87 (1.95)	-1.86	69.69	.067	-0.45 [-0.92, 0.03]	t
Attention Switching	4.74 (1.93)	5.39 (1.92)	-1.48	74.96	.142	-0.34 [-0.80, 0.11]	
<i>Gold-MSI</i>							
General ME	5.68 (0.50)	2.74 (1.07)	15.38	52.28	<.001	4.25 [3.27, 5.23]	***
Active Engagement	4.94 (0.82)	2.95 (1.19)	8.50	65.23	<.001	2.11 [1.50, 2.70]	***
Formal Education	5.94 (0.56)	1.71 (0.68)	29.79	71.75	<.001	7.03 [5.79, 8.27]	***
Emotion	5.88 (0.74)	4.95 (1.32)	3.79	57.60	<.001	1.00 [0.45, 1.54]	***
Singing	5.33 (0.84)	2.84 (1.26)	10.21	64.23	<.001	2.55 [1.89, 3.20]	***
Perception	6.33 (0.50)	4.22 (1.49)	8.25	45.16	<.001	2.45 [1.68, 3.22]	***
<i>PROMS</i>							
Pitch	0.27 (0.06)	0.18 (0.06)	6.23	74.97	<.001	1.44 [0.93, 1.94]	***
Melody	0.23 (0.10)	0.07 (0.08)	9.68	74.95	<.001	2.24 [1.65, 2.81]	***
Timbre	0.32 (0.08)	0.26 (0.09)	2.91	73.47	.004	0.68 [0.21, 1.15]	**
Rhythm	0.32 (0.08)	0.27 (0.08)	3.35	74.99	.001	0.77 [0.30, 1.24]	**

Note. Descriptive values show mean ratings for the PANAS (Breyer & Bluemke, 2016), the Big Five Domains (Rammstedt et al., 2018), and the Gold-MSI (Müllensiefen et al., 2014). AQ scores were calculated based on Hoekstra et al. (2008) and Baron-Cohen et al. (2001).

^a Note that original degrees of freedom were 75 but were corrected due to unequal variance.

Figure 6.3.: Boxplots depicting correct responses per Morph Type separately for musicians and non-musicians



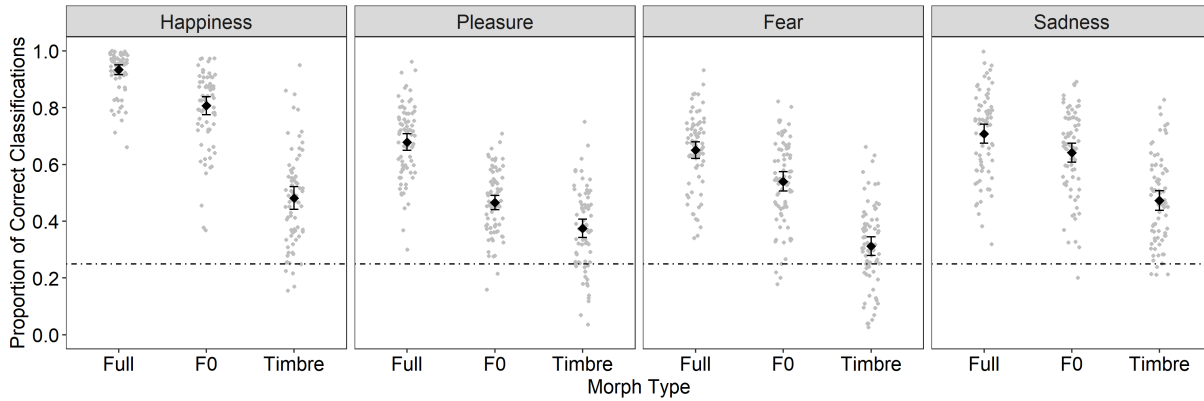
Note. The dotted line represents guessing rate at .25.

Follow-up analyses of the Morph Type effect revealed that performance was best in the Full condition, followed by the F0 and then the Timbre condition (Full vs. F0: $|t(76)| = 20.12$, $p < .001$, $d = 2.31$ [1.88, 2.74], F0 vs Timbre: $|t(76)| = 22.34$, $p < .001$, $d = 2.56$ [2.10, 3.03], Full vs Timbre: $|t(76)| = 38.50$, $p < .001$, $d = 4.43$ [3.69, 5.15]). This main effect of Morph Type was also found for all emotions separately (all $F_s(2, 152) > 116.05$, $p < .001$), although it differed slightly between emotions, as suggested by the interaction (see Figure 6.4, for all post-hoc tests, refer to <https://osf.io/5tczs/>).

To address our specific interest in the relative importance of F0 and Timbre for the different emotions, we calculated the performance difference_{F0-Tbr} for each emotion separately. Performance difference was largest for Happiness ($M = 0.33 \pm 0.02$ SEM), followed by Fear ($M = 0.23 \pm 0.02$), Sadness ($M = 0.17 \pm 0.02$), and Pleasure ($M = 0.09 \pm 0.02$; all pairwise comparisons $|ts(76)| \geq 2.79$, $p_{corrected} \leq .006$, $ds \geq 0.32$ [0.09, 0.55]). Using unbiased hit rates H_u , the performance difference between Sadness and Pleasure was comparable ($|t(76)| = 1.34$, $p = 0.184$, $d = 0.15$ [-0.07, 0.38]).

Classification of averaged stimuli and confusion data In addition to the proportion of correct responses, we calculated confusion data for each Emotion and Morph Type, this time including the averaged stimuli. The response matrices are displayed in Figure 6.5. A planned analysis of the averaged stimuli revealed that they were most often classified as expressing sadness, followed by pleasure, happiness and fear (sadness vs. pleasure: $|t(76)| = 3.56$, $p < .001$, $d = 0.41$ [0.17, 0.64]; pleasure vs. happiness: $|t(76)| = 3.40$, $p = .001$, $d = 0.39$ [0.16, 0.63]; happiness vs. fear:

Figure 6.4.: Mean proportion of correct classifications per Emotion and Morph Type



Note. Whiskers represent 95%-confidence intervals. Grey dots represent individual participants' data. The dotted line represents guessing rate at .25

$|t(76)| = 0.17, p = .867, d = 0.02 [-0.21, 0.24]$, p-values corrected). There was no significant effect of Group. Please refer to Figures B.3 and B.4 for a presentation of confusion data separated by Group.

Figure 6.5.: Confusion matrices for each Emotion for the three Morph Types

	Full					F0					Timbre				
	Hap	Ple	Fea	Sad	Avg	Hap	Ple	Fea	Sad		Hap	Ple	Fea	Sad	
Sad	2	13	19	71	38	6	26	27	64		19	27	33	47	
Fea	2	4	65	16	18	6	10	54	15		16	14	32	20	
Ple	2	68	6	9	27	7	47	8	15		16	38	19	21	
Hap	94	15	10	4	18	81	17	11	6		48	21	17	12	

Note. Numbers represent the proportion of classification responses per Emotion and Morph Type. Hap = happiness, Ple = pleasure, Fea = fear, Sad = sadness, Avg = average.

6.3.3. Links between musical skills and vocal emotion perception

In a subsequent exploratory analysis, we calculated Spearman correlations between vocal emotion perception performance and both the PROMS music perception performance and the Gold-MSI self-rated musicality. The results are shown in Tables 6.2 and 6.3, as well as Figures B.5 and B.6.

Correlations between the PROMS and vocal emotion perception Of particular interest, we obtained a strong correlation between the overall vocal emotion recognition performance and average PROMS performance. This correlation also emerged in a separate analysis of the control

Table 6.2.: Correlations between vocal emotion recognition and music perception performance

	PROMS _{Avg}	Pitch	Melody	Timbre	Rhythm
VER _{Avg}	.44 (<.001)	.22 (.090)	.39 (.002)	.22 (.084)	.35 (.005)
Full-Morphs	.47 (<.001)	.28 (.028)	.44 (<.001)	.29 (.023)	.27 (.028)
F0-Morphs	.35 (.005)	.10 (.434)	.32 (.011)	.15 (.278)	.35 (.005)
Timbre-Morphs	.13 (.322)	.11 (.424)	.08 (.523)	.07 (.542)	.13 (.322)

Note. VER = Vocal Emotion Recognition performance. *p*-values were adjusted for multiple comparisons using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995)

group ($r(36) = 0.48, p = .002$), but was non-significant in musicians ($r(37) = 0.22, p = .117$), possibly due to reduced variance. Performance in the Full-morph condition correlated with all subtests of the PROMS. Interestingly, there was also a more specific link between the F0 morph condition and the Melody subtest, suggesting that both tasks tap into similar abilities. There was no link between the timbre morph condition and the timbre subtest.

In the next step, we explored these above correlations in more detail to examine a potential role of musical training. Specifically, we calculated partial correlations to control for formal musical education (Kim, 2015). The correlations between VER_{Avg} and PROMS_{Avg} ($r(75) = 0.41, p < .001$), Full-Morphs and Melody ($r(75) = 0.35, p = .002$), Full Morphs and Timbre ($r(75) = 0.24, p = .036$), and F0-Morphs and Melody ($r(75) = 0.31, p = .006$) remained significant. Correlations of Full-morph performance with Pitch and Rhythm turned non-significant when controlling for formal musical education ($rs(75) \leq 0.22, ps \geq .055$).

Table 6.3.: Correlations between vocal emotion recognition and self-rated musicality

	General Musicality	Active Engagement	Formal Education	Emotion	Singing	Perception
VER _{Avg}	.30 (.035)	.21 (.147)	.21 (.147)	.02 (.865)	.31 (.035)	.28 (.041)
Full-Morphs	.40 (.004)	.28 (.041)	.28 (.041)	.10 (.555)	.41 (.004)	.34 (.021)
F0-Morphs	.21 (.147)	.11 (.555)	.12 (.508)	-.04 (.788)	.23 (.120)	.19 (.168)
Timbre-Morphs	.08 (.677)	.06 (.741)	.05 (.748)	-.02 (.865)	.06 (.748)	.09 (.621)

Note. VER = Vocal Emotion Recognition performance. *p*-values were adjusted for multiple comparisons using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995)

Correlations between the Gold-MSI and vocal emotion perception There was a correlation between vocal emotion perception performance and self-rated musicality, even when controlled for formal musical education ($r(75) = 0.28, p = .014$). Further, self-rated singing abilities were linked to increased sensitivity towards vocal emotions (controlled for formal musical education: $r(75) = 0.23, p = .041$).

Correlations between personality traits and vocal emotion perception To rule out that the performance difference between musicians and non-musicians could be attributed to one of the personality traits that differed between groups, we correlated them with averaged vocal emotion performance. The results were non-significant for openness ($r(75) = 0.13, p = .269$) but entailed a marginally positive association between neuroticism and vocal emotion perception ($r(75) = 0.22, p = .051$). None of the AQ scales correlated significantly with vocal emotion perception performance (all $p_s \geq .078$).

6.4. Discussion

In this study, we replicated earlier works showing that musicians outperform non-musicians in vocal emotion perception. Further, we investigated the role of different acoustic cues underpinning vocal emotion perception across listener groups and emotional categories. Our findings highlight the special role of pitch contour (F0), i.e. the melody of vocal emotions. On the one hand, musicians displayed a specific advantage for this cue. On the other hand, pitch contour seemed to be the perceptually dominant parameter across all emotional categories. In what follows, we will discuss these findings in more detail.

6.4.1. The musicality benefit for vocal emotion perception – a matter of auditory sensitivity?

Although a benefit of musicality for vocal emotion perception has been reported before (M. Martins et al., 2021; Nussbaum & Schweinberger, 2021), the present study offers an important contribution to this literature. This is because we considered in detail a number of methodological limitations to previous work, including a clear specification of “musicality”, appropriately powered sample sizes, and controls for confounding variables such as cognitive functioning (Lima & Castro, 2011; Thompson et al., 2004; Trimmer & Cuddy, 2008). In addressing these limitations, the present data offer original and strong evidence for transfer benefits from music to voice perception.

Most importantly, our study reveals novel insights into the role of acoustic cues underpinning these benefits. We found that musicians were specifically tuned to the melody of vocal emotions, in that they displayed a cue-specific advantage for pitch contour (F0), but not for timbre. While previous studies reported correlational links between pitch sensitivity and vocal emotion perception (Globerson et al., 2013; Lima & Castro, 2011), we present the first causal evidence that is based on voice stimuli which were directly acoustically manipulated (Arias et al., 2021).

In line with the general tenor in the literature, our findings suggest that the link between musicality and vocal emotion perception is mediated by low-level auditory sensitivity (Correia et al., 2022; Lima & Castro, 2011; Lolli et al., 2015; M. Martins et al., 2021) and pitch sensitivity in particular (Globerson et al., 2013). In fact, the link between auditory sensitivity in music and voice perception even holds in the absence of formal musical training and when correlations are controlled for formal musical education. These findings converge with data from Correia et al. (2022) who found that the association between music training and vocal emotion perception is

fully mediated by auditory and music perception skills. It also fits well with data from individuals with congenital amusia, whose pitch perception deficits predict emotion perception problems (Lima et al., 2016; Lolli et al., 2015; Thompson et al., 2012). Although ours and this latter work do not rule out potential music training effects (Fuller et al., 2018; Good et al., 2017; Thompson et al., 2004), they suggest that differences in auditory sensitivity might prepare some individuals to excel in and enjoy musical activities while also enhancing their vocal socio-affective skills.

Looking at the different subtests of the PROMS allowed us to assess the relevance of specific musical skills for vocal emotional processing in more detail. Both “Pitch” and “Melody” subtests target pitch perception, but “Pitch” measures pitch discrimination of two static tones, whereas “Melody” requires the tracking of changes in pitch contour over time (Law & Zentner, 2012; Zentner & Strauss, 2017). Similar to the PROMS “Pitch” task, the PROMS “Timbre” task measures the ability to discriminate the timbre of two static tones. However, the PROMS “Melody” task lacks a timbre equivalent, requiring the tracking of dynamic timbre cues. The “Rhythm” subtest, by contrast, again requires sensitivity to how acoustic events evolve over time.

In our data, vocal emotion perception was consistently linked to performance in tests that examined dynamic rather than static acoustic processing. Specifically, both the “Melody” and “Rhythm” subtests but not the “Pitch” and “Timbre” subtests correlated significantly with overall vocal emotion perception. Thus, for predicting emotion recognition success, tracking acoustic changes over time seems more relevant than representing temporally isolated acoustic features (Juslin & Laukka, 2003). This seems intuitive, as vocal cues are also dynamically evolving over time. Accordingly, we found that the vocal F0 condition correlated with “Melody”, as these tasks share similar demands on the perceptual system, but not with “Pitch”. Similarly, Globerson et al. (2013) found that vocal prosody recognition could be predicted by the ability to detect dynamic pitch changes, but not by static pitch discrimination. For timbre, the static music task failed to correlate with the vocal timbre condition. Maybe with a music test requiring tracking of timbre features over time a link to vocal timbre perception would become apparent.

6.4.2. Emotional communication in music and voice – same code, same task?

It has been long established that emotions have similar acoustic signatures in voices and music (Juslin & Laukka, 2003). Further, they are perceived in similar ways (Schirmer et al., 2002; Steinbeis & Koelsch, 2011), and processed by shared networks (Escoffier et al., 2013; Frühholz et al., 2016; Peretz et al., 2015). The current investigation further strengthens the notion that auditory sensitivity in both domains is linked in listeners. Can we therefore conclude that emotions share the same characteristics and functions in these domains? In traditional models of nonverbal behavior, emotional prosody has been understood in the context of a sender-receiver perspective, where an emotional message is coded into a signal and the signal is sent with the, perhaps implicit, intent/expectation of being decoded by the receiver (Bänziger et al., 2015; Shariff & Tracy, 2011). Yet, more recently, nonverbal behaviors have been conceptualized more broadly. Accordingly, emotions in voices may not necessarily be a “message” to another person, but may serve as tool

to navigate or influence one's social environment (Schirmer et al., 2022). A fear scream, for example, by sounding unpleasant might serve as a defense mechanism that effectively deters an assailant (Bachorowski & Owren, 1995; Schirmer et al., 2022). These viewpoints do not necessarily exclude each other - auditory emotions are presumably both signals and tools. However, the degree to which different auditory channels serve these functions could differ between music and voices: While vocal emotions result from an agent's current emotional state, musical emotions result perhaps from a more deliberate/explicit communication process. Composers purposefully translate feeling states or intentions into sounds so as to reach an audience. Moreover, music interpreters and performers explicitly reflect on what might be a composer's emotional message as part of their rehearsal work and training. Further, music consumption in Western cultures is predominated by settings with a clear sender/receiver distinction. By contrast, vocal emotions can be found in interactions in which individuals take on more reciprocal roles when behaving nonverbally. Taken together, although vocal and musical emotions share intriguing similarities, they may serve somewhat different functions with the latter being perhaps more intentional in nature.

On a side note, conceptualizing vocal emotions as tools may challenge the ecological validity of explicit emotion categorization tasks, since they do not entirely capture the way vocal emotions are "used" in daily life (Schirmer et al., 2022). However, it may be expected that musicians can cope better with such an explicit categorizing of emotions, because this approximates their analytic work with music. In the course of practicing a musical piece, emotion categories are often specifically identified, and their expression is expressly pursued. Therefore, future research should probe musicality benefits for vocal emotion perception using implicit measures and brain responses, so as to ascertain that these benefits are not strictly measure dependent (I. Martins et al., 2022).

6.4.3. The relative importance of pitch contour (F0) and timbre

In the present data, we found pitch contour (F0) to be more important than timbre for successful recognition across all emotional categories. This finding is in line with early work highlighting the perceptual dominance of pitch cues in vocal emotions (Banse & Scherer, 1996; Juslin & Laukka, 2003). However, performance in the F0 condition, with timbre rendered uninformative, was still worse than in the Full condition, suggesting that timbre carries unique emotional information as well (Grichkovtsova et al., 2012; Nussbaum, Schirmer & Schweinberger, 2022). This is also reflected in the emotion-specific perceptual dominance of F0, which was calculated as the performance difference_{F0-Tbr} for each emotion separately: The biggest dominance of F0 over timbre cues was found for happiness, whereas the smallest F0 dominance emerged for pleasure and sadness. This finding could be related to studies that highlight the importance of timbre for the perception of sadness (Grichkovtsova et al., 2012) and pleasure (Nussbaum, Schirmer & Schweinberger, 2022, Chapter 5). Minor differences between studies in the relative importance of both acoustic cues for these emotions may be due to their use of different emotional voice databases and the fact that voices can vary substantially in how they are affected by, and communicate, emotions (Spackman et al., 2009).

6.4.4. Constraints on generality, and future directions

Although the present study has a number of methodological strengths, certain choices in sample and design pose limitations and set directions for further research. One aspect that should be kept in mind is that the present study investigated vocal emotion perception from brief pseudoword stimuli, such that further studies with longer utterances of emotional voices (e.g., sentences or pseudosentences) will be needed to reveal the generality of the present findings. Regarding the sample, we acknowledge targeting a population socialized in Western music culture. Additionally, participants were native or fluent German speakers to ensure that the pseudowords used in the study were not perceived as semantically meaningful. Therefore, our findings may not generalize to individuals with a different musical culture or language background (Morrison & Demorest, 2009). Indeed, one would wish to see similar studies conducted with other, more diverse samples.

Further concerning the sample, we note that, despite our best efforts to ensure group comparability, musicians and non-musicians differed in terms of neuroticism and autistic traits. Because these traits did not correlate with vocal emotion perception in the present study, they are unlikely to explain the benefit of musicality. Nevertheless, the differential link between musicality and autistic traits seems worth exploring in more detail, as other studies reported relationships between autistic traits and voice identity perception (Skuk et al., 2019) as well as emotional processing (Di Yang et al., 2022). While not differing on the total AQ score, musicians seem to score lower on the social communication domain, but higher on the attention to detail domain, when compared with non-musicians. The idea of insular talents such as musical aptitude in people with clinical levels of autism is not new (Heaton et al., 1998). Further, autistic traits appear to correlate with pitch perception and absolute pitch in particular (Bonnell et al., 2003; Wenhart et al., 2019). In non-clinical populations, musical skills have been linked to detail-oriented processing (Wenhart & Altenmüller, 2019). However, to date, it is not fully understood how different aspects of autistic traits affect musical aptitude and musical experiences (Sivathanasan et al., 2022), which could be worth exploring in the future.

Additionally, a particularly interesting comparison for future research would be that between singers and instrumentalists. Our sample was too dominated by instrumentalists to allow for a meaningful analysis of subgroups. Nevertheless, we observed a correlation between self-rated singing abilities and emotion recognition performance. This seems intuitive, since singing provides the form of musical expression that is most closely related to vocal emotions. However, it should be noted that the only study that has compared instrumental vs. singing classes suggested that singing could actually interfere with vocal emotion perception (Thompson et al., 2004). This was unexpected, even for the authors, and the degree to which this finding is generated by methodological constraints has been intensely debated (Lima & Castro, 2011; Lolli et al., 2015; Nussbaum & Schweinberger, 2021; Thompson et al., 2004). Of interest in this context, a recent study observed similar brain responses to emotional sounds in singers and instrumentalists (I. Martins et al., 2022). On balance, the available literature does not paint a consistent picture concerning singers vs instrumentalists, and this issue deserves more systematic investigation.

6.5. Summary and outlook

Here, we report a robust advantage for musicians when compared with non-musicians in vocal emotion perception. Moreover, we show, using a novel voice manipulation approach, that pitch contour (F0) information plays a more important role than timbre across emotions and listeners and explains the musicality advantage. Further exploratory analyses revealed a link between auditory sensitivity in voices and music, especially for pitch cues. This link persists in the absence of formal musical training, suggesting that natural auditory sensitivity, rather than formal music training, drives the transfer benefits of musicality in the context of vocal emotion perception. Future research should expand these findings by comparing different listener subgroups such as singers vs. instrumentalists. The possible role of individual differences in personality and autistic traits for the complex interplay between musicality and vocal emotion perception might be another promising path for future exploration.

7. Electrophysiological correlates of vocal emotional processing in musicians and non-musicians

Abstract

Musicians outperform non-musicians in vocal emotion perception, but the underlying mechanisms are still debated. Performance measures highlight the importance of auditory sensitivity towards emotional voice cues. In particular, musicians seem to be sensitive to the pitch contour (F0) of vocal emotions. However, it remains unclear whether and how these group differences in acoustic processing are reflected at the brain level. To address this, we compared musicians' (N = 39) and non-musicians' (N = 39) event-related potentials (ERPs) to acoustically manipulated voices. We used parameter-specific voice morphing to create and present vocal stimuli that conveyed happiness, fear, pleasure, or sadness, either in both acoustic cues, or selectively in either F0 contour or timbre only. Although the fronto-central P200 and N400 components were modulated by the F0 and timbre manipulation, prominent group differences between musicians and non-musicians were neither observed in these ERP components nor in the following centro-parietal LPP. However, there was a correlation between individual music perception skills and the overall P200 amplitude. Additionally, exploratory analyses revealed group differences in later time windows (> 700 ms past voice onset) for sadness and happiness, potentially suggesting differences in appraisal processes. In sum, while this study did not reveal prominent group differences in early ERPs to vocal emotions, music perception skills seem to affect electrophysiological responses to vocal emotions following the early acoustic analysis of sounds.

7.1. Introduction

In the EEG-study of Chapter 5, we observed parameter-specific modulations of the P200 and the N400 in a fronto-central ROI. These ERP modulations seemed to reflect the relative importance of F0 and timbre for different emotions. In the present study, we wanted to expand these findings by studying individual differences. Specifically, we explored how these ERP modulations may be modified by musicality. In their behavioral performance (Chapter 6), musicians displayed specific sensitivity towards F0 cues. Furthermore, we observed a correlation between vocal emotion perception and musical hearing abilities that even persisted in the absence of formal musical training. As discussed in Chapter 6, these findings suggest that the musicality benefit is largely acoustic-bound. Therefore, musicality may modulate early stages of vocal emotional processing, which involve the acoustic analysis and integration of emotional voice cues (Schirmer & Kotz,

2006). However, behavioral measures alone give no insight into the timing and neural processes underlying the musicality benefit for vocal emotions. In particular, it remains possible that differences between musicians and non-musicians would emerge at later processing stages and could be related to cognitive and more top-down regulated evaluation of the acoustic patterns. To address this question, we conducted an ERP study. Therefore, we re-invited the musicians and non-musicians recruited for the online study in Chapter 6 to the lab and recorded their EEG, while they listened to the vocal emotional stimuli again. In what follows, I will review ERP effects suggesting differential electrophysiological responses in musicians and non-musicians across different types of auditory stimuli. Subsequently, I outline the rationale and hypotheses of the study.

7.1.1. Auditory evoked potentials related to musical expertise

There is much evidence that effects of musicality can be observed in the electrophysiological brain response to auditory stimuli (Kraus & Chandrasekaran, 2010; Pantev & Herholz, 2011). Most insight stems from cross-sectional designs comparing musicians to non-musicians. Musicality has been found to modulate the N100, the P200, and the mismatch negativity (MMN) in response to musical stimuli (Chartrand et al., 2008; Pantev & Herholz, 2011; Pantev et al., 1998). At a general level, one intriguing finding was that the cortical responses tended to be strongest when musicians listened to their own instrument of expertise, which was taken as indication for a training effect (Pantev et al., 2001). At a more specific level of individual ERP components, evidence on the N100 is somewhat inconclusive, as several studies failed to find differences between musicians and non-musicians, when listening to music tones vs. pure tones (Lütkenhöner et al., 2006; Shahin et al., 2005). The P200, instead, seems to be a robust marker of musical expertise and differences between musicians and non-musicians seem to become more pronounced with increasing complexity of the musical stimuli (Chartrand et al., 2008; Shahin et al., 2003, 2005). However, these P200 modulations are still poorly understood, as linking them to performance outcomes has proven to be challenging (Sheehan et al., 2005). As it stands, the P200 may reflect an unspecific effect of auditory expertise in response to complex sounds (Chartrand et al., 2008). Similarly, greater MMN effects in response to musical stimuli have been taken as evidence for superior auditory change detection and pre-attentive processing in musicians compared to non-musicians (Koelsch et al., 1999).

Of importance, musical expertise modulates auditory evoked brain responses beyond the musical domain. Differences between musicians and non-musicians have been reported for speech (Besson et al., 2007; Kaganovich et al., 2013; Schön et al., 2004) and nonverbal vocal expressions (Strait et al., 2009). A feature these findings have in common is that group differences are typically most pronounced in acoustically more complex conditions. For example, Besson et al. (2007) reviewed several studies targeting the detection of pitch incongruities in speech stimuli. Differences between musicians and non-musicians became usually apparent in conditions with weak incongruity, that was harder to detect, but not in strong incongruity conditions which are easier to perceive. While group effects were consistently visible as larger positivity in musicians between 100 and 300 ms past voice onset, they were also found in a later positivity between 400 to 700 ms.

Concerning musicality effects on vocal emotion perception, electrophysiological evidence is relatively inconsistent, which perhaps is partially due to variability between studies in terms of samples, stimuli, and tasks. In a study by Pinheiro et al. (2015), a reduced frontocentral P50 (~ 50 ms) was observed in musicians compared to non-musicians, but effects were similar for neutral and emotional prosody. Likewise, Rigoulot et al. (2015) could not find a clear pattern concerning emotional processing of voices in musicians and non-musicians. I. Martins et al. (2022) studied ERP responses to music and nonverbal vocalizations in an implicit listening task. They reported differences between musicians and non-musicians in terms of larger amplitudes seen in musicians at central and fronto-central electrodes in the P200, the P300, and the LPP in response to musical stimuli, but not to emotional vocalizations. To which degree these findings generalize to explicit listening tasks and emotional prosody perception remains unresolved. This relative lack of evidence in electrophysiological data stands in contrast to the much larger body of literature reporting consistent musicality benefits for vocal emotion perception in the behavioral domain (cf. Chapter 3). Likewise, in Chapter 6, we found a robust performance difference between musicians and non-musicians in the behavioral data, which was specifically related to pitch contour processing. In the present study, we explored the same sample with regard to electrophysiological measures. While ERPs related to enhanced pitch processing in musicians have been studied for music and speech stimuli (Besson et al., 2007), this has – to the best of our knowledge – never been done in the context of vocal emotions. Here, we aim to close this gap. To this end, we targeted ERP components that have been previously shown to be modulated by acoustic cues and musical expertise (the P200, the N400 and the LPP) in different contexts. The next section provides further details on the rationale of this EEG study.

7.1.2. Rationale of the study

This EEG study was designed as both a replication and an extension of the EEG study reported in Chapter 5. The replication efforts mainly targeted the parameter-specific modulations of the P200 and the N400 that we had observed when comparing responses to F0 and timbre for different emotions. EEG data in Chapter 5 had been analyzed with an exploratory approach, that resulted in the identification of a fronto-central ROI. Therefore, in this study, we primarily focused on this cluster with a more hypothesis-driven approach. The extension of the previous EEG study targeted several aspects: First, we investigated individual differences by comparing musicians and non-musicians. Note that all participants of this study had previously contributed data to the behavioral study reported in Chapter 6, such that we had the opportunity to explore potential relationships between individual ERP findings and independent behavioral observation. Second, we explored whether previous findings would generalize to a slightly different stimulus set with a different morphing-approach (see below). Third, alongside the P200 and the N400, we explored an additional centro-parietal component in a later time interval, which has been referred to in the literature as the late positive potential (LPP, Schirmer & Kotz, 2006). The LPP has been linked to the elaborative processing of vocal emotions (Hajcak & Foti, 2020; I. Martins et al., 2022), and an LPP with similar latency and topography has also been implicated in the processing of facial emotions (Schupp et al., 2004).

These extensions entail three important design adjustments of this study compared to the one reported in Chapter 5: First, we recruited a different sample. Most of the participants were not from the student population, and in fact were slightly older (~8-10 years). Unlike the previous EEG study, there was an almost equal number of male and female participants in the sample. Second, we used a different stimulus set, employing a modified voice morphing approach to create the acoustically controlled stimuli. Specifically, we used emotional averages instead of neutral voices as reference condition. We opted for this stimulus set because we could show in Chapter 4 that F0 and Timbre morphs that were created with this reference condition did not differ systematically in perceived naturalness. In parallel, the key feature of the parameter-specific voice morphs, which is the expression of emotional quality through specific vocal cues only, was preserved in these stimuli. Note also that these stimuli were identical to the ones used in the behavioral study in Chapter 6. Third, we changed the response format to allow for a meaningful assessment of later ERP components. In Chapter 5, participants performed an emotion classification task in each trial and could enter their response directly after voice onset. As this design introduced motor confounds in later time intervals, we had to restrain analysis to 500 ms past voice onset. In the present study, to permit the analysis of later time intervals, participants listened passively to the presented stimuli. However, to ensure their attention to the emotional quality of sounds, they were prompted for an emotion classification in about 10% of the trials past voice offset.

7.1.3. Hypotheses and analysis plan

Based on the findings reported in Chapter 5, we expected differential modulation of ERPs (P200 and N400) for F0 vs. Timbre conditions. In Chapter 5, the amplitude difference of the P200 and the N400 between the F0 and Timbre condition matched the relative importance of these cues for behavioral performance, and depended on the emotional category. For pleasure only, Timbre seemed more informative than F0, which was also reflected in opposite ERP patterns compared to the other emotions. In the behavioral performance of the current sample (Chapter 6), the pattern is slightly different: F0 seemed more informative for all emotions, although the relative importance of Timbre (in terms of a smaller magnitude for this difference between F0 and Timbre) again was observed for pleasure. However, making specific predictions about how this behavioral pattern would be reflected in the present ERP data was challenging, since ERP and behavioral data came from two different sessions with slightly different tasks (emotional classification in 100% vs 10% of trials).

EEG-studies targeting effects of musicality on vocal emotional processing are rare, and tend to be inconsistent. By contrast, the behavioral findings in Chapter 6 presented a very consistent picture highlighting the specific sensitivity for F0 cues in musicians. If this F0-sensitivity would be reflected in the ERPs, one would expect an interaction of musicality and morph type. A modulation of early ERPs (P200) would suggest that musicality affects automatic and largely bottom-up processes of acoustic analysis and emotional integration. Modulations in later ERPs (N400 and LPP) would suggest an involvement of higher-order and more controlled processes.

In a first step, the data analysis was closely aligned with the findings from Chapter 5, and therefore focused on the P200 [150, 250 ms], N400 [300, 400 ms] in a fronto-central ROI. In addition, we analyzed the LPP [400, 700 ms] in a centro-parietal ROI. In a second step, we report some exploratory analyses, such as cluster-based permutation tests and correlation analyses. Please note that, in view of this unique set of data, further EEG analyses (e.g., time-frequency analyses) are currently planned, but are beyond the scope of this dissertation.

7.2. Method

7.2.1. Participants

The study consisted of two parts: all participants first completed the online study (reported in Chapter 6) and were subsequently invited into our laboratory for an EEG session. Due to the Covid-19 pandemic, it was virtually impossible to control the time interval between the online and the lab experiment, which therefore ranged between several hours and almost two months. Of the 81 participants who completed the online session, 80 came to the EEG session. EEG data of two participants (one musician, one non-musician) had to be excluded due to bad data quality (extensive drifts and muscle artifacts in both cases). Please note that the samples of the online study and EEG-study overlapped with few exceptions only (online: 81 participants, four exclusions; EEG: 80 participants, two different exclusions). A detailed summary of the sample characteristics can be found in the Appendix, Tables A.6, A.7, A.8, and A.10.

Musicians Data from 39 musicians entered analysis (18 male, 21 female, aged 20 to 42 years [$M = 29.9$; $SD = 5.48$]). Mean onset age of musical training was 7 years ($SD = 2.54$, 4 – 17 years).

Non-musicians Data from 39 non-musicians entered analysis (19 male, 20 female, aged 19 to 48 years [$M = 30.5$; $SD = 6.34$]).

7.2.2. Stimuli

The stimulus material was the same as in Chapter 6.

7.2.3. Design

EEG-setup The EEG-setup was identical to the one described in Chapter 5.

Procedure Participants were instructed to listen to the presented voices and pay attention to vocally expressed emotion. To ensure attention, participants were asked to classify the emotion in about 10% of the trials. Assignment of emotions to response keys was identical to the online study for each participant. Each trial started with a white fixation cross centered on a black screen. After 1000 ± 100 ms (jittered randomly), the cross changed into green and a vocal stimulus started playing, followed by 2000 ms of silence during which the green fixation cross stayed on the

screen. Only in those 10% of trials with a response prompt, a screen displaying the four response options appeared, which lasted until the participant entered a response. Then the next trial started. The experiment started with 20 practice trials encompassing stimuli not used thereafter. Subsequently, all 312 experimental stimuli were presented once in random order, and then again in a different random order, resulting in 624 trials. Individual self-paced breaks were encouraged between blocks of 78 trials. The duration of the experiment was about 50 to 60 minutes and, including pre- and post-processing, participants were around 90 minutes in the lab.

7.2.4. EEG-data processing and analysis

EEG data analysis was done using EEGLAB (Delorme & Makeig, 2004) in Matlab R2020a (MATLAB, 2020). Preprocessing steps such as epoching, down-sampling, re-referencing, filtering, artifact rejection/correction, and CSD transformation were identical to the ones reported in Chapter 5. ERPs were derived by averaging epochs for each condition and participant. In one recording (musician), about 70 trials were lost due to a malfunction of the headphones. In another one (non-musician), around 30 trials were lost due to extensive coughing of the participant. After visual inspection, ERPs of both datasets were found to be of sufficient quality and were therefore kept for analysis. In total, a minimum of 32 trials and an average of 46.2 trials per condition (out of a possible maximum of 48) and participant entered statistical analysis. The condition with averaged emotions was excluded from the data analysis, to focus on the three morph types (Full, F0, Timbre) x four emotions (happiness, pleasure, fear, and sadness).

In the first part of the analysis, we focused on the fronto-central cluster we identified in Chapter 5 [F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2]. We quantified mean amplitudes of the P200 [150, 250] and the N400-like negativity [300, 400]. In addition, we analyzed a later interval ranging from 400 to 700 ms, which we refer to as LPP [400 700]. The LPP has a more centro-parietal distribution (Hajcak & Foti, 2020), which is why we used a different ROI, shifted to parietal electrodes [C1, Cz, C2, CP1, CPz, CP2, P1, Pz, P2] to quantify this component.

Subsequently, we ran exploratory cluster-based permutation tests on all 64 electrodes in a latency range from 0 to 1000 ms; using the FieldTrip toolbox (Maris & Oostenveld, 2007; Oostenveld et al., 2011). The analyses were done separately for each emotion using the Monte Carlo method with 1000 permutations and a minimum cluster size of two channels.

Participants were prompted for behavioral emotional classification response in 10% of the trials. No analysis was planned for these behavioral responses, as these trials were picked fully randomly and therefore varied between participants and conditions. However, we provide some descriptive data and visualization of these trials at the end of the results section.

7.3. Results

7.3.1. Analysis of the P200, the N200, and the LPP

Mean amplitudes of the **P200**, **N400** and the **LPP** were analyzed in three different $3 \times 4 \times 3$ ANOVAs with the between-subject factor Group (Musicians, Non-musicians), and the within-subject factors Emotion (happiness, pleasure, fear, and sadness) and Morph Type (Full, F0, and Timbre). A summary of all main effects and interactions is displayed in Table 7.1. In all ERPs, an interaction of **Emotion x Morph Type** was observed (see Figure 7.1 for the P200 and the N400 and Figure 7.2 for the LPP). In the N400 and the LPP, there were also main effects of **Emotion** and **Morph Type**. There were no significant main effects or interactions involving **Group**.

In a follow-up analysis, we specifically tested the difference between **F0** and **Timbre** for each emotion separately. For the P200, there was a significant difference in Pleasure, $|t(77)| = 2.71$, $p = .008$, $d = -0.31$ $[-0.08, -0.54]$, and Fear, $|t(77)| = 2.43$, $p = .017$, $d = 0.28$ $[0.05, 0.50]$. For the N400, there were no significant effects for F0 vs. Timbre. For the LPP, there was only a marginal effect for Happiness, $|t(77)| = 1.86$, $p = .066$, $d = 0.21$ $[-0.01, 0.44]$.

Table 7.1.: Results of the $3 \times 4 \times 3$ mixed-effects ANOVAs on mean amplitudes of the P200, the N400 and the LPP

	df1 2	P200			N400			LPP		
		F	p	ω_p^2	F	p	ω_p^2	F	p	ω_p^2
Group (Gr)	1 76	0.78	.381	<.01	0.01	.932	.01	0.52	.474	<.20
Emotion (Emo)	3 228	1.61	.189	<.01	6.30	<.001	.06	7.51	<.001	.08
MType	2 152	0.23	.792	.01	3.41	.036	.03	6.09	.003	.06
Gr x Emo	3 228	0.65	.582	<.01	0.89	.448	<.01	1.61	.189	.02
Gr x MType	2 152	0.98	.377	<.01	0.01	.986	.10	0.12	.885	<.01
Emo x MType	6 456	4.05	.001	.04	5.12	<.001	.05	4.23	.001	.05
Gr x Emo x MType	6 456	1.01	.421	<.01	0.95	.462	<.01	1.45	.197	.02

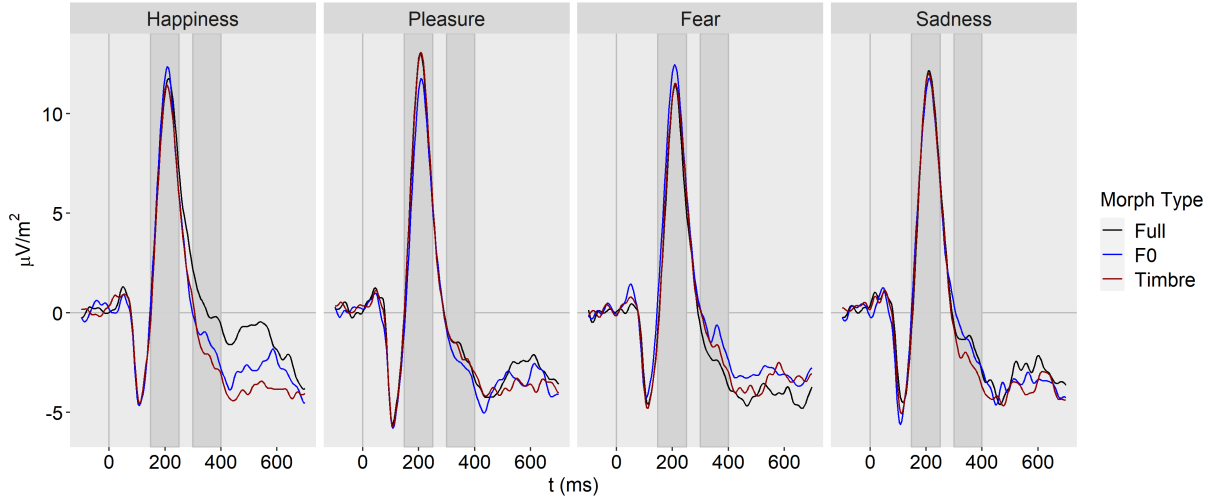
Note. Gr = Group, Emo = Emotion, MType = Morph Type.

7.3.2. Correlations between ERP amplitudes and the PROMS

In a subsequent exploratory analysis, we correlated the ERP amplitudes with the averaged music perception performance in the PROMS (for details, see Chapter 6). We included only data from participants that were kept for behavioral data analysis in Chapter 6. Thus, correlations were based on 74 participants. Average performance in the PROMS was positively correlated with the amplitude of the P200 ($r(74) = 0.26$, $p = .025$), but not the N400 ($r(74) = 0.07$, $p = .535$) or the LPP ($r(74) = 0.17$, $p = .150$), see Figure 7.3. In a second analysis, we correlated the averaged PROMS with the amplitude difference between the F0 and the Timbre conditions for the P200,

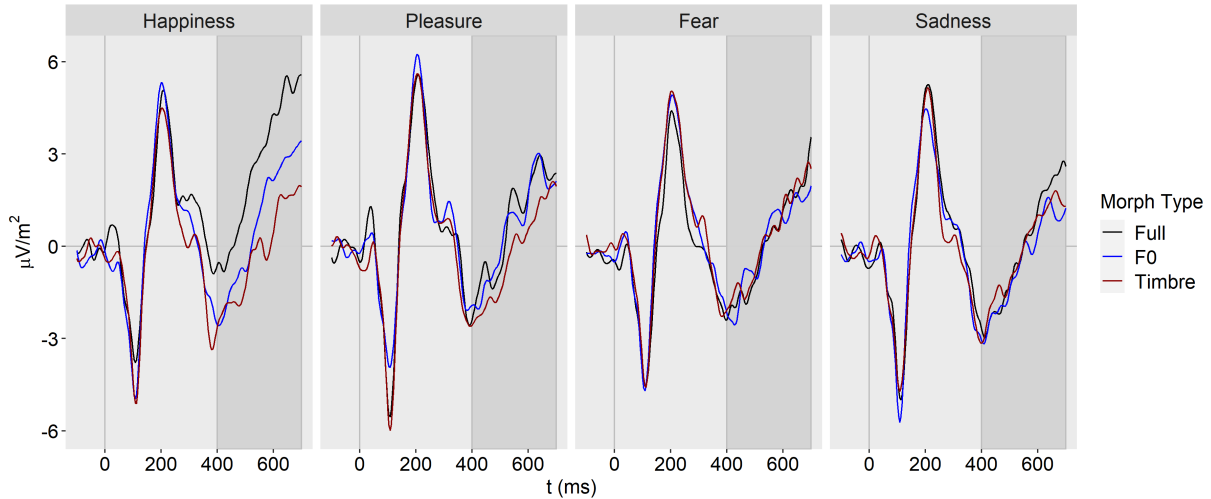
N400, and the LPP. However, the F0 vs. Timbre difference was not linked to musical expertise in any of the ERP components ($rs(74) \leq 0.11, ps \geq .309$).

Figure 7.1.: ERPs separately for Emotion and Morph Type – fronto-central ROI



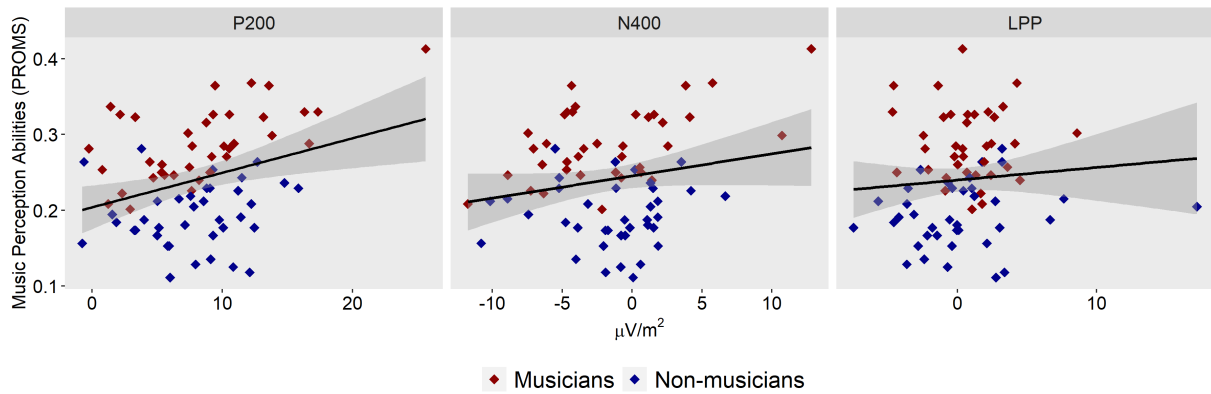
Note. Averages are collapsed across $[F1, Fz, F2, FC1, FCz, FC2, C1, Cz$ and $C2]$. Gray shaded areas illustrate the time window of the P200 [150, 250] and the N400-like negativity [300, 400].

Figure 7.2.: ERPs separately for Emotion and Morph Type – centro-parietal ROI



Note. Averages are collapsed across $[C1, Cz, C2, CP1, CPz, CP2, P1, Pz, P2]$. The gray shaded area illustrates the time window of the LPP [400, 700]. The time interval for the quantification of the LPP was chosen prior to data analysis based on previous literature (Hajcak & Foti, 2020). However, after visual inspection, it was noted that the LPP peaked later (~ 800 ms) in the present data. Please refer to Figure B.7 in the Appendix for an additional visualization.

Figure 7.3.: Relationship between music perception abilities (PROMS) and ERP amplitudes



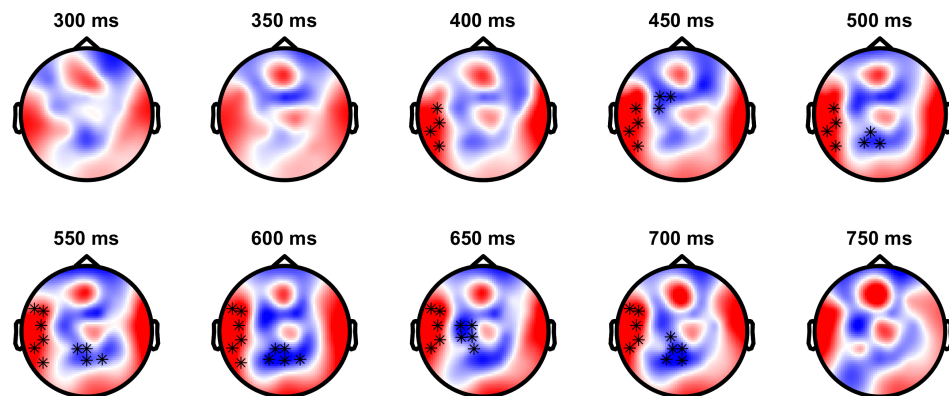
Note. Data points represent data of individual participants. The black line illustrates the linear regression, the shaded grey area around it the standard error.

7.3.3. Nonparametric cluster-based permutation tests

In another exploratory analysis, we ran cluster-based permutation tests to compare (1) F0 vs. timbre morphs and (2 & 3) musicians vs. non-musicians across all electrodes and timepoints until 1000 ms past voice onset. This was done to scan for any effects outside the ROIs.

First, we compared the ERPs in the F0 vs. the timbre conditions for each emotion separately. No clusters were found for pleasure, fear, and sadness. For happiness, the test revealed a significant difference between the F0 and the Timbre condition ($p < .05$), in a left-lateralized cluster between 360 and 720 ms, followed by a central cluster which appeared around 420 ms (Figure 7.4).

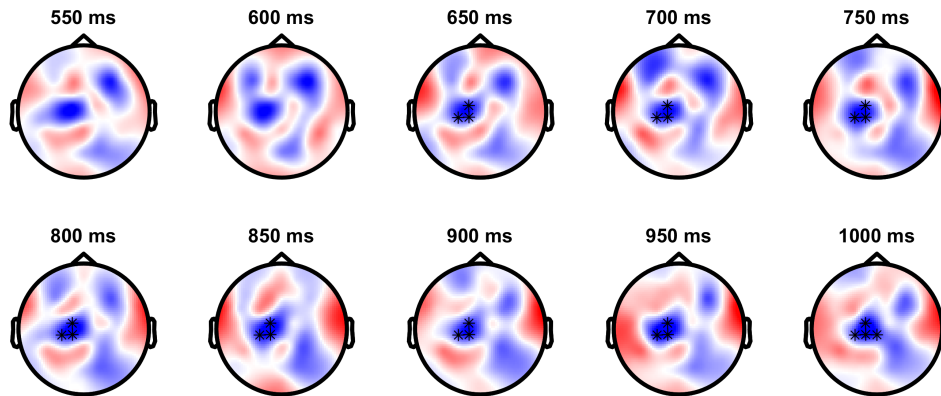
Figure 7.4.: Happiness – Scalp topographies of the contrast between F0 and timbre



Note. Clusters of significant differences are indicated by the black asterisks.

Second, we compared musicians vs. non-musicians for each emotion separately, averaged across morph types. No group differences were found for happiness, pleasure, and fear. For sadness, however, a late central cluster appeared between 640 until the end of the analyzed time window (Figure 7.5).

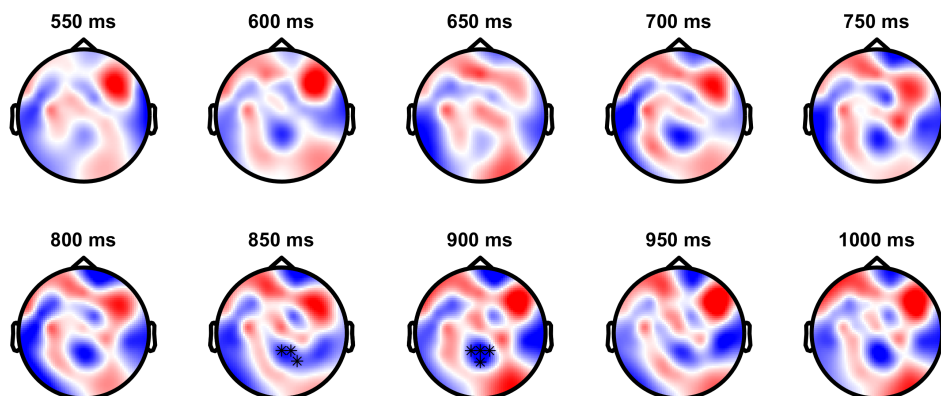
Figure 7.5.: Sadness – Scalp topographies of the contrast between musicians and non-musicians, averaged across morph types



Note. Clusters of significant differences are indicated by the black asterisks.

Third, we calculated difference waves between the F0 and Timbre conditions. These difference waves were compared between musicians and non-musicians, again for each emotion separately. No group differences were found for pleasure, fear and sadness. For happiness, however, a late central cluster appeared between 845 and 915 ms (Figure 7.6).

Figure 7.6.: Happiness – Scalp topographies of the contrast between musicians and non-musicians for F0-Timbre difference waves

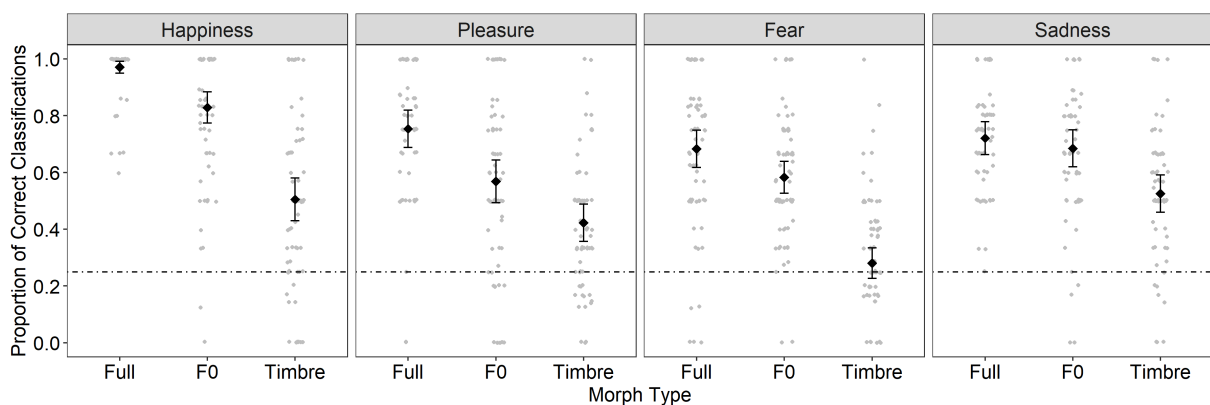


Note. Clusters of significant differences are indicated by the black asterisks.

7.3.4. Behavioral classification task

In about 10% of the trials, participants were prompted to classify the emotion expressed in the voice. This prompt was fully random, so the number of response trials differed across participants and conditions. The number of actual response trials ranged from 44 to 83 between participants ($M = 62$; $SD = 8.31$). The average proportion of correct classifications was $M = 0.61$ ($SD = 0.08$), ranging from 0.38 to 0.81. For Full morphs – supposedly the easiest condition – mean performance was $M = 0.78$ ($SD = 0.11$), ranging from 0.42 to 1.00. Thus, all participants classified the emotion above chance level (.25), suggesting that they paid sufficient attention to the expressed emotion of our stimuli. Figure 7.7 provides a visualization of correct classifications per Emotion and Morph Type, which resembles the pattern observed in Chapter 6, Figure 6.4, suggesting that participants responded in a similar way in the lab and the online session. However, due to the reduced and imbalanced dataset here, we refrained from statistical analysis and base this interpretation on visual inspection only.

Figure 7.7.: Mean proportion of correct responses per Emotion and Morph Type



Note. Whiskers represent 95% confidence intervals. Grey dots represent individual participants' data. The dotted line represents guessing rate at .25.

7.4. Discussion

The aim of the present study was to replicate and extend the findings of the EEP study reported in Chapter 5. Whereas the modulations of ERPs in response to F0 and timbre were partly replicated, the ERP findings did not reveal a conclusive pattern with regard to musicality effects. However, in a subsequent exploratory analysis, two patterns emerged that could be of potential interest: First, musical listening expertise, measured with the PROMS, was correlated with the P200 amplitude, across emotions and morph types. Second, cluster-based permutation tests revealed differences between musicians and non-musicians in later time intervals (> 640 ms), but only for specific emotions and contrasts: When comparing the electrophysiological response averaged across all morph types, we found a group difference for sadness. When focusing on the contrast between F0 and timbre, we found a group difference for happiness. Both clusters were

located in centro-parietal regions. In what follows, we will first discuss to which degree the role of F0 and Timbre is reflected reliably in electrophysiological responses. Second, we will discuss our findings on musicality in more detail.

7.4.1. Electrophysiological correlates of F0 and timbre processing

In the EEG study reported in Chapter 5, manipulation of F0 and timbre affected the amplitude of the P200 and the N400. The precise amplitude differences depended on the emotional category and reflected the relative importance of F0 vs timbre for behavioral performance. In the present study, we also observed robust interactions of the emotional category with the morphing condition in the P200, the N400 and the LPP. Thus, parameter-specific manipulation of vocal emotions seems to modulate both early and later electrophysiological responses. Furthermore, effects differed as a function of emotional category, similar to findings in Chapter 5. However, when we specifically focused on the comparison of F0 vs timbre within different emotions, patterns were less conclusive:

For happiness, we only observed a marginal LPP-effect in the ROI. This is in contrast to the big effect we previously observed in the P200 and N400, as well as the big performance difference between the two conditions observed both in Chapter 5 and 6. However, when we scanned for F0 vs. timbre differences beyond the ROI using a cluster-based permutation analysis (Figure 7.4), we observed two pronounced late clusters, in left-lateralized and centro-posterior regions. This suggests substantial differences between F0 and timbre for happiness, which were not observed for any of the other emotions. These clusters are reminiscent of the ones observed in Chapter 5 but have different timing and shifted spatial distribution. For pleasure, we observed an effect at the time of the P200 in the ROI, similar to the one in Chapter 5. The P200 amplitude was bigger in the timbre condition, compared to F0. In Chapter 5, this coincided with a behavioral advantage for timbre as well, which did not seem to be the case here (cf. Chapter 6). Note, however, that EEG and behavioral data were collected in two different sessions here and therefore cannot be linked directly. For fear, an effect was observed in the P200. In Chapter 5, this was observed later, at the time of the N400, but in a similar direction. For sadness, no differences between F0 and timbre were found. Similarly, we first did not find any effects for sadness after running the cluster-based permutation test on the previous data. A small difference for sadness only appeared when we specifically focused on the ROI.

While the present findings clearly show that manipulation of F0 and timbre in vocal emotions modulates different ERP components associated with their emotional integration and cognitive evaluation, they do not fully replicate the patterns we observed in the previous study (Chapter 5). Deviating findings could be the consequence of key changes in the design. Most importantly, we used a different stimulus set, with morphs based on averaged emotions as a reference category. On a behavioral level, stimuli resulted in a very similar performance pattern as the ones with neutral voices as reference, except for the role of timbre in pleasure (Figure 5.2 and 6.4). Nevertheless, these stimuli differed with regard to their acoustic composition and subsequently their perceived naturalness (Chapter 4). In voice morphs with neutral reference, which were used in the first

EEG study, naturalness was reduced in the timbre compared to the F0 condition, creating a potential confound. In voice morphs with averaged emotions as references, used in the second EEG study, naturalness of timbre and F0 morphs was comparable. While the behavioral measures seem to be remarkably robust against these factors (cf. Chapter 4), this does not necessarily hold for the electrophysiological correlates. For example, Schirmer and Gunter (2017) found that the N100, the P200, the N400, and the LPP were affected by the “voice-like” quality of stimuli, which interacted with the emotional processing. Therefore, effects of naturalness, (i.e. “human-likeness”) of voices should be considered, especially when manipulated stimulus material is used to study electrophysiological outcomes.

Another contributing factor may have been the behavioral task. In both studies, we explicitly instructed participants to focus on the expressed emotions. However, in the first study, participants had to enter a response in every trial and could do that right after voice onset. In the present study, responses were only collected in 10% of the trials, after voice offset. Thus, during presentations of the voices, participants could not know yet whether they would have to make a response later. These small changes could have resulted in a top-down modulation of neural activity. In fact, several studies reported that the direction of listeners’ attention can modulate ERPs related to vocal emotional processing. For example, sex differences were found in paradigms which targeted pre-attentive processing of vocal emotions (i.e. as indicated by the MMN), but were no longer observed when listeners’ attention was directed to the emotional prosody of voices (Schirmer, Kotz & Friederici, 2005; Schirmer, Striano & Friederici, 2005). Further, Schirmer and Kotz (2003) found an interference effect of emotional prosody on judgments of word valence, reflected in a larger N400 in emotionally incongruent conditions, but not vice versa. Paulmann et al. (2013), in contrast, compared ERPs in response to emotional prosody while listeners either rated the speakers’ arousal (explicit condition) or their own (implicit condition), but did not find any task effects. Thus, although the behavioral task and attentional focus of listeners can affect brain responses to vocal emotions, the precise mechanisms are not yet fully understood. In the present paradigm, the listeners’ focus may have been less on conscious emotion recognition, which could have made the acoustic manipulation less impactful.

In summary, the present findings suggest that the processing of F0 and timbre is reflected in brain responses to vocal emotions. However, a systematic pattern remains elusive, as the observed parameter-specific modulations seem amenable to features of the stimuli and the participants’ task. Future research which quantifies the contribution of stimulus and design factors is therefore necessary for a more specific understanding of F0 and timbre effects in the context of EEG experiments.

7.4.2. Electrophysiological correlates of musical expertise

In the present data, no group differences between musicians and non-musicians were observed for the P200, the N200 and the LPP. This finding is in contrast with the growing body of literature showing a reliable benefit of musicality for vocal emotion perception in the behavioral

domain (cf. Chapter 3). Instead, our results add to the inconclusive literature that target the electrophysiological correlates of this benefit (I. Martins et al., 2022; Pinheiro et al., 2015; Rigoulot et al., 2015). While reliable ERP differences between musicians and non-musicians could be identified for music and speech stimuli (Chartrand et al., 2008; Pantev & Herholz, 2011), this has not yet been successful for vocal emotions. For example, I. Martins et al. (2022) found that differences in musicians and non-musicians were displayed in saliency detection (P200), attention allocation (P3) and elaborative processing (LPP) of emotions in music, but not for voices. The authors attributed these findings to the privileged status of musical stimuli for musicians. The null-findings in our study seem to support this view. However, three exploratory findings suggest that musicians' brains may respond differently to vocal emotions nevertheless:

First, we found a correlation between the P200 amplitude and music perception abilities. As this analysis takes individual differences of musicality within groups into account, it has more statistical power than the group comparison. The finding is in line with previous data suggesting that the P200 amplitude is a reliable, but unspecific marker of auditory expertise (Chartrand et al., 2008). However, this effect could be either related to voice and emotional processing, or just reflect differences in generic attention towards the stimulus material. Further, correlations were not controlled for multiple comparisons and therefore await further replication.

Second, we observed ERP amplitude differences between musicians and non-musicians after running cluster-based permutation tests, but for sadness only. This may seem surprising, as musicians did not seem to display a specific advantage for sad stimuli in the behavioral data. It is, however, in line with an fMRI study by Park et al. (2015), that also reported an effect for sad stimuli in the brain, but not at the behavioral level. Park et al. (2015) observed differences in frontal areas associated with higher order functions, such as evaluative judgements or empathic engagement, which fits our observation of differences in a rather late time interval (>640 ms). The emotional specificity of this effect, however, remains subject to speculation. Park et al. (2015) hypothesized that sadness may be of "higher affective saliency" to musicians, resulting in a unique implicit representation that is traceable in neural markers, but not necessarily in behavioral outcomes. However, as this unique role of sadness was predicted neither by us nor by Park et al. (2015), this hypothesis awaits further testing.

Third, we explored the differential processing of F0 vs timbre cues. The behavioral findings reported in Chapter 6 showed that musicians seemed to be particularly tuned to emotional pitch cues. Using EEG, we wanted to explore if this benefit resides in an early integration of emotional voice cues or later elaborative processes. In the absence of any parameter-specific group differences in the pre-defined ROIs, we compared groups on the F0 vs. timbre contrast across all electrodes and timepoints. There was only one cluster for happiness in a late interval (~ 850 - 900 ms). This suggests that differences are observed in later and more elaborative processes, similar to the findings for sadness above. It is possible that effects of musicality emerge only during attentive and effortful processing of vocal emotions and depend less on pre-attentive mechanisms. In that case, the behavioral task may have hindered the detection of group differences,

as listeners' attention was somewhat shifted away from conscious decision making. With the acoustic manipulation being less impactful overall, group differences regarding F0 and timbre may have been too small to be detected, except for happiness, which displayed a pronounced difference between F0 and timbre, both at the behavioral as well as at the electrophysiological level. Note, however, that these findings are exploratory in nature and need further replication.

In summary, our findings did not reveal consistent differences between musicians and non-musicians. However, some additional exploratory analyses suggest that group differences may still be traceable in the electrophysiological responses nevertheless. Potentially, musicality effects could be revealed more consistently with sufficiently powered analyses and more suited electrophysiological markers, which we will discuss in the next section.

7.4.3. Future research

Research linking musicality to different ERP components has yielded conflicting results, potentially rendering ERPs insufficient as markers of musical sensitivity. Future research efforts may therefore explore other electrophysiological measures and analyses that have proven valuable for comparing musicians and non-musicians. One candidate is the Frequency following Response (FFR) in the brainstem (Kraus & Chandrasekaran, 2010). This neural oscillation measure is correlated with the age at onset and years of musical training, suggesting an influence of training and experience. Strait et al. (2009) found that musicians and non-musicians differ in their brainstem response to emotional infant cries.

In fact, an increasing body of studies now reports musicality-related effects on oscillatory brain activity (Bidelman, 2017; Shahin et al., 2008; Sorati & Behne, 2019; Trainor et al., 2009). For musical stimuli, gamma-band activity (30-100 Hz) has been identified as a marker of musical expertise (Shahin et al., 2008; Trainor et al., 2009). To the best of our knowledge, only one study investigated neural oscillations in musicians and non-musicians for vocal emotional sounds (Nolden et al., 2017). Differences were observed for theta- (4–8 Hz) and alpha-activity (8–12 Hz), but data was based on two speakers only. As our data suggest that the stimuli can affect electrophysiological measures, it may be insightful to explore these effects in neural oscillations using more diverse material in the future. Finally, great potential lies in the application of machine-based decoding algorithms for EEG data, which can have higher sensitivity for small effects than conventional ERP analysis (Grootswagers et al., 2017). In this framework, a classifier would be trained to distinguish between emotions based on EEG patterns. Decoding performance could then be compared between musicians or non-musicians.

7.4.4. Summary

In the present study, we explored electrophysiological correlates of acoustically manipulated emotional voices and compared them between musicians and non-musicians. While the acoustic manipulation of F0 and timbre modulated ERP components (P200, N400, LPP), findings were less consistent and only partly replicated the patterns observed in a previous EEG study. These

differences may be the results of several design adjustments, suggesting that the present ERP findings are prone to experimental features such as the specific vocal samples or the participants' task. Regarding musical expertise, results were inconclusive. There were no group differences in the P200, N400 or LPP overall. However, exploratory analyses revealed a correlation between musical listening abilities and the P200 amplitude, as well as differences between musicians and non-musicians in response to sadness and happiness in later time windows. These findings may suggest that musicians' brains respond differently to vocal emotions compared to non-musicians, nevertheless, but this claim requires further empirical validation.

8. General Discussion

In this dissertation, I addressed three main questions:

- (1) What is the contribution of different acoustic cues to vocal emotion perception?**
- (2) How does musicality affect the processing of emotional voice cues?**
- (3) Is parameter-specific voice morphing a suitable tool to study the processing of vocal emotions?**

Each of the following three sections will be dedicated to one of these questions. In each section, I discuss the empirical evidence, important implications, and potential limitations in detail, before I conclude with a short summary of the main findings. Subsequently, I outline how this work contributes to our understanding of vocal emotion perception and propose potential directions for future research.

8.1. The role of F0 and timbre for the processing of vocal emotions

A main objective of this dissertation was to shed light on the role of different acoustic cues for the processing of vocal emotions, with a specific focus on F0 and timbre. I probed the relative rather than the absolute contribution of these cues, an approach which is arguably more appropriate when cues are highly intercorrelated (refer to section 2.1.1 in the Introduction for a theoretical consideration). To this end, I used parameter-specific voice morphing to create stimuli that expressed emotions via F0 contour, timbre, or both, while other cues were held constant at a non-informative level. Compared to previous efforts that tried linking emotions to their underlying acoustics, the current approach is mainly distinguished by two aspects: First, it allows a quantification of the “unique contribution” of each cue, that is the emotional information in one cue that cannot be compensated by the other one. Second, the experimental manipulation of vocal emotions allows to establish a causal relationship between acoustic cues and perceptual outcomes, in contrast to purely correlational designs. In this dissertation, I was interested in recognition performance as well as the time-critical neural processing of vocal emotions. For this reason, I gathered behavioral and electrophysiological data; in two different samples and using slightly different stimulus sets (Chapters 5, 6, and 7). In what follows, I will first discuss the behavioral findings in more detail, then reflect on the implication of the electrophysiological data, and finally conclude with a consideration of limitations and open questions.

8.1.1. The role of F0 and timbre for emotion recognition

Overall, the empirical evidence shows that both F0 contour and timbre express **important and unique information** that signal emotional quality: On the one hand, emotional inferences were somewhat compromised in both the F0 and timbre conditions compared to the condition with full acoustic information, suggesting that both cues are needed for best possible performance. On the other hand, recognition in both F0 and timbre conditions was above chance for almost all emotions. This suggests that both cues can be used for successful emotional inferences in a partly interchangeable fashion, in line with Brunswik's lens model (Brunswik, 1956). These findings are not only consistent with previous analyses, which reported that both F0 and timbre cues predict recognition performance (Juslin & Laukka, 2003; Scherer, 2018), but go beyond the predominantly correlational literature, and establish a causal link of these cues with emotional inferences.

Despite the consensus that both cues carry significant emotional information, it has been debated whether one may still be more important than the other (Juslin & Laukka, 2003). Previous acoustic analyses of natural voice recordings highlighted the importance of F0 cues for vocal emotion perception, but acknowledged that the neglect of timbre features in many research paradigms could have biased these results (Scherer, 1986). With the present voice morphing approach, I strived to address this issue through an equivalent manipulation of both cues. Nevertheless, on average, **F0 contour was still found to be more informative than timbre**, supporting the prevailing view in the literature. Note that some studies using fully synthesized voices suggest a predominant role of timbre (there mostly referred to as voice quality) compared to F0 cues (Gobl, 2003; Yanushevskaya et al., 2018). However, these findings are only comparable to the degree to which these synthesized stimuli resemble human emotion expression.

Importantly, the present data suggest that the contribution of F0 and timbre differs as a **function of the emotional category**. In happiness and fear, F0 seems to be far more important than timbre, as indicated by a marked performance difference. In sadness and pleasure, this F0-dominance is less profound, or even reversed. Specifically, this pattern was consistent across the two studies reported in Chapter 5 and 6, with one exception: In pleasure, timbre was found to be more important than F0 in the first study (Chapter 5). In the second study, although performance in the F0 condition was elevated to a degree that it overtook timbre performance (Chapter 6), the performance difference was nevertheless smaller than for the other emotions. Therefore, the relative contribution of cues may slightly vary between specific stimulus sets and/or samples. In addition to the proportion of accurate responses, the patterns of misclassification can be a valuable source of information, which again were similar across the two studies: In the Full conditions, confusions occurred primarily within the same valence category, thus between happiness vs. pleasure, and between fear vs. sadness. In the parameter-specific conditions, all emotions were most frequently confused with sadness, an effect that was most pronounced in the timbre condition. One may speculate that the acoustic manipulation affected emotional intensity, and in case of doubt, sadness was picked by listeners as the least intense option.

While these data paint a consistent picture in itself, their comparability to other empirical findings is challenging due to the unique morphing approach and the rather uncommon selection of emotional categories (note that I included pleasure to balance the valence and arousal across the four emotions). Nevertheless, a number of selected publications reveal relevant parallels. First, Grichkovtsova et al. (2012) compared the contribution of timbre and prosodic contour (encompassing F0, loudness and temporal patterns) for different emotions using a prosody transplantation method and observed a marked performance difference for happiness in benefit of prosodic contour. In parallel, Waaramaa et al. (2008) found that happiness is poorly recognized in monopitched voices with limited F0 variation. Thus, the predominant role of F0 for the perception of happiness seems to be empirically well supported. Concerning our two emotions with lesser intensity – sadness and pleasure – several findings highlight the importance of timbre cues (refer to rating data in Appendix C for information on intensity of the present stimulus set). In sadness, Grichkovtsova et al. (2012) found timbre to be more important than prosodic contour, which is only partly in line with the present results and suggests again that the precise pattern depends on the stimulus set and the specific acoustic manipulations. In ambiguous vocalizations such as gasps and moans, listeners rely on timbre cues to distinguish between a pleased and a hurt voice (Anikin, 2020), highlighting the importance of timbre for the perception of pleasure. Although our data still suggests a predominance of F0 cues for these emotions, findings converge in showing that timbre is relatively more important in these less intense emotions as compared to happiness and fear.

These emotion-specific patterns prompted researchers to speculate whether F0 and timbre may serve distinct functions in emotional signaling (Ladd et al., 1985; Laukkanen et al., 1997; Tursunov et al., 2019). Mainly, two hypotheses have been considered: First, as F0 was consistently found to reflect unspecific arousal (Bachorowski, 1999; Brück et al., 2011), it was assumed that **timbre may be more effective in conveying valence**. Accordingly, automatic emotion classification of vocal sounds could be substantially improved in the valence domain via the incorporation of timbre cues (Tursunov et al., 2019). Furthermore, accurate valence perception was remarkably preserved in monopitched stimuli with very limited F0 information (Waaramaa et al., 2008). However, a rigid functional connection between F0-arousal and timbre-valence is at odds with the present data, where it should have become apparent in the patterns of misclassifications: As discussed in Chapter 5, in the F0 conditions, participants should have mixed up high arousal emotions with other high arousal emotions (i.e. happiness and fear) and low arousal emotions with other low arousal emotions (i.e. pleasure and sadness). Likewise, in the timbre condition, mix-ups should have happened primarily within rather than across positive (i.e. happiness and pleasure) and negative emotions (i.e. fear and sadness). Neither was observed in the present data. Instead, the available evidence matches better with a second hypothesis, suggesting that **timbre may be more effective in signaling milder affective states** (Gobl, 2003). Such mild emotions may be naturally more ambiguous with regard to their F0 contour, shifting listeners' attention towards timbre (Anikin, 2020; Gobl, 2003; Yanushevskaya et al., 2018). This would explain the increased contribution of timbre to the perception of sadness and pleasure as compared to happiness and fear. Accordingly, one potential implication of the present findings is that timbre

cues may deserve more attention in future research on the perception of subtle prosocial emotions in particular (Sauter, 2017). A future research project could investigate this hypothesis in induced emotions, which are usually more subtle and less intense.

However, although the present data suggest some systematic use of timbre and F0, a too rigid framework linking timbre and F0 to distinct functions is neither supported empirically nor conceptually. The lack of empirical evidence may be addressed by future research, but efforts trying to link acoustic cues to fixed functions could lead to similar frustrations as have the efforts to describe different emotional categories in terms of universal acoustic patterns (as discussed in Chapter 2.1.1), because – on a conceptual level - they disregard the variability and flexibility inherent to vocal expressions. To some degree, F0 and timbre may simply signal partly redundant information which are picked up by listeners in a flexible manner for a common means: the successful perception of emotion.

8.1.2. Electrophysiological correlates of emotional F0 and timbre cues

In two EEG studies, I explored how specific emotional voice cues modulate ERP responses in listeners. In the first EEG study (Chapter 5), I observed parameter-specific effects in the P200 and the N400 components in a fronto-central region of interest (ROI). These modulations mapped onto the relative importance of cues for behavioral performance. For example, the big performance difference between F0 and timbre for happiness was also reflected in a big amplitude difference in the P200 and the N400. These components have been linked to emotional integration and top-down modulated cognitive evaluation of acoustic stimuli (Schirmer & Kotz, 2006). In fact, the amplitude difference between F0 and timbre in the N400 predicted behavioral performance. However, in the second EEG study, which included musicians and non-musicians (Chapter 7), these findings were only partially replicated and appeared less conclusive. Parameter-specific modulations of ERPs within the fronto-central ROI were smaller or even absent for some emotional categories. After expansion to the whole electrode array, exploratory analyses revealed two pronounced late clusters for happiness, in left-lateralized and central/posterior regions. Taken together, these findings suggest that ERPs are affected by acoustic manipulation of vocal emotions, but specific effects can vary across studies.

In retrospect, variation in the effects observed here may be explained by changes in stimuli and/or task. As shown in Chapter 4, stimuli in the second ERP study differed from those in the first EEG study with regards to their perceived naturalness. In the first EEG study using neutral reference voices, naturalness was reduced in the Timbre compared to the F0 condition, creating a potential confound. In the second EEG study with averaged emotions as reference, naturalness of Timbre and F0 morphs was comparable. Although the behavioral measures were found to be remarkably unaffected by these differences (detailed discussion in section 8.3), they could have impacted on the electrophysiological responses (Schirmer & Gunter, 2017). Conceptual models on voice perception propose that voices are processed in neural networks which are not activated by other types of auditory stimuli (Belin et al., 2004, 2011). Empirical data, in contrast, showed

that auditory networks are not organized by sound class, but correlate with sound knowledge and experience (Schirmer et al., 2012). Either way, effects of unnatural voice features on brain responses seem plausible, as such voices sound both less human-like and as well as less familiar to listeners, potentially resulting in different temporal and spatial ERP effects. Based on the present data, however, these explanations are only speculative and require to be tested in future research.

Another key difference that could explain the inconsistent findings across the two studies is the behavioral task during the EEG recording: In the first study, participants performed an emotion classification task and could enter their response directly after voice onset. In the second study, they were only prompted for the emotion classification in 10% of the trials after voice offset. Although in both studies participants were instructed to pay attention to the expressed vocal emotion, this difference could have resulted in delays in emotional appraisal, and a correspondingly greater contribution of top-down modulation of neural activity, subsequently modulating the ERP components (Schirmer, Kotz & Friederici, 2005; Schirmer et al., 2002).

Taken together, the electrophysiological findings show that ERPs can be modulated by the acoustic manipulation of emotional voice cues. However, these modulations seem to be susceptible to design features and stimulus properties, making them hard to replicate. Understanding the impact of these experimental choices in more detail may reveal a more systematic pattern of parameter-specific ERP modulations in the future.

8.1.3. Open questions

While the present work offers original insight into the role of F0 and timbre for vocal emotion processing, several questions remain unanswered which could serve as starting points for future research efforts. The first question is inherent to the morphing procedure: the manipulation of specific voice cues inevitably breaks the natural co-occurrence of F0 and timbre cues. Thus, it allows an assessment of F0 and timbre in isolation, but gives less insight into their **potential interaction** (Chartrand & Belin, 2006; Singh & Hirsh, 1992). While previous findings suggest that timbre and F0 information interact on a perceptual level in vocal emotions (Gobl, 2003; Ilie & Thompson, 2011; Spackman et al., 2009; Yanushevskaya et al., 2018), the present data do not provide strong evidence for this claim. A big interaction effect would have become apparent if performance were close to chance in the F0 and timbre conditions, but clearly above chance in the full condition. This pattern was not observed for any of the emotional categories. Nevertheless, a potential interaction may be implied by the minor differences observed between the two studies (Chapter 5 and 6): Although patterns appeared to be mostly comparable, performance in the F0 condition of pleasure was elevated in the second one. Both datasets used the exact same emotional information in all stimuli, but differed with regard to the reference stimuli which contributed the emotionally non-informative portion. This performance difference might suggest that it made an impact which timbre information was coupled with the F0 contour of pleasure. Beyond this observation, the present paradigm offers only limited insight into the possible interdependence of F0 and timbre. In principle, however, voice morphing could be used to develop a paradigm which

allows a quantification of their individual vs. combined contribution. For example, instead of rendering specific cues completely uninformative, one could flexibly vary the emotional information expressed by both F0 and timbre to explore how they influence each other.

A related question concerns the role of different timbre parameters, such as harmonics-to-noise ratio or spectral energy distribution. Here, all these parameters were morphed in conjunction, but future research could target their isolated contribution (Piazza et al., 2018). Finally, I investigated vocal emotion perception from brief pseudoword stimuli only, such that further studies with longer utterances of emotional voices (e.g., sentences or pseudosentences) would be valuable to reveal the generality of the present findings.

8.1.4. Summary and conclusion: What is the contribution of different acoustic cues to vocal emotion perception?

In this dissertation, I showed that both F0 and timbre provide unique information that allows listeners to infer vocal emotions, although overall, F0 seems to be the predominant parameter. Their precise contribution, however, depends on the emotional category: In happiness and fear – emotions of high intensity –, F0 seems to be far more important than timbre, as indicated by a marked performance difference. In less intense emotions – sadness and pleasure – the contribution of F0 and timbre was more balanced, suggesting that timbre may be more effective in signaling milder affective states. In the electrophysiological data, I observed F0- and timbre-related effects in both early and later ERP components, suggesting that these cues modulate several neural processes associated with vocal emotion perception. In summary, these data show that F0 and timbre signal both unique as well as partly redundant emotional information, which can be picked up by listeners in a flexible manner to infer the expressed emotion.

8.2. Links between musicality and vocal emotion perception

Several findings suggest a link between musicality and vocal emotion perception, but evidence is heterogenous and thus offers only limited insight into the mechanisms underlying this link. In this regard, the contribution of this dissertation is threefold: First, I provided a systematic review of the existing literature in Chapter 3. Second, I assessed how musicians and non-musicians differed in their use of acoustic cues to infer vocal emotions (Chapter 6). Third, I recorded ERPs of both groups to explore how electrophysiological responses to vocal emotions would be modulated by musical expertise (Chapter 7). Overall, the behavioral data reveal a consistent picture: both previous findings and the present work support the notion that musicians outperform non-musicians in vocal emotion perception. Importantly, this benefit seems to be mediated by low-level auditory sensitivity, and a privileged processing of vocal pitch cues in particular. Electrophysiological correlates, by comparison, were less conclusive. There were no group differences in the P200, N400 or LPP. However, individual music perception skills correlated with the overall P200 amplitude. Further, exploratory analyses revealed group differences in

later time windows (> 600 ms past voice onset) for sadness and happiness. In what follows, I will review the role of auditory sensitivity for emotion perception in more detail. Subsequently, I will discuss to which degree active musical engagement vs. a natural musical aptitude may contribute to the benefit in vocal emotion perception. Finally, I will critically reflect on the electrophysiological patterns and open questions, which could be not answered conclusively in the present work.

8.2.1. Sensitive to melodies? – How musicality benefits the processing of vocal emotions

The study reported in Chapter 6 presents clear evidence that musicians display a specific **advantage for emotional pitch cues**, but not for timbre, in voices. Importantly, individuals with high musical abilities are not merely more sensitive to F0 contours than non-musicians (Strait et al., 2009), they also seem to rely on them to a larger degree than on timbre cues when making prosodic judgements (Cui & Kuang, 2019). In that vein, they are more proficient at making use of the more dominant acoustic cue for emotional signaling (refer to previous Discussion section 8.1). While this was already implied by previous literature reporting correlations between pitch processing and vocal emotion perception (Globerson et al., 2013; Lima & Castro, 2011), a distinctive feature of the present work is that it offers original causal evidence that is based on a direct acoustic manipulation of voice stimuli. While this study focused on highly trained musicians, its results closely mirror published findings on the tail-end of the musicality spectrum: as discussed in Chapter 3, difficulties in pitch perception have been consistently linked to vocal emotion perception problems in individuals with amusia (Lima et al., 2016; Lolli et al., 2015; Pralus et al., 2019; Thompson et al., 2004). As a limitation, however, it has to be noted that the criteria for the presence of amusia were based on pitch perception performance only in some of these studies, thus neglecting the potential impact of other cues. Further, findings from the low end of the musicality spectrum do not readily lend themselves to the conclusion that emotional benefits at the high end are underpinned by the same mechanisms. For example, while M. Martins et al. (2021) argued that the emotional benefits observed in highly trained musicians are restricted to the auditory domain, emotional difficulties observed in people with amusia may be more widespread across modalities (Lima et al., 2016). In that sense, the present findings are therefore not merely a replication of the patterns observed in amusia shifted to a different performance level, but constitute qualitatively distinct evidence that sensitivity to pitch cues mediates the musicality benefit in vocal emotions.

Furthermore, correlational patterns based on the PROMS music perception test highlight the importance of **dynamic rather than static auditory processing** for emotional inferences. The PROMS measures different subcomponents of musical sensitivity. Importantly, those subcomponents which require the tracking of acoustic information over time predicted emotion perception performance, whereas the ones which are based on comparison of static auditory snapshots did not (Chapter 6). This pattern is reminiscent of a previous report by Globerson et al. (2015), who found that emotion perception was linked to dynamic pitch change detection, but not to static pitch discrimination performance. As tracking of musical features over time is a crucial

component of music performance and perception (Strait et al., 2009), there is ample evidence for superior perception of timing and rhythmic patterns in highly trained musicians (Kraus & Chandrasekaran, 2010). To which degree this superior tracking in the time domain relates to vocal emotion perception remains unresolved. The present work provides correlational evidence for a strong link between rhythm perception in music and emotion recognition in voices. However, with the focus being on F0 vs timbre, timing information was held constant across emotions in the present research, thus preventing a causal assessment of this cue. In principle, however, timing information can be manipulated conveniently using parameter-specific voice morphing, offering a potential road for future research.

Besides auditory sensitivity, it has been debated to which degree **higher-level supramodal skills** such as empathy, emotion differentiation (i.e. the ability to discriminate between similar/subtle emotional qualities) or decision making are involved in the musicality benefit for vocal emotions. In that regard, the existing literature is inconsistent (Chapter 3 and Chapter 6). While Correia et al. (2022) suggested that the musicality benefit is fully mediated by auditory perception skills, Trimmer and Cuddy (2008) provided strong evidence against a link between auditory sensitivity and emotion perception, and interpreted the musicality benefit by means of non-auditory supramodal emotion skills instead. Yet, this claim is at odds with several findings that could indicate that the performance difference between musicians and non-musicians is restricted to the auditory modality (Correia et al., 2022; Twaite, 2016; Weijkamp & Sadakata, 2017). Furthermore, Farmer et al. (2020) reported that despite enhanced perception, feeling of others' emotions was unaffected, which makes differences in empathy seem unlikely. In the present study, there was a somewhat surprising correlational pattern, which could be interpreted in a similar vein: while I observed a consistent correlation of vocal emotion recognition with music perception skills, a correlation with the Emotion-subscale of the Gold-MSI was absent. This subscale covers "active behaviors related to emotional responses to music" (Müllensiefen et al., 2014), and thus rather high-level and declarative forms of emotional engagement. The absence of any correlation with vocal emotion perception renders the impact of supramodal emotional reasoning unlikely and is consistent with the notion of **low-level auditory sensitivity underlying the observed benefit in musicians**.

8.2.2. The role of musical training vs natural auditory sensitivity

Musical skills emerge as a result of both training, i.e. through active musical engagement, and aptitude, i.e. pre-existing differences in natural auditory sensitivity. Therefore, the mere presence of a performance difference between musicians and non-musicians does not allow inferences about the underlying causal mechanisms. Ideally, this question would be resolved with longitudinal randomized studies. However, as only very few of such studies have been conducted in the field of vocal emotion perception (for details, refer to Chapter 3), the contribution of nature vs. nurture aspects to the musicality benefit remains a matter of debate.

As discussed in Chapter 3, several authors argued that **active engagement in a musical task** over a longer time period is a crucial factor for the development of auditory skills. Brain data suggest that enhanced connectivity in the auditory-motor domains and a synchronized co-activation lead to a more sophisticated representation of complex sounds in auditory networks, subsequently enhancing auditory sensitivity (Kraus & Chandrasekaran, 2010; Lappe et al., 2008; Palomar-García et al., 2017). In fact, two studies targeted this effect for vocal emotion perception in CI users and compared auditory-motor to auditory-only musical training. Intriguingly, both studies observed improvements for the auditory-motor intervention only (Chari et al., 2020; Fuller et al., 2018). These findings advocate for active musical engagement in the context of hearing rehabilitation with a CI, where individuals face the challenge of massive postimplantation adaptation to degraded auditory input. However, they may not generalize to the normal-hearing population, where evidence for an effect of active musical engagement is far less conclusive (M. Martins et al., 2021). Although several studies claim that musical training affects emotion perception performance, findings are heterogeneous, conflicting, and limited by methodological flaws (for a more detailed discussion, see Chapter 3 and M. Martins et al., 2021).

The available evidence fits much better with the notion of **natural auditory sensitivity**, which could facilitate an inclination in individuals to pursue a musical career while also enhancing their voice perception skills. This idea is corroborated by research on two special groups of listeners: Individuals with amusia and “naturally good musicians”. Amusia has been consistently linked to deficits in vocal emotion perception (cf. Chapter 3), suggesting that auditory sensitivity to both music and voices may be mediated by an innate genetic factor. “Naturally good musicians” are people with excellent music perception abilities in the absence of formal musical training (Correia et al., 2022; Mankel & Bidelman, 2018). Indeed, vocal emotion capacities of these “naturally good musicians” equaled the performance of highly trained musicians (Correia et al., 2022). Our results in Chapter 6 fully support this finding: the link between auditory sensitivity in music and vocal emotion perception persisted in the absence of formal musical training, and when correlations were controlled for formal musical education. However, although there seems to be a high consensus for natural auditory sensitivity to be involved in vocal emotional skills, a fully conclusive picture may only be achieved through longitudinal studies in the future.

8.2.3. Electrophysiological correlates

The existing neuroscientific literature on the brain mechanisms underlying the musicality benefit for vocal emotions is sparse and inconclusive (see Chapter 3). To address this gap, I recorded ERPs of musicians and non-musicians and compared the P200, the N400 and the LPP. However, no group differences were observed in any of these components, nor did group modulate ERP effects of the F0 and Timbre manipulations. These findings seem somewhat at odds with the large body of literature showing a reliable benefit of musicality for vocal emotion perception (Chapter 3) and the clear behavioral pattern reported in Chapter 6 in particular. Instead, these results fit into a series of inconclusive electrophysiological findings (I. Martins et al., 2022; Pinheiro et al., 2015; Rigoulot et al., 2015). While there is evidence for ERP differences between musicians

and non-musicians – and even between classical and jazz musicians - for music and speech stimuli (Bianco et al., 2018; Chartrand et al., 2008; Pantev & Herholz, 2011), such differences fail to reliably extend to vocal emotions. In particular, I. Martins et al. (2022) found group differences in the P200, the P300 and the LPP in response to emotional music, but not to non-verbal emotional vocalizations, similar to the present null-findings for emotional prosody. Hence, although musicians show a behavioral benefit in emotion recognition performance, this does not seem to be reflected in ERP components associated with the processing of vocal emotional sounds.

However, some exploratory findings may imply that a musician’s brain responds differently to vocal emotions. There was a correlation between the P200 amplitude and music perception abilities, supporting the notion that the P200 can be an unspecific marker of auditory expertise (Chartrand et al., 2008). Further, extended analyses across all electrodes and time points revealed differences between musicians and non-musicians in later time ranges (> 600 ms). Comparison of groups in the ERP responses averaged across morph types revealed a difference for sadness only. Although this may seem surprising given the lack of such an emotion-specific pattern in the behavioral data, this pattern may be reminiscent of the one obtained in a published fMRI study (Park et al., 2015). Park et al. (2015) speculated that sadness may be of higher saliency to musicians compared to non-musicians, resulting in a unique neural representation (for a detailed discussion, refer to Chapter 7). In addition, an analysis exploring group differences in the processing of F0 vs timbre cues revealed a difference for happiness, also at a later time interval (> 800 ms), suggesting modulations related to the more controlled cognitive evaluation of the acoustic input. However, as this finding was unpredicted and exploratory, these claims need further empirical validation.

In light of the behavioral finding that the musicality benefit for vocal emotion perception seems to be driven by auditory sensitivity, one might have expected modulations in earlier ERP components associated with acoustic analysis and emotional integration (Schirmer & Kotz, 2006). The correlation of the P200 with music perception abilities fits this assumption. However, the present data further suggest differences in rather late and likely more elaborative processes. Perhaps, the differences in auditory sensitivity observed in the behavioral data are not necessarily shaped by early processes alone, but by later processes such as cognitive evaluation and conscious decision making. Musicians may be more acquainted with explicit emotional reasoning about auditory input, as it is part of their analytical work with music. This, in turn, could alter the way acoustic features are processed and represented in the brain, to facilitate the access for conscious decision making. After all, listening is not a passive process, but shaped by the way listeners filter and interpret the incoming sounds (Denham & Winkler, 2020). Musicians may profit here from a very fine-grained representation of subtle acoustic differences and more efficient decision making (Lima & Castro, 2011).

Taken together, the insight into vocal emotional processing in musicians and non-musicians provided by the present electrophysiological data is inconclusive and limited to exploratory and unpredicted findings. Nevertheless, this does not rule out that a musicians’ brain responds

differently to vocal emotions. Promising next steps may be addressing some of the questions raised by these findings or exploring electrophysiological correlates other than ERPs that have been proven valuable in research on effects of musical expertise, such as oscillatory responses (Mankel & Bidelman, 2018; Nolden et al., 2017; Strait et al., 2009).

8.2.4. Limitations and open questions

The present findings may have some limitations and raise open questions which could be addressed in future research. Please note that some broader open questions will be discussed in the general outlook in section 8.4.

When studying differences between musicians and non-musicians, the **comparability of groups** is always an important aspect. The recruited samples were carefully matched in age, distribution of sexes, socioeconomic background, and affective states (as measured by the PANAS). The latter is noteworthy because depressive symptoms can affect emotion recognition performance (Nilsson & Sundberg, 1985). Note that data collection took place during the COVID-19 pandemic, which might have put participants – and professional musicians in particular – at risk for mental health issues due to the precarious occupational situation. It is therefore important to note that none of the participants reported any current mood problems, and no group differences were found for affective measures. However, slightly higher levels of openness and neuroticism were observed in musicians compared to non-musicians. While a link between openness and musicality has been reported before (Corrigall et al., 2013; Schellenberg, 2016), research on other personality traits is sparse. Furthermore, the differential link observed between musicality and autistic traits seems worth exploring in more detail. While the overall AQ did not differ, musicians scored lower on the social communication domain, but higher on the attention to detail domain, compared with non-musicians. Clinical levels of autism have been frequently linked to insular talents such as musical aptitude and absolute pitch perception (Bonnell et al., 2003; Heaton et al., 1998; Wenhart & Altenmüller, 2019). However, it is unclear how autistic traits in the non-clinical spectrum affect musical experiences, and how this links to vocal emotion perception.

Another potential point of criticism could be that I only assessed formal education instead of an objective measurement of **cognitive abilities**. The relationship between musicality and general intelligence has been the focus of extensive research efforts (Schellenberg, 2001; Vincenzi et al., 2022). I omitted an objective measure because it would have prolonged the already extensive testing duration. Most of the participants of the non-musical control group were recruited at the university and either pursued or had completed a PhD. I therefore argue that the chance that the present findings were affected by a substantial difference in cognitive function to the disadvantage of the control group is very unlikely.

Finally, I targeted a population socialized in **Western music culture** and with a German language background in particular, to ensure that the pseudowords used in the study were not perceived as semantically meaningful. The role and complexity of pitch, harmonic and rhythmic

features can vary tremendously across different musical styles (Morrison & Demorest, 2009). Further, research using synthesized emotional voices suggests that the relative reliance on F0 and timbre cues for emotion inferences depends on a listeners' language background (Yanushevskaya et al., 2018). Hence, the present findings do not necessarily generalize across different cultures and therefore need to be replicated in more diverse samples.

8.2.5. Summary and conclusion: How does musicality affect the processing of emotional voice cues?

In this dissertation, I replicated the musicality benefit for vocal emotion perception in a sample comparing (semi-)professional musicians and non-musicians. Importantly, I showed that musicians are particularly tuned to the melody of vocal emotions - they outperformed non-musicians in the Full condition and when emotions were expressed by the pitch contour only, but not when they were expressed by vocal timbre. Further, the link between auditory sensitivity towards melodies and vocal emotional skills even persists in the absence of any musical training, suggesting a predisposition in individuals to exploit melodic patterns in both music and voices. In contrast to the clear behavioral results, the electrophysiological correlates were inconclusive, but several exploratory findings imply that musicians' brains may respond differently to vocal emotions. In summary, the present data offer original and strong evidence for transfer benefits from music to emotional voice perception, by highlighting auditory sensitivity as one of the driving factors.

8.3. Reflections on parameter-specific voice morphing

Another objective of this work was to assess the validity of parameter-specific voice morphing for the study of vocal emotion perception. A strength of this technique is that it permits the independent manipulation of specific voice cues, which allows insight into their functional role beyond correlational patterns. But despite its potential, it also bears two central caveats: First, the preparational work is technically very challenging and requires manual steps that can be time-consuming and error-prone. Second, the re-combination of different parameter weights for re-synthesis inevitable breaks the natural covariation of acoustic cues (Assmann & Katz, 2000), which is a property that can be seen as an advantage or a disadvantage, as it can lead to acoustic distortions. The preservation of acoustic quality in resynthesized voice morphs, however, is one of the key requirements stated by Kawahara and Skuk (2018) in order to produce ecologically valid stimuli. This problem may not be particularly relevant for a wide range of applications, and as long as parameters with moderate values are recombined, which results in sufficiently natural sounding voices. Parameter-specific voice morphing has therefore been successfully applied in studies on vocal sex or age (Pernet & Belin, 2012; Skuk et al., 2015, 2020). Vocal emotions, in contrast, can be characterized by rather extreme acoustic features, which is why their re-combination can substantially affect the perceived naturalness (i.e. human-likeness) of resulting voice morphs. However, as research on the "acoustic code" of vocal emotions could profit substantially from an employment of parameter-specific voice morphing (Arias et al., 2021), it is worthwhile to

pusue this technology, but to include a critical reflection on its validity with respect to voice naturalness. I addressed this problem from two angles: On the one hand, I explored different voice morphing protocols/approaches to improve stimulus quality. On the other hand, in the awareness that a certain degree of distortion is inherent to all approaches, I assessed the degree to which reduction of naturalness in the resulting voice morphs would disrupt vocal emotional processing. In what follows, I will first discuss the empirical findings on naturalness and its effect on emotional processing in parameter-specific voice morphs. Subsequently, I converted the experience I gained during my practical work with voice morphing into specific recommendations to foster a valid and successful employment of this technology in future research projects.

8.3.1. Perceived naturalness of parameter-specific voice morphs

The two experiments reported in Chapter 4 explicitly targeted the naturalness of emotional voice morphs, created with two different references: neutral voices and the average of emotional voices. The reference stimuli contributed the emotionally non-informative portion of the parameter-specific voice morphs. As expected from subjective listening impression, parameter-specific voice morphing affected perceived naturalness of the stimulus material. However, in the stimulus set with the neutral reference, naturalness was reduced in the Timbre compared to the F0 condition, creating a potential confound. Using averaged emotions as reference, naturalness of Timbre and F0 morphs was comparable, making them more suited for future research. This pattern suggests that although an impact on perceived naturalness in parameter-specific voice morphing may not be avoided completely, there are degrees of freedom one can explore to improve the stimulus quality.

The literature reviewed in Chapter 4 painted a very consistent picture on two acoustic features that affect the perceived naturalness in voices: fundamental frequency variation and the covariation between F0 and formant frequencies. Both features were also reflected in the present findings: **Fundamental frequency variation** was the strongest predictor of naturalness ratings in a regression analysis. Further, it was most likely the driving factor behind the improvement of the Timbre stimuli when averaged emotions were used as reference: In the Timbre condition, the emotional timbre is combined with the F0 contour of the non-emotional reference category. In this specific dataset, neutral voice quality was expressed by a very monotonous pitch, an impression which was confirmed by acoustic analysis, displayed in Table A.1 and A.2. The averaged emotions displayed a much greater F0 variance, which probably elevated the perceived naturalness of Timbre morphs.

The importance of **covariation of F0 and formants** could be probed with the parameter-specific voice morphing procedure itself: The re-combination of these cues from different voices breaks their natural covariation and inevitably results in a mismatch between fundamental frequency and formant frequencies in the F0 and the Timbre condition. Overall, these data clearly show that this re-combination affects perceived naturalness - but not in all cases: in F0 morphs with neutral reference, perceived naturalness was unaffected, as it was comparable to Full morphs. The specific circumstances under which the naturalness of voice morphs remains intact despite a

re-combination of F0 and timbre features is of particular interest, as it would allow to exploit the potentials of voice morphing without compromising ecological validity. It may be worthwhile to target this aspect in future research. Furthermore, the perception of naturalness relies on acoustic features which were particularly affected by **voice averaging**: despite its larger F0 variation, averaged voices were rated as far less natural than neutral ones. It should be kept in mind that voice averaging is in its infancy and currently very prone to artifacts such as reduced aperiodicity, higher harmonics-to-noise ratio, and temporal imprecision (Bruckert et al., 2010). Thus, both reference types had limitations. Ideal reference voices might have been original non-emotional audio recordings which nevertheless display sufficient fundamental frequency variation.

Taken together, findings from Chapter 4 strongly suggest that voice morphing can reduce perceived naturalness of the stimulus material and can pose a serious threat of confound if morphing conditions are affected differently. This highlights the importance of a conscious and critical reflection of the methodological choices, and advocates an explicit and objective validation of the stimulus material in an independent sample, instead of relying on the subjective listening impression of researchers only. Such rating data can provide arguments for or against the comparability of conditions and hence the validity of the voice morphing approach in general. Further, it allows to assess the interaction of naturalness with the actual research target. Therefore, the focus of the next section will be on the interplay of naturalness and emotional voice processing.

8.3.2. The role of stimulus naturalness for emotional voice processing

Humans display a strong tendency to infer emotions even in highly artificial settings and non-living objects (refer to Chapter 4 for a more detailed discussion). In line with this idea, the findings in Chapter 4 showed that perception of emotionality was remarkably robust against any distortion of voice naturalness. Similar results have been reported in the facial domain (Calder et al., 2000). The behavioral emotion perception performances reported in Chapter 5 and 6 offer additional insight: In Chapter 5, participants classified emotions morphed with the neutral reference, whereas in Chapter 6, the average reference was used. A direct comparison of the Figures 5.2 and 6.4 show that the behavioral performance was highly similar for both stimulus sets (with the exception of pleasure in the F0 condition; for a detailed discussion of this finding, please refer to section 8.1). This similarity is remarkable in light of the profound differences in naturalness I observed between conditions, supporting the notion that emotional processing can suppress disruptive effects of unnatural and artificial features to the degree that behavioral measures are largely unaffected.

However, electrophysiological data paint a different picture. In the EEG study reported in Chapter 5, parameter-specific voice morphs created with the neutral reference were associated with consistent modulations of the P200 and the N400. In fact, these modulations were reminiscent of the behavioral performance differences in size and direction. However, in a second EEG study using voice morphs created with the average reference (reported in Chapter 7), these patterns did not fully replicate. Across all emotions, ERP difference between the F0 and Timbre conditions were much smaller and less consistent with the behavioral data. It is

possible that the ERP differences between F0 and Timbre conditions with neutral reference observed in Chapter 5 are the result of the strong confound with naturalness, instead of emotional processes per se; whereas using the dataset with averaged reference, in which this confound was eliminated, resulted in much smaller ERP modulations. Thus, one may speculate that even though reduced naturalness does not affect behavioral outcomes, it still leaves a “neural mark”. On the one hand, reduced human-likeness of these stimuli may disrupt the recruitment of neural networks which predominantly respond to human voices (Belin et al., 2011). On the other hand, listeners’ limited experience with these stimuli could affect their brain responses (Schirmer et al., 2012). In line with these ideas, ERP effects related to the human-likeness of the stimulus material have been reported for both faces and voices (Schindler et al., 2017; Schirmer & Gunter, 2017).

It has to be kept in mind, however, that this explanation is purely speculative at this point, as other key differences between the studies could be responsible for the diverging outcomes as well (for an in-depths comparison, refer to Chapter 7). First, the stimulus sets probably differed with regard to their acoustic features beyond naturalness. Second, the recruited samples differed slightly regarding age and professional background. Third, although both studies guided attention towards the emotional features of the sound, participants performed different tasks. Finally, links between behavioral data and the second EEG study should be made with caution, as they were recorded in two different sessions. Therefore, a systematic investigation of ERP modulations related to naturalness in emotional voice morphs remains pending. Nevertheless, these data highlight the importance of considering possible confounds, especially in designs with electrophysiological outcomes, as the effects could be far more detrimental than in behavioral paradigms.

8.3.3. Emotional voice morphing – practical recommendations and future applications

The present findings advocate parameter-specific voice morphing as a suitable tool to study the processing of vocal emotions, but only if conducted appropriately. For future research projects, the following recommendations based on my personal experience as well as the empirical data may guide through three important steps of the stimulus preparation: the recording/selection of original voices, the morphing pipeline in Tandem-STRAIGHT (Kawahara et al., 2008) itself, and the assessment of the resulting acoustic quality.

First, during the **recording/selection of original voice material**, it can be recommended to ensure substantial F0 variation in the stimuli. In vocal emotional research, this ties back to a bigger question, namely what a “neutral” voice sounds like. From personal experience, speakers usually come with a clear mental depiction of e.g. an angry or a happy voice, but are sometimes confused when they are instructed to use a “neutral voice”, which subsequently prompts them to use an overly monotonous tone. Considering the present findings, this should be avoided. Stimuli which lack sufficient quality from the start, like the vocal averages used here, are insufficient as well. Ideally, one would obtain original non-emotional voices with sufficient F0 variation.

Subsequently, **voice-morphing using Tandem-STRAIGHT** requires conscientiousness and a critical mind of the researcher. The recommendations described by Kawahara and Skuk (2018) offer a good starting point. As the preprocessing of voices requires some manual steps, detailed documentation is crucial to make this process as replicable and transparent as possible. For critical choices, such as the removal of artifacts, it may be beneficial to employ two-person protocols. Finally, it is strongly recommended to provide an independent **measure of acoustic quality** other than subjective impression of the researcher. It has to be kept in mind that the evaluation of naturalness does not only depend on aspects of the voices but also of the listeners. Researchers who are exposed to their stimuli on a daily basis may develop perceptual biases after prolonged adaptation, potentially clouding their judgement with respect to stimulus quality (Kloth et al., 2017; Webster & MacLin, 1999).

In the future, it may be worthwhile to explore parameter-specific voice morphing in vocal emotions with **different intensities**. As discussed in section 2.1, actor portrayals may display exaggerated emotions with extreme acoustic features. When these features are re-combined, extreme values could affect naturalness to a larger degree than moderate ones. Recordings of induced emotions are assumed to be less intense and an ecologically more valid depiction of emotional expressions in real life (Scherer, 2003). Therefore, they may be the more suitable starting material. Once ecological validity of stimuli is ensured, the acoustic flexibility of voice morphing opens the door to several **future applications** beyond basic research. For example, it allows the development of tailored training protocols. The emotional information expressed by a specific parameter could be flexibly adjusted to train individuals with specific deficits. In amusia, a specific impairment in pitch perception is frequently discussed as the potential reason for emotion perception difficulties (Lima et al., 2016; Thompson et al., 2012), but that does not seem to hold for every individual (Lagrois & Peretz, 2019). After obtaining an idiosyncratic profile of auditory abilities, parameter-specific voice morphing could be used to create effective training material targeting individual needs, fine-tuned to an appropriate level of difficulty. The emotional information expressed by specific cues could be flexibly adjusted and even caricatured to make them more salient. The key disadvantage, however, lies in the extensive manual preprocessing, limiting vocal material to pre-recorded stimuli. Furthermore, voice morphing in its current form does not allow real-time voice manipulation. Real-time applications that perform an online exaggeration of emotional acoustic information could help to improve the daily life of individuals with sensory deficits. However, online caricaturing of all acoustic features in an emotional utterance is technically very challenging. As a starting point, one could target F0 contour only, which was the dominant parameter for emotion perception in the present work. Online-tracking of the F0 contour is already a standard feature in a range of software applications (e.g. in any standard music tuning app, such as Soundcorset, <https://soundcorset.com/>), and therefore its real-time caricaturing may become feasible in the near future.

8.3.4. Summary and conclusion: Is parameter-specific voice morphing a suitable tool to study the processing of vocal emotions?

In this dissertation, I gathered convincing empirical evidence showing that parameter-specific voice morphing is a suitable tool to study the processing of vocal emotions. Crucially, this technology allows causal insight into the role of different acoustic cues, making a valuable contribution to the predominantly correlational literature trying to uncover the acoustic code of vocal emotions. However, I also found that voice morphing can affect the perceived naturalness of the resulting stimulus material, which could be a threat to ecological validity. While behavioral measures of emotion perception were remarkably robust against any distortions of voice naturalness, electrophysiological correlates could be critically affected. Therefore, great care should always be dedicated to the acoustic quality of morphs, ideally by providing objective measures and a critical reflection of the consequences of methodological choices. If conducted appropriately, parameter-specific voice morphing can be an extremely powerful tool, with the potential to open many new doors to the understanding of vocal emotion perception.

8.4. Directions for future research

Several open questions and potential limitations discussed in this dissertation offer exciting directions for future research.

As the present work specifically focused on F0 and timbre, stimuli were controlled for the potential influence of other cues. Therefore, an open question concerns the contribution of **amplitude (i.e. loudness)** and **temporal characteristics** to emotional inferences. While previous acoustic analyses suggest a central role of amplitude and timing (Juslin & Laukka, 2003), evidence using acoustically manipulated material is sparse. However, Ilie and Thompson (2011) reported an impact of amplitude and speech rate manipulation on ratings of valence and arousal. Further, Chen et al. (2012) found that sound amplitude modification had a significant impact on the processing of angry voices, both on the behavioral and the electrophysiological level. Based on synthesized voices, Yanushevskaya et al. (2013) argued that while loudness alone is relatively ineffective for emotional signaling, it seems to unfold some potential in appropriate combination with other vocal cues. The present data cannot speak to the impact of loudness, since all stimuli were amplitude-normalized. However, they make an indirect case for the temporal unfolding of cues. Both F0 and timbre information change over time and this dynamic aspect seems to play an important role. On the one hand, F0 variation appeared to be a strong predictor of emotionality ratings (Chapter 4). On the other hand, emotion recognition skills were consistently correlated with rhythm perception skills in musical stimuli (Chapter 6). In fact, musical skills, which rely on an integration of acoustic features over time, were the only ones which were predictive of vocal emotion performance, while the ones which are based on an evaluation of auditory snapshots were not. Taken together, while these findings hint at a major role of temporal cues in emotional processing, a more systematic investigation remains pending.

Considering the perceptual flexibility among listeners, one possibility is to try to pinpoint **systematic differences between individuals**. The present work makes an initial contribution to such an approach, by comparing musicians and non-musicians. Additionally, Yanushevskaya et al. (2018) suggested that the relative weighting of F0 vs. timbre cues depends on the individuals' language background. Finally, Schneider, Sluming, Roberts, Scherg et al. (2005) provided evidence that individuals can be grouped as "fundamental pitch listeners" and "spectral listeners", with a strong neural basis in the primary auditory cortex (Heschl's gyrus) for this distinction. In the future, it would be worth investigating how these perceptual modes relate to the processing of F0 and timbre in emotional voices.

Following up on the previous point, it should be acknowledged that the simple categorization of individuals as "musicians" does **not make them a homogenous group**. In fact, there are not only different levels of musical expertise and diverse cultures, but there is also great variety in terms of genres, styles, professions, and forms of expression, within the scope of the Western music system and beyond. Therefore, treating musicians as a single group provides an insufficient representation of this heterogeneity. Indeed, there is ample evidence for instrument-specific neural modulations in individuals (Chartrand et al., 2008; Kraus & Chandrasekaran, 2010; Pantev et al., 2001), and different neurocognitive profiles related to different forms of musical expertise (Schneider, Sluming, Roberts, Bleck & Rupp, 2005; Tervaniemi, 2009). In the context of vocal emotion perception, this variability could be addressed by taking a closer look at several subgroups of musicians.

A particularly interesting comparison is the one between **singers and instrumentalists**. As singing provides the form of musical expression that is most closely related to vocal emotions, it could be assumed that singing fosters vocal emotion perception abilities to a larger degree than instrumental activities. However, the empirical evidence is inconclusive. In the present data, self-rated singing abilities were correlated with vocal emotion recognition performance, similar to Correia et al. (2022), although both findings were exploratory. One step further, Greenspon and Montanaro (2023) provided evidence that objective measures of singing ability predict vocal emotion perception. In contrast, in a music-intervention study, Thompson et al. (2004) observed that singing lessons may actually interfere with vocal emotional processing, whereas keyboard lessons had a positive effect. Another recent study observed similar brain responses to vocal emotions in singers and instrumentalists (I. Martins et al., 2022). Thus, at this point, the available literature on the comparison of singers and instrumentalists is inconsistent and requires further systematic investigation.

Furthermore, it is noteworthy that the present investigations targeted **professional/semi-professional musicians**, and therefore explicitly excluded **musical amateurs**. This was done to ensure a maximum difference between the groups of musicians and non-musicians in terms of musical expertise. However, there is accumulating evidence that differences between professionals and amateurs are not only of quantitative, but of qualitative nature. For example, recent findings suggest that professional musicians and non-musicians are comparable in terms of general

intelligence, whereas non-professional musical amateurs have higher scores (Vincenzi et al., 2022). Additionally, musical engagement as leisure activity was found to have a larger protective effect on “brain aging” than musical engagement as a professional (Rogenmoser et al., 2018). This also seems to be reflected in general health, which was found to be better in amateurs than professionals (Bonde et al., 2018). Amateurs may profit from musical activity to a larger degree, because it provides an enrichment to their professional occupation, and they may experience less pressure during music performances. Further, in the context of the COVID-19 pandemic, amateurs may have been less threatened by precarious occupational circumstances. It is unclear, however, how amateurs and professionals may diverge with regard to vocal emotion perception. Further, it could be interesting to explore whether the underlying mechanisms, which were identified for professional musicians – auditory sensitivity and pitch processing in particular –, would also hold for musical amateurs.

Despite accumulating evidence, we are far away from understanding the link between musicality and vocal emotional processing to the full extent, and many questions remain open: With pitch processing playing an important role, what could be the contribution of a musical skill called absolute pitch? Does the musicality benefit for emotional processing expand to vocal emotions expressed by speakers from other cultures and languages? What is the role of active maintenance of musical skills? And finally, given the tight connection between vocal perception and production (A. W. Young et al., 2020), could musicians be also more proficient in the expression of vocal emotions?

In addition to all these open questions regarding musicality, I also acknowledge that this dissertation only scratched the surface in our efforts to understand the role of **voice naturalness** for emotional processing and beyond. On an empirical level, the present dissertation clearly illustrates the demand for more systematic and large-scale investigations on the behavioral outcomes, but especially the neural correlates of voice naturalness. To this end, it may be worthwhile to explore different ways to de-confound naturalness from other vocal characteristics. That is, the goal is not only to vary information such as emotionality without effects on naturalness, but also to vary the degree of naturalness without effects on other characteristics of the voice. On the conceptual level, it should be noted that naturalness is a multi-faceted concept, which could profit from precise definitions and clear-cut differentiation in the literature. For example, voice naturalness may be defined as “human-like”, as it was done in the present work, but it could also be understood as “odd” or “rare”. As these conceptualizations could have different empirical implications, they should be reflected on and made explicit, not only to the scientific community but also to potential participants and raters in experimental instructions. In the future, well-conducted research on the impact of voice naturalness could make an important contribution in a world that is more and more relying on interaction with non-human agents, making it very important to understand the interplay of emotion perception and artificial voice features.

8.5. Summary and conclusion

In this dissertation, I addressed three main gaps in the field of vocal emotion perception.

First, I quantified the relative contribution of F0 and timbre cues to the perception of different emotions and explored their electrophysiological correlates. To this end, I used acoustically manipulated voices by means of parameter-specific voice morphing. My empirical findings make several valuable contributions to the existing body of literature: I provided causal evidence that both F0 and timbre carry unique information that allow emotional inferences, although F0 seems to be the more dominant parameter overall. Their precise contribution, however, varies as a function of emotional category. In emotions with high intensity, the dominance of F0 is stronger, whereas in emotions with lower intensity, the roles of F0 and timbre are more balanced. The electrophysiological data revealed F0- and timbre-specific modulations in several ERP components, such as the P200 and the N400.

Second, I explored how musicality affects the processing of emotional voice cues. I started by providing a review on the literature linking musicality to emotion perception and subsequently showed that musicians have a benefit in vocal emotion perception compared to non-musicians, which seems to be rooted in auditory sensitivity. The present data not only replicated previous findings but offer original insight into the special role of pitch cues: musicians outperformed non-musicians when emotions were expressed by the pitch contour only, but not when they were expressed by vocal timbre. Thus, musicians seem to be particularly tuned to the melody of vocal emotions. In addition, I found that the link between auditory sensitivity and vocal emotion perception performance persisted in the absence of formal musical training, suggesting a strong role of a predisposition in individuals to exploit melodic patterns in both music and voices. Although the electrophysiological patterns were less conclusive, they imply that musicality may modulate brain responses to vocal emotions. In summary, this dissertation provides strong evidence for transfer benefits from musicality to vocal emotion perception, highlighting auditory sensitivity and pitch perception in particular as important underlying mechanisms.

Third, I critically reflected whether parameter-specific voice morphing would qualify as a valid tool to study the processing of vocal emotions. I identified distortions in voice naturalness resulting from extreme acoustic manipulations as one of the major threats to the ecological validity of the stimulus material produced with this technique. To address this problem, I gathered explicit data on the perception of naturalness and assessed its impact on emotion perception. The results suggested that while voice morphing does affect the perceived naturalness of stimuli, behavioral measures of emotion perception were found to be remarkably robust against these distortions. However, unnatural voice features could affect electrophysiological correlates. The present work provides convincing evidence that parameter-specific voice morphing is a valid tool for vocal emotional research. At the same time, it only offers a starting point in understanding naturalness as a concept in person perception, which I hope will prompt future initiatives striving for a more systematic understanding, both conceptually and empirically.

In summary, this dissertation expands the conceptual and empirical knowledge of vocal emotion processing, regarding the underlying acoustic parameters, electrophysiological correlates, and individual differences with a specific focus on musicality. This way, it contributes to the understanding of several “common everyday” and yet extraordinary qualities in humans: the use of our voices, the ability to express and perceive emotions, and the capacity to make music.

A. Supplemental Tables

Table A.1.: Summary of the acoustic characteristics of female voice morphs separately for each Emotion, Morph Type and Reference Type

	Morphs	Ref	F0 _{Mean}	F0 _{SD}	F0 _{Glide}	FormDisp	HNR	Dur
<i>Happiness</i>								
	Full	AVG	348	98	-112	993	19	799
	Full	NEU	343	94	-104	1002	18	696
	F0	AVG	348	98	-112	1096	20	799
	F0	NEU	343	94	-104	1057	19	696
	Timbre	AVG	247	25	-37	981	19	799
	Timbre	NEU	197	11	1	982	18	695
<i>Pleasure</i>								
	Full	AVG	185	21	-32	1131	19	799
	Full	NEU	184	20	-31	1131	19	696
	F0	AVG	185	21	-32	1094	19	799
	F0	NEU	184	20	-31	1053	17	696
	Timbre	AVG	247	25	-37	1122	20	799
	Timbre	NEU	197	11	1	1124	20	695
<i>Fear</i>								
	Full	AVG	288	30	28	1112	21	800
	Full	NEU	284	30	30	1109	20	696
	F0	AVG	288	30	28	1093	21	800
	F0	NEU	284	30	30	1054	18	696
	Timbre	AVG	247	25	-37	1120	21	799
	Timbre	NEU	197	11	1	1109	20	695
<i>Sadness</i>								
	Full	AVG	219	19	-39	1090	22	799
	Full	NEU	222	20	-29	1090	22	696
	F0	AVG	219	19	-39	1097	21	799
	F0	NEU	222	20	-29	1053	20	696
	Timbre	AVG	247	25	-37	1085	22	799
	Timbre	NEU	197	11	1	1086	21	695
<i>Average</i>	Full	AVG	247	25	-39	1094	22	799
<i>Neutral</i>	Full	NEU	197	11	0	1054	21	695

Table A.2.: Summary of the acoustic characteristics of male voice morphs separately for each Emotion, Morph Type and Reference Type

	Morphs	Ref	F0 _{Mean}	F0 _{SD}	F0 _{Glide}	FormDisp	HNR	Dur
<i>Happiness</i>								
	Full	AVG	259	89	-74	999	17	762
	Full	NEU	256	88	-86	990	16	644
	F0	AVG	259	89	-74	1037	17	762
	F0	NEU	256	88	-86	971	16	644
	Timbre	AVG	158	21	-43	985	15	762
	Timbre	NEU	110	4	0	988	14	643
<i>Pleasure</i>								
	Full	AVG	121	18	-32	1064	14	763
	Full	NEU	122	18	-38	1067	14	646
	F0	AVG	121	18	-32	1046	15	763
	F0	NEU	122	18	-38	976	14	646
	Timbre	AVG	158	21	-43	1058	14	762
	Timbre	NEU	110	4	0	1068	14	643
<i>Fear</i>								
	Full	AVG	191	23	-19	1077	17	764
	Full	NEU	189	23	-25	1066	16	645
	F0	AVG	191	23	-19	1046	17	764
	F0	NEU	189	23	-25	972	15	645
	Timbre	AVG	158	21	-43	1074	17	762
	Timbre	NEU	110	4	0	1070	15	643
<i>Sadness</i>								
	Full	AVG	122	14	-47	1040	16	763
	Full	NEU	124	15	-40	1040	15	642
	F0	AVG	122	14	-47	1049	16	763
	F0	NEU	124	15	-40	969	14	642
	Timbre	AVG	158	21	-43	1033	16	762
	Timbre	NEU	110	4	0	1041	15	643
<i>Average</i>	Full	AVG	158	21	-43	1047	17	762
<i>Neutral</i>	Full	NEU	110	4	0	971	17	643

Note. All acoustical parameters of Table A.1 and Table A.2 were adapted from McAleer et al. (2014) and extracted using PRAAT software (Boersma, 2018) and MATLAB (2020). $F0_{Glide} = F0_{End} - F0_{Start}$; Formant Dispersion (FormDisp): ratio between consecutive formant means (F1 to F4); HNR (harmonics-to-noise ratio) was extracted with the cross-correlation method in Praat. AVG/NEU: average/neutral reference, Dur: Duration in ms.

Table A.3.: Summary of key mappings to (a) emotions and (b) pseudowords. Participants were assigned to key mappings via their participation number

(a)	“s”	“d”	“k”	“l”
1	happiness	pleasure	sadness	fear
2	sadness	fear	happiness	pleasure
3	pleasure	happiness	fear	sadness
4	fear	sadness	pleasure	happiness
(b)	“s”	“d”	“k”	“l”
pseudoword	/belam/	/molen/	/namil/	/loman/

Note. Participants were instructed explicitly to press the keys “s” and “d” with their left index- and middle-finger and the keys “k” and “l” with their right index- and middle-finger.

Table A.4.: Descriptive data of questionnaires

	M	SD	Min	Max
AQ				
Total	14.90	4.92	7	25
Attention To Detail	5.00	1.95	2	9
Social	9.90	4.05	4	21
BFI				
Extraversion	7.23	1.80	4	10
Agreeableness	6.67	1.64	3	10
Conscientiousness	6.97	1.63	3	10
Neuroticism	6.31	2.12	2	10
Openness	8.26	1.77	3	10
Gold-MSI				
Global Score	2.29	0.62	1.22	3.71
Active Engagement	3.85	1.16	1.67	6.00
Musical Training	3.32	1.40	1.00	5.71
Emotions	5.80	0.70	3.83	7.00
Perceptual Abilities	5.13	1.06	2.67	6.67
Singing Abilities	3.93	1.24	1.57	6.57

Note. For further information and interpretation of the values, please refer to Baron-Cohen et al. (2001) and Hoekstra et al. (2008) for the Autism Quotient Questionnaire (AQ); to Rammstedt and John (2007) for the 10-item personality inventory measuring the Big-Five domains (BFI); and to Müllensiefen et al. (2014) for the Goldsmiths Musical Sophistication Index (Gold-MSI).

Table A.5.: Pearson correlations between questionnaire data and vocal emotion recognition performance and for each Morph Type separately

	Vocal Emotion Recognition			
	Averaged	Full	F0	Tbr
<i>AQ</i>				
Total Score	-0.12 (.481)	-0.06 (.699)	-0.23 (.156)	0.02 (.911)
AttentionToDetail	-0.07 (.678)	-0.08 (.630)	-0.19 (.250)	0.13 (.417)
Social	-0.11 (.512)	-0.04 (.811)	-0.19 (.244)	-0.04 (.801)
<i>BFI</i>				
Extraversion	0.07 (.670)	0.00 (.993)	0.13 (.422)	0.06 (.734)
Agreeableness	0.44 (.005)	0.47 (.003)	0.32 (.049)	0.25 (.119)
Conscientiousness	-0.21 (.189)	-0.19 (.235)	0.11 (.520)	-0.47 (.003)
Neuroticism	0.12 (.469)	0.12 (.451)	-0.01 (.928)	0.19 (.256)
Openness	0.20 (.215)	0.16 (.340)	0.14 (.407)	0.22 (.182)
<i>Gold-MSI</i>				
Global Score	-0.03 (.870)	-0.09 (.588)	0.10 (.560)	-0.06 (.733)
Active Engagement	-0.11 (.494)	-0.21 (.196)	-0.09 (.593)	0.08 (.623)
Musical Training	-0.05 (.783)	-0.07 (.655)	0.09 (.606)	-0.12 (.455)
Emotions	0.08 (.619)	-0.01 (.960)	0.05 (.772)	0.20 (.214)
Perceptual Abilities	0.06 (.730)	-0.02 (.910)	0.17 (.310)	0.01 (.952)
Singing Abilities	0.09 (.600)	0.09 (.610)	0.19 (.250)	-0.08 (.629)

Note. "Averaged" = averaged data across all Morph Types. *p*-values (two tailed) in parentheses.

Table A.6.: List of reported instruments by musicians and non-musicians

Musicians		Non-Musicians	
Klavier (piano)	11	Gesang (singing)	5
Orgel (organ)	7 6	Klavier (piano)	3
Gesang (singing)	7	Violine (violin)	2
Gitarre (guitar)	3 4	Gitarre (guitar)	1
Violine (violin)	2	Blockflöte (flute)	1
Klarinette (clarinet)	2	Tamburin (tambourine)	1
Violoncello (cello)	1	Fürst-Pless-Horn (horn)	1 0
Kontrabass (double bass)	1		
Blockflöte (flute)	1		
Oboe (oboe)	1		
Querflöte (flute)	1		
Trompete (trumpet)	1		
Dudelsack (bagpipe)	1		

Note. These data include both samples from Chapter 6 and Chapter 7, which only differ with regard to very few participants. If two numbers are reported, the first corresponds to Chapter 6 and the second corresponds to Chapter 7.

Table A.7.: Socioeconomic background of participants

Income (€)			Education		Degree					
	M	C		M	C		M	C		
<1750	13	7	keine (none)	0	0	keine (none)	1	0		
1750-2500	6 7	8	Schüler (pupil)	0	0	Schüler (pupil)	0	0		
2500-3500	9 8	15	Hauptschule (secondary school)	0	0	inAusbildung (under training)	6 5	6 5		
3500-5000	5	6 7	Mittelschule secondary school)	1	0	Lehre (traineeship)	2	1		
>5000	6	2	Fachschule (technical college)	0	4	Fachschule (technical college)	1	4		
			Abitur (A-levels)			38	34 35	Meister (master as craftsmen)	0	1
								Bachelor (Bachelor)	8 9	4 6
								Fachhochschule (polytechnic degree)	4	2
								Master/Diplom (Master/Diploma)	14	16
								Promotion (PhD)	3	4
Sample Chapter 6:										
$\chi^2(4) = 5.66, p = 0.226$			$\chi^2(2) = 5.21, p = 0.074$			$\chi^2(8) = 6.40, p = 0.603$				
Sample Chapter 7:										
$\chi^2(4) = 6.33, p = 0.176$			$\chi^2(2) = 5.12, p = 0.077$			$\chi^2(8) = 5.68, p = 0.683$				

Note. This table presents the number of individuals belonging to different income, education, and degree categories. These data include both samples from Chapter 6 and Chapter 7, which only differ with regard to very few participants. If two numbers are reported, the first corresponds to Chapter 6 and the second corresponds to Chapter 7. We tested group differences between musicians (M) and non-musicians (C) using a Chi-square test and show the results in the last line of this table. Please note that the response options “Education” (i.e. the type of school) and “Degree” (i.e. the highest professional qualification) were tailored to the German educational system and are therefore difficult to translate. Further, please note that “Fachschule” and “Abitur” are similar as they both enable a person to pursue a university degree (with a few more constraints for a “Fachschule” degree). We therefore consider the trend observed for the “Education” factor merely as an artefact of the response format. M = Musicians, C = Controls/Non-Musicians.

Table A.8.: Characteristics of participants - Demography, personality, and musicality (EEG Study - Chapter 7)

	Musicians	Non-Musicians					
	M (SD)	M (SD)	t	df*	p	Cohens d	
Age	29.9 (5.5)	30.5 (6.3)	-0.46	74.47	.645	-0.11 [-0.56, 0.35]	
<i>PANAS</i>							
positive Affect	3.36 (0.60)	3.08 (0.66)	1.99	75.42	.051	0.46 [0.00, 0.91]	t
negative Affect	1.71 (0.46)	1.52 (0.70)	1.39	66.06	.168	0.34 [-0.14, 0.83]	
<i>Big Five</i>							
Openness	4.15 (0.47)	3.77 (0.78)	2.55	62.63	.013	0.65 [0.13, 1.15]	*
Conscientiousness	3.49 (0.75)	3.68 (0.76)	-1.10	75.98	.272	-0.25 [-0.70, 0.20]	
Extraversion	3.48 (0.66)	3.34 (0.74)	0.85	74.66	.397	0.20 [-0.26, 0.65]	
Agreeableness	3.97 (0.58)	3.78 (0.66)	1.33	74.86	.187	0.31 [-0.15, 0.76]	
Neuroticism	2.90 (0.63)	2.64 (0.84)	1.54	70.61	.127	0.37 [-0.10, 0.85]	
<i>AQ</i>							
Total	15.64 (4.88)	17.69 (6.36)	-1.60	71.23	.115	-0.38 [-0.85, 0.09]	
Attention to Detail	5.46 (2.01)	4.35 (2.05)	2.39	75.96	.019	0.55 [0.09, 1.01]	*
Social	10.17 (4.65)	13.33 (6.49)	-2.47	68.90	.016	-0.59 [-1.08, -0.11]	*
Social Skills	1.41 (1.65)	2.56 (2.60)	-2.34	64.28	.023	-0.58 [-1.08, -0.08]	*
Communication	1.87 (1.64)	2.51 (1.74)	-1.67	75.72	.100	-0.38 [-0.84, 0.07]	
Imagination	2.10 (1.52)	2.82 (1.94)	-1.81	71.77	.073	-0.43 [-0.90, 0.04]	t
Attention Switching	4.79 (1.90)	5.43 (2.06)	-1.43	75.68	.156	-0.33 [-0.78, 0.12]	
<i>Gold-MSI</i>							
General ME	5.65 (0.51)	2.71 (1.02)	16.00	55.77	<.001	4.28 [3.33, 5.23]	***
Active Engagement	4.94 (0.78)	2.98 (1.20)	8.56	65.51	<.001	2.11 [1.50, 2.71]	***
Formal Education	5.90 (0.57)	1.67 (0.63)	31.14	75.02	<.001	7.19 [5.75, 8.42]	***
Emotion	5.86 (0.70)	4.94 (1.31)	3.86	58.58	<.001	1.01 [0.46, 1.55]	***
Singing	5.29 (0.84)	2.82 (1.22)	10.44	67.47	<.001	2.54 [1.90, 3.18]	***
Perception	6.30 (0.50)	4.20 (1.45)	8.56	46.98	<.001	2.50 [1.73, 3.25]	***
<i>PROMS</i>							
Pitch	0.27 (0.07)	0.18 (0.07)	5.80	75.92	<.001	1.33 [0.83, 1.82]	***
Melody	0.23 (0.07)	0.07 (0.07)	9.98	75.98	<.001	2.29 [1.71, 2.86]	***
Timbre	0.32 (0.08)	0.26 (0.09)	3.03	74.29	.003	0.70 [0.23, 1.17]	**
Rhythm	0.32 (0.08)	0.26 (0.07)	3.54	75.98	.001	0.81 [0.34, 1.28]	**

Note. Descriptive values show mean ratings for the PANAS (Breyer & Bluemke, 2016), the Big-Five Domains (Rammstedt et al., 2018), and the Gold-MSI (Müllensiefen et al., 2014). AQ scores were calculated based on Hoekstra et al. (2008) and Baron-Cohen et al. (2001).

^a Note that original degrees of freedom were 76 but were corrected due to unequal variance.

Table A.9.: Summary of response key mappings to emotions

	“d”	“f”	“j”	“k”
CB 1	happiness	pleasure	sadness	fear
CB 2	sadness	fear	happiness	pleasure
CB 3	pleasure	happiness	fear	sadness
CB 4	fear	sadness	pleasure	happiness

Note. Participants were instructed explicitly to press the keys “d” and “f” with their left index- and middle-finger and the keys “j” and “k” with their right index- and middle-finger. CB = counterbalancing condition. Response mappings were identical for each participant in Chapter 6 and Chapter 7.

Table A.10.: Participant assignment to the different response key mappings

	Musicians	Non-Musicians
CB 1	10	7
CB 2	10 11	7
CB 3	8 7	15 16
CB 4	11	9

Note. Participants were randomly assigned to key mappings in the online study (Chapter 6) and later received the same mapping in their EEG session (Chapter 7). These data include both samples from Chapter 6 and Chapter 7, which only differ with regard to very few participants. If two numbers are reported, the first corresponds to Chapter 6 and the second corresponds to Chapter 7. CB = counterbalancing condition.

B. Supplemental Figures

Figure B.1.: Channel locations of the 64-channel setup used for the EEG experiments reported in Chapter 5 and 7

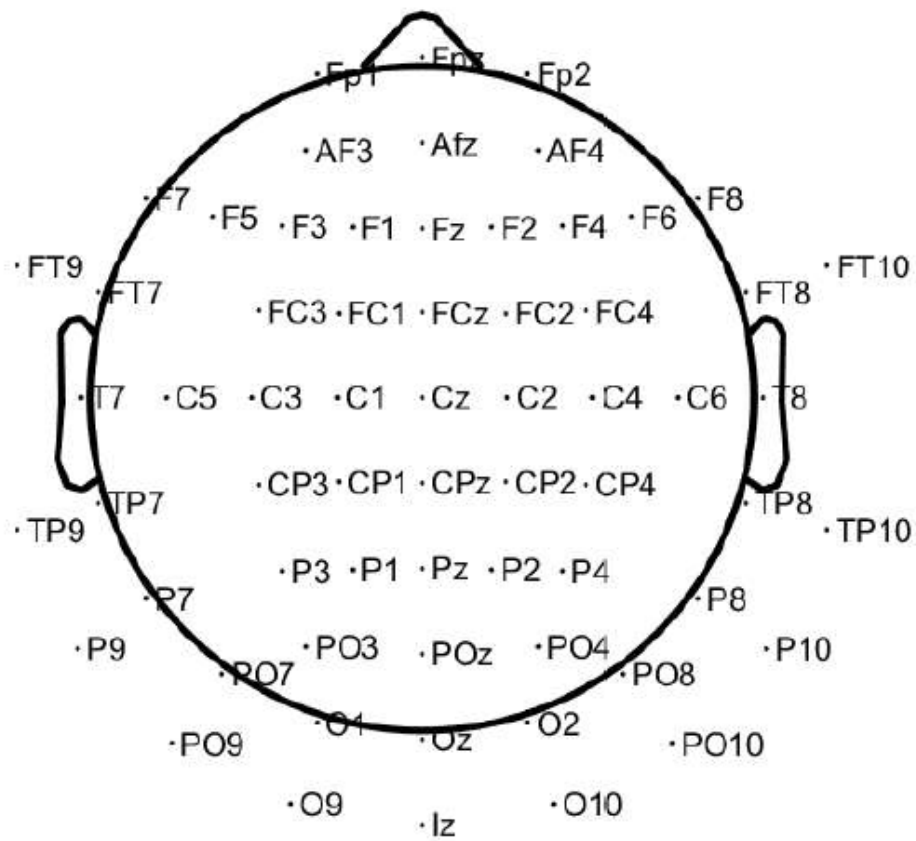
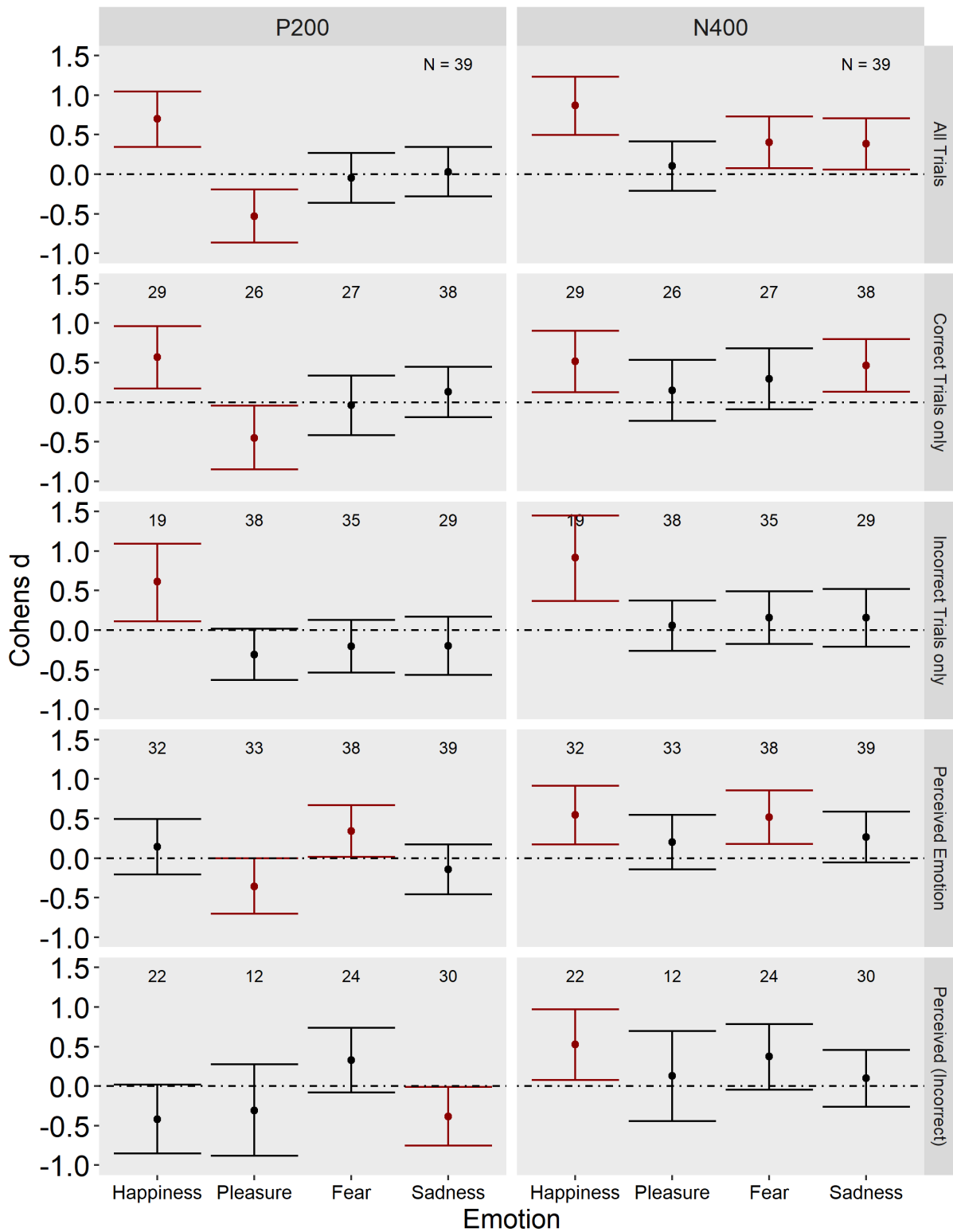


Figure B.2.: Effect sizes of the F0 vs. Timbre contrast for the P200 and N400 in different subsets of trials



Note. Data from the experiment reported in Chapter 5. Perceived Emotion = emotions grouped according to the answer of participants (e.g. all stimuli that were perceived as expressing happiness); Perceived incorrect = incorrect trials grouped according to the answer of participants; N = number of participants that contributed data (e.g. not all participants had correct Timbre and F0 trials, to calculate the difference).

Figure B.3.: Confusion matrices for each emotion for the three morph types – musicians only

		Full							F0						Timbre			
Classification Proportion in %	Sad	1	12	18	75	40	Sad	7	28	26	66	Sad	20	28	35	48		
	Fea	2	5	68	15	18	Fea	6	10	59	15	Fea	16	15	32	21		
	Ple	2	70	4	8	27	Ple	7	49	7	16	Ple	18	38	18	21		
	Hap	95	13	9	2	15	Hap	79	14	8	3	Hap	46	18	14	10		
		Hap	Ple	Fea	Sad	Avg		Hap	Ple	Fea	Sad		Hap	Ple	Fea	Sad		
									Emotion									

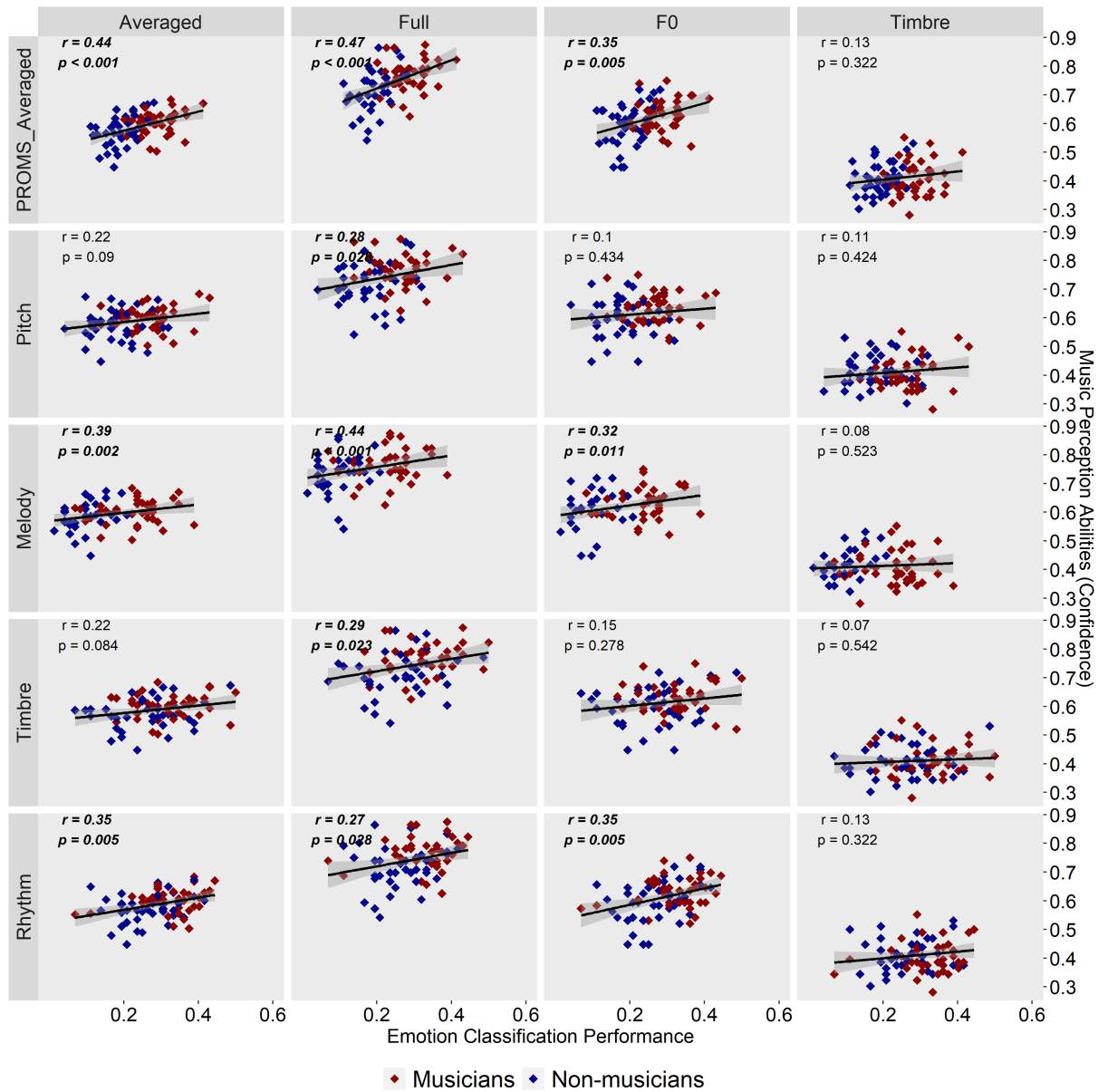
Note. Numbers represent the proportion of classification responses per Emotion and Morph Type, averaged across musicians. Hap = happiness, Ple = pleasure, Fea = fear, Sad = sadness, Avg = average.

Figure B.4.: Confusion matrices for each emotion for the three morph types – non-musicians only

		Full							F0						Timbre			
Classification Proportion in %	Sad	2	13	19	67	36	Sad	5	25	28	62	Sad	17	27	30	46		
	Fea	3	4	62	17	17	Fea	6	10	50	15	Fea	17	13	31	19		
	Ple	3	66	8	10	26	Ple	7	45	9	13	Ple	15	37	20	21		
	Hap	92	16	12	5	21	Hap	82	20	13	9	Hap	51	23	19	13		
		Hap	Ple	Fea	Sad	Avg		Hap	Ple	Fea	Sad		Hap	Ple	Fea	Sad		
									Emotion									

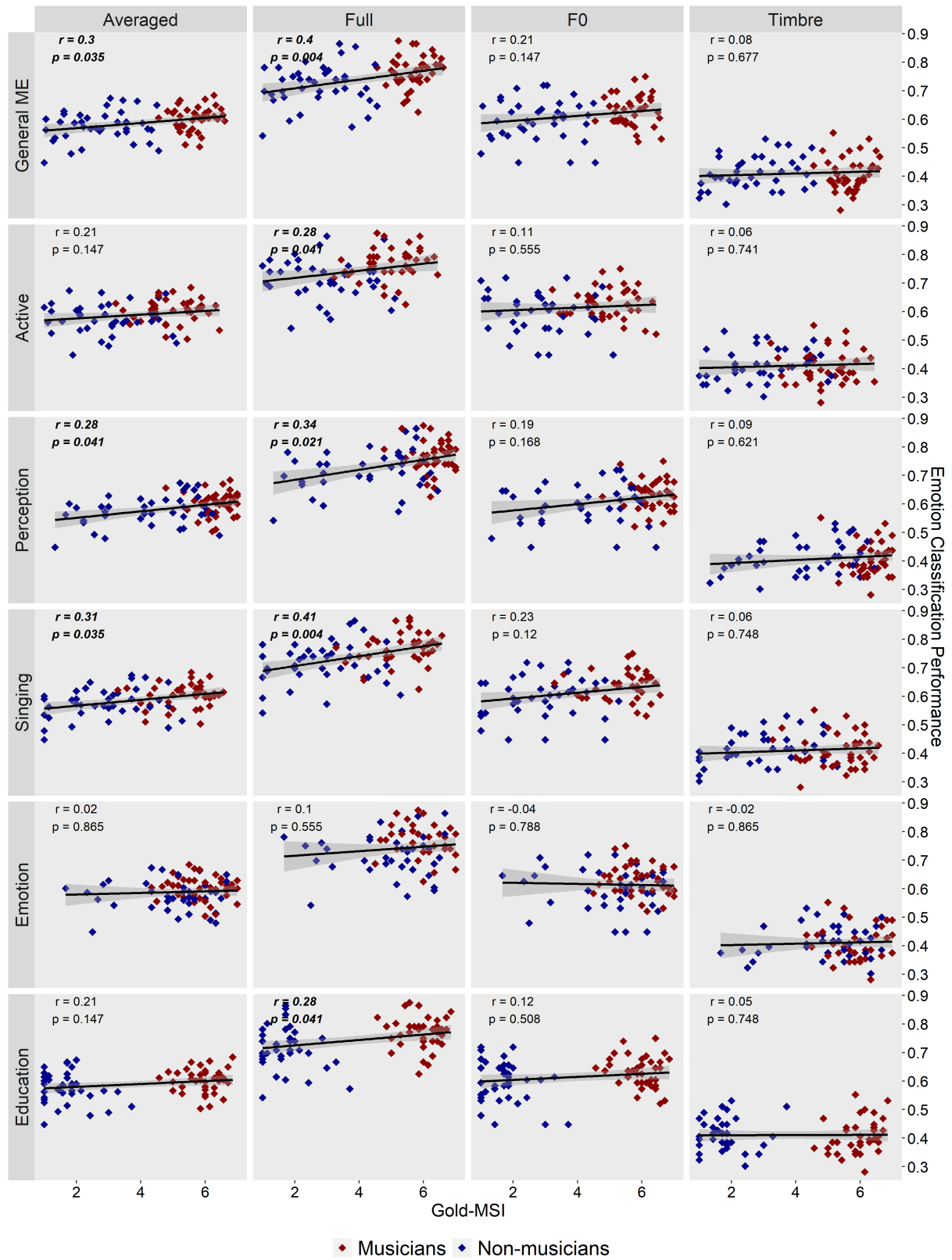
Note. Numbers represent the proportion of classification responses per Emotion and Morph Type, averaged across musicians. Hap = happiness, Ple = pleasure, Fea = fear, Sad = sadness, Avg = average.

Figure B.5.: Correlation between emotion classification performance and music perception abilities (PROMS)



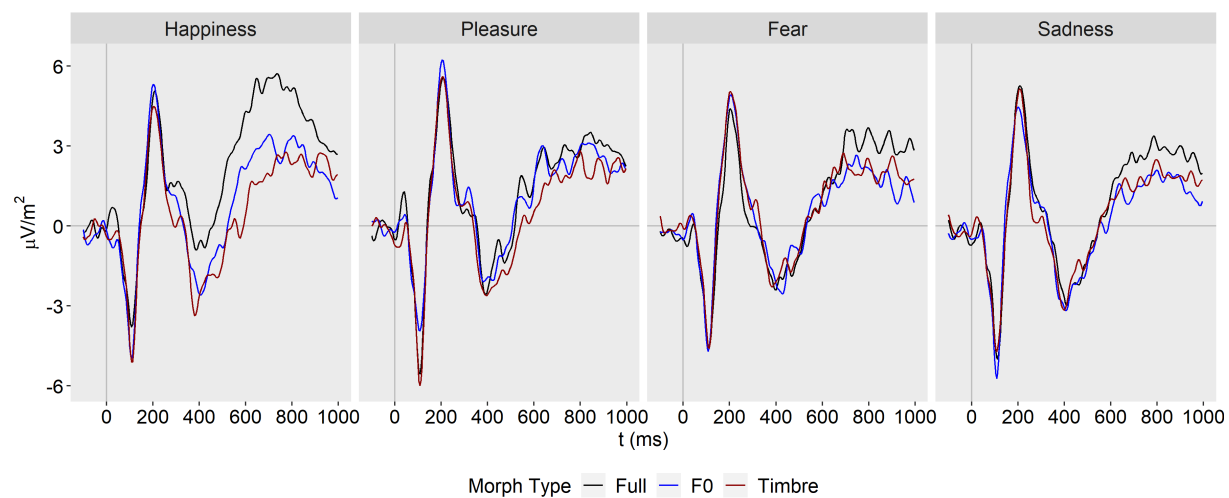
Note. The y-axis shows the different subtests of the PROMS (Pitch, Melody, Timbre, and Rhythm) as well as the averaged performance across all subtests (PROMS Averaged). The x-axis shows the vocal emotion classification performance separately for each Morph Type (Full, F0 and Timbre) and averaged across Morph Types (Averaged). p-values were adjusted for multiple comparisons using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995; false discovery rate set to 0.05, number of tests = 20).

Figure B.6.: Correlation between emotion classification performance and self-rated music skills (Gold-MSI)



Note. The y-axis shows the different subscores of the Gold-MSI. The x-axis shows the vocal emotion classification performance separately for each Morph Type and averaged across Morph Types (Averaged). p-values were adjusted for multiple comparisons using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995; false discovery rate set to 0.05, number of tests = 24).

Figure B.7.: ERPs separately for Emotion and Morph Type - centro-parietal ROI [-200, 1000]



Note. Averages are collapsed across [C1, Cz, C2, CP1, CPz, CP2, P1, Pz, P2].

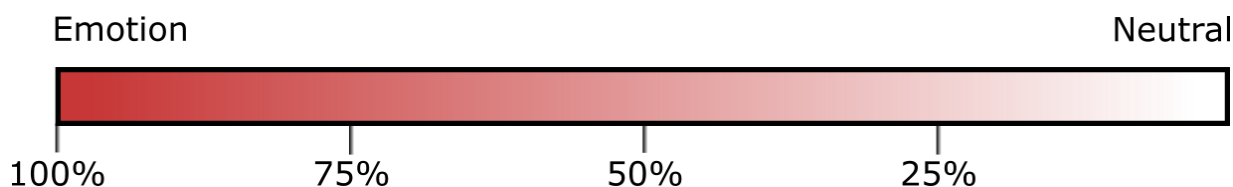
C. Supplemental Rating Data

Motivation/Rationale

Through voice morphing, the original stimulus material is manipulated in various ways. Some of the steps intentionally result in perceivable changes, others should result in perceptually identical voices (Kawahara & Skuk, 2018). In the Tandem-STRAIGHT voice morphing framework, the first processing step comprises the decomposition of the original voice material into source and filter information, subsequently allowing resynthesis of voices with different parameter modifications. However, if no parameter modifications are introduced, the resynthesis should result in a voice that is perceptually identical to the original stimulus. In the present dataset with emotional utterance this means that these resynthesized voices (herein referred to as continuum endpoints or 100% morphs) should be reliably identified as the intended emotion (**Assumption 1**). We then morphed the emotional voices with their neutral counterparts (within the same speaker and pseudoword) and created morphs in 25%-steps (refer to Figure C.1). We hypothesized, that with decreasing proportion of the emotional voices in the morphed stimuli, perceived emotional intensity and emotion classification accuracy would decrease (**Assumption 2**).

We designed this rating study test our Assumptions 1 and 2. Please note that we only used **full morphs** (encompassing all Tandem-STRAIGHT parameters) in this rating study to validate the morphing continua we created. Sufficient quality of the full morphs was considered as the crucial prerequisite to create parameter-specific voice morphs, since parameter-specific voice morphs are created from the same continua as the full morphs.

Figure C.1.: Illustration of created voice morphs



Note. Emotional voice morphs were created based on morphing continua between all emotional voices and the corresponding neutral voice of the same speaker and pseudowords. For each emotion continuum, four voice morphs encompassing 25%, 50%, 75% and 100% of the emotional voice were created.

Method

Raters

20 raters participated in the study (13 females, 7 males, aged 19 to 30 years [$M = 21.3$; $Mdn = 20$; $SD = 3.2$], 3 left-handed). Data was collected in October 2019.

Stimuli

Original Audio Recordings We selected original audio recordings from a database of vocal actor portrayals provided by Sascha Frühholz. For the present study, we used recordings from 8 speakers (4 male, 4 female) uttering 7 emotions (anger, disgust, fear, happiness, pleasure, sadness, and surprise) in 4 different pseudowords (/molen/, /loman/, /belam/, /namil/).

Voice Morphing Using the Tandem-STRAIGHT software (Kawahara et al., 2008, 2013), we created morphing trajectories between each emotion and the neutral expression of the same speaker and pseudoword; for a more detailed description refer to Kawahara and Skuk (2018). For each of these morphing trajectories, we created 4 voice morphs with different morph levels (all full-morphs encompassing all vocal parameters) lying on the 25%, 50%, 75% and 100% points of the neutral-emotion continuum (refer to Figure C.1). Using PRAAT (Boersma, 2018), we root-mean-square normalized all stimuli to 70 dB SPL. In total, this procedure resulted in $8 \text{ (speaker)} \times 7 \text{ (emotion)} \times 4 \text{ (pseudoword)} \times 4 \text{ (morph level)} = 896$ stimuli.

Procedure

After being informed about the purpose of the study and giving informed consent, participants were seated in front of the experimental computer in a quiet room. Up to three participants were tested simultaneously. All instructions were presented via computer screen. Participants' task was to listen to each stimulus and to decide on the emotional category and the emotional intensity in a forced choice format. Emotional categories were anger, disgust, fear, happiness, pleasure, sadness, and surprise. Regarding emotional intensity, participants could choose between 1 (very weak) – 4 (very strong). Responses were entered via mouse click.

Each trial started with a fixation cross for 500 ms. Then the target stimulus was presented, while the fixation cross remained on the screen. Afterwards, the first rating was prompted ("Which of the following emotion did the voice contain?"), which remained on the screen until participants entered their choice. Then, the second rating was presented ("To what extent did the voice contain the specified emotion?") and again it remained on the screen until the participant responded. Afterwards, the next trials started. Stimuli were presented in randomized order.

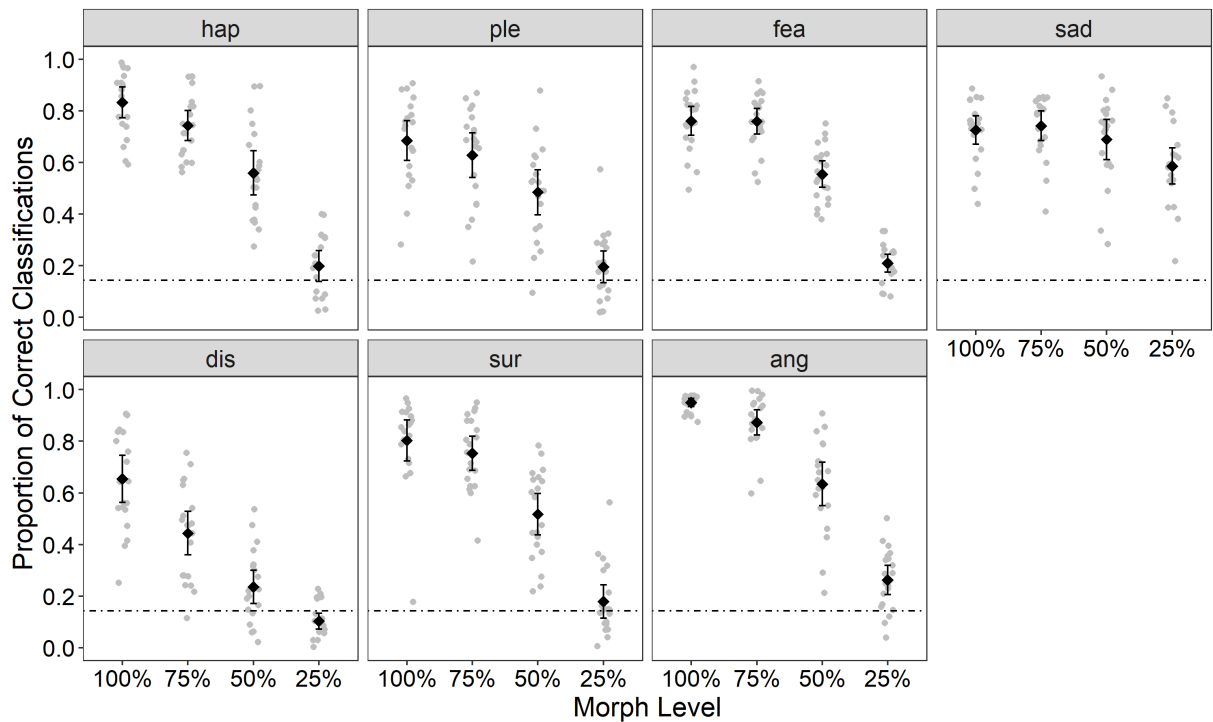
In the beginning, participants performed 8 practice trials to acquaint them with the experimental procedure, using stimuli that were not used in the main experiment later. Then they completed 10 blocks of 90 trials (the last block being a little shorter with just 86 trials), resulting in a total of 896 trials. Between blocks, participants were encouraged to take a short break. In total, duration of the experiment was about 60-80 minutes.

Results

Emotion classification

Emotion classification accuracy was analyzed in a 7 x 4 mixed-effects ANOVA with the within-subject factors Emotion (happiness, pleasure, fear, sadness, disgust, surprise, and anger) and Morph Level (100%, 75%, 50%, 25%). Both main effects and the interaction were significant (Emotion: $F(6, 114) = 19.59, p < .001, \eta_p^2 = 0.508$; Morph Level: $F(3, 57) = 912.35, p < .001, \eta_p^2 = 0.980$; Emotion x Morph Level: $F(18, 342) = 18.48, p < .001, \eta_p^2 = 0.493$; see Figure C.2).

Figure C.2.: Proportion of correct Classifications for each Emotion at different Morph Levels



Note. The dashed line represents the guessing rate of 14%. hap = happiness, ple = pleasure, fea = fear, sad = sadness, dis = disgust, sur = surprise, ang = anger.

Post-hoc analyses showed that the classification performances decreased as a function of Morph Level; 100% vs 75%: $t(19) = 9.53, p < .001$; 75% vs 50%: $t(19) = 22.80, p < .001$; and 50% vs 25%: $t(19) = 23.33, p < .001$. This pattern was also observed for each Emotion separately, $|ts(19)| \geq 2.42, ps \leq .026$, with the exception of the 100% vs 75% contrast in Fear, Sadness, and Surprise, $|ts(19)| \leq 1.96, ps \geq .064$.

Patterns of misclassifications are displayed in Figure C.3. Note the increase of misclassifications as sadness at the Morph Levels 50% and 25%.

Figure C.3.: Patterns of misclassifications for each Emotion separately for the four Morph Levels

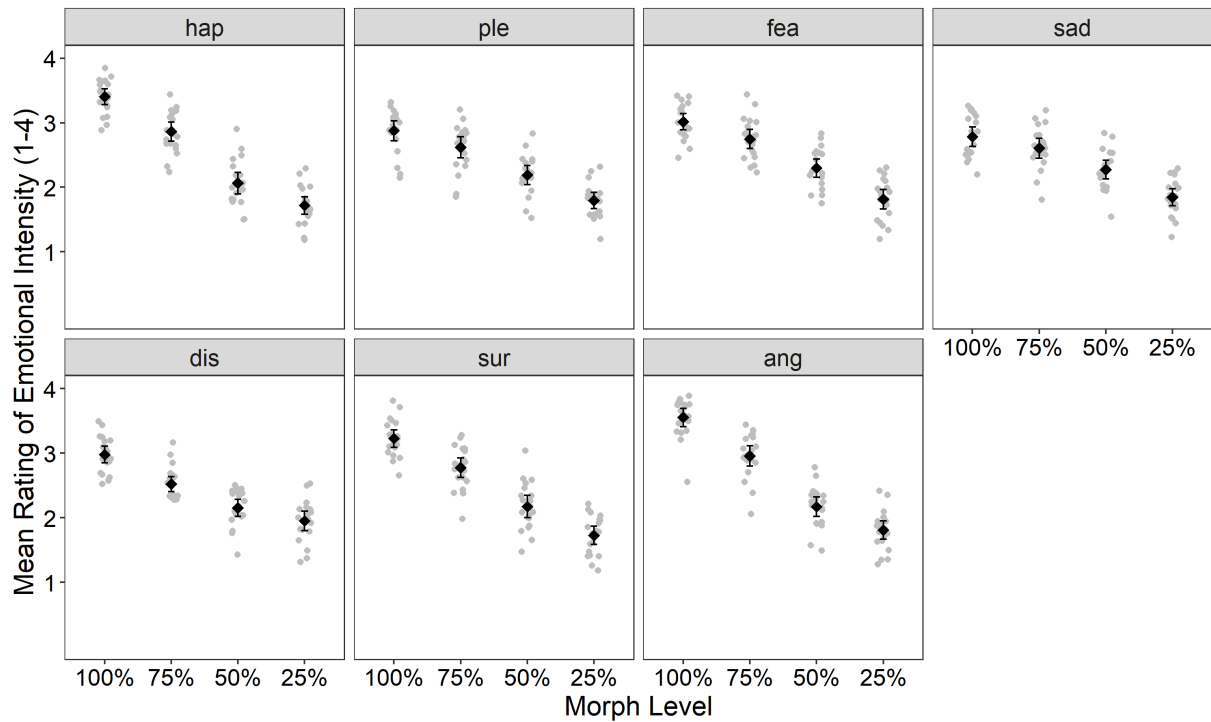
		100%							75%						
Mean Proportion of Responses in %	ang -	4	3	1	1	10	1	95	6	2	1	0	9	2	87
	sur -	11	2	4	0	3	80	0	15	3	3	0	3	75	2
	dis -	0	9	4	4	65	0	1	2	10	2	4	44	1	2
	sad -	0	6	7	72	12	0	0	1	12	11	74	23	1	2
	fea -	0	3	76	16	3	4	1	1	4	76	16	7	9	2
	ple -	1	68	0	6	3	0	0	1	63	1	5	6	0	0
	hap -	83	8	8	2	3	13	2	74	7	6	1	8	12	5
		50%							25%						
ang -	12	2	2	0	8	3	63	16	7	10	5	10	9	26	
sur -	8	2	3	2	2	52	2	5	3	6	4	2	18	3	
dis -	6	8	5	5	24	2	5	10	9	8	8	10	9	9	
sad -	9	26	26	69	42	6	12	31	46	46	59	57	26	39	
fea -	5	6	55	15	9	21	7	8	7	21	10	6	19	9	
ple -	4	48	3	5	6	0	2	10	20	5	6	5	7	5	
hap -	56	7	6	4	9	15	8	20	9	5	8	9	12	9	
		hap	ple	fea	sad	dis	sur	ang	hap	ple	fea	sad	dis	sur	ang
		Emotion													

Note. Numbers represent the proportion of classification responses per Emotion and Morph Type. hap = happiness, ple = pleasure, fea = fear, sad = sadness, dis = disgust, sur = surprise, ang = anger.

Ratings of emotional intensity

Rating of Emotional Intensity were coded from 1 (very weak) to 4 (very strong) and aggregated across speaker and pseudoword. A 7 x 4 mixed-effects ANOVA with the within-subject factors Emotion (happiness, pleasure, fear, sadness, disgust, surprise, and anger) and Morph Level (100%, 75%, 50%, 25%) revealed both significant main effects and the interaction (Emotion: $F(6, 114) = 6.28, p < .001, \eta_p^2 = 0.248$; Morph Level: $F(3, 57) = 506.02, p < .001, \eta_p^2 = 0.964$; Emotion x Morph Level: $F(18, 342) = 22.78, p < .001, \eta_p^2 = 0.545$; see Figure C.4).

Figure C.4.: Mean Intensity Ratings for each Emotion at different Morph Levels



Note. *hap* = happiness, *ple* = pleasure, *fea* = fear, *sad* = sadness, *dis* = disgust, *sur* = surprise, *ang* = anger.

Similar to the emotion classification performance, ratings decreased as a function of Morph Level: 100% vs 75%: $t(19) = 16.77, p < .001$; 75% vs 50%: $t(19) = 24.83, p < .001$; and 50% vs 25%: $t(19) = 15.59, p < .001$; for each Emotion separately: $|ts(19)| \geq 3.73, ps \leq .001$.

Emotion classification and intensity ratings in emotions with morph level 100%

The perception of the emotional stimuli with Morph Level 100% were of particular interest since they should be perceptually identical to the original emotional voices and thus should be reliably perceived as the intended emotion. As displayed in Figure 2, all emotions were identified with the highest rate at 100%. In fact, in the 100%-morphs, the proportion of correct classification was above .65 in all emotions (happiness: $0.83 \pm 0.03SE$, pleasure: 0.68 ± 0.04 , fear: 0.76 ± 0.03 , sadness: 0.72 ± 0.03 , disgust: 0.65 ± 0.04 ; surprise: 0.80 ± 0.04 , and anger: 0.95 ± 0.01). Furthermore, despite being rated highest in intensity compared to the other morph levels, there were differences between emotions in terms of perceived intensity at the 100% morph level (main effect of Emotion in a one-way ANOVA: $F(6, 114) = 34.713, p < .001$). Planned comparisons revealed that happiness was perceived as more intense than pleasure ($t(19) = 9.57, p < .001$, $M = 3.40 \pm 0.06$ and $M = 2.88 \pm 0.07$ for happiness and pleasure, respectively) and that fear was perceived as more intense than sadness ($t(19) = 6.58, p < .001$, $M = 3.01 \pm 0.06$ and $M = 2.78 \pm 0.07$ for fear and sadness, respectively).

Conclusion

The rating study confirmed our Assumptions 1 and 2: Emotional continuum endpoints (being the 100%-morphs) were perceived reliably as the intended emotion. Further, with decreasing emotional proportion in the voices, perceived intensity and classification accuracy decreased. This rating study was designed to validate our morphing approach and ensure the quality of the morphing continua. These were then not just used to create full morphs but also parameter-specific ones. We further showed that although intensity ratings varied profoundly as a function of morph level (refer to Figure C.4), we could still observe differences in perceived intensity in the 100%-morphs between different emotions.

D. References

- Adam Auton. (2021). Red blue colormap (MATLAB Central File Exchange., Ed.).
- Akkermans, J., Schapiro, R., Müllensiefen, D., Jakubowski, K., Shanahan, D., Baker, D., Busch, V., Lothwesen, K., Elvers, P., Fischinger, T., Schlemmer, K., & Frieler, K. (2019). Decoding emotions in expressive music performances: A multi-lab replication and extension study. *Cognition & Emotion*, 33(6), 1099–1118. <https://doi.org/10.1080/02699931.2018.1541312>
- Alku, P., Tiitinen, H., & Näätänen, R. (1999). A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology*, 110(8), 1329–1333. [https://doi.org/10.1016/S1388-2457\(99\)00088-7](https://doi.org/10.1016/S1388-2457(99)00088-7)
- American Psychological Association. (2020). *Publication manual of the american psychological association*. <https://doi.org/10.1037/0000165-000>
- Amorim, M., Roberto, M. S., Kotz, S. A., & Pinheiro, A. P. (2022). The perceived salience of vocal emotions is dampened in non-clinical auditory verbal hallucinations. *Cognitive Neuropsychiatry*, 27(2-3), 169–182. <https://doi.org/10.1080/13546805.2021.1949972>
- Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with parkinson's disease. *Journal of Speech, Language, and Hearing Research*, 58(4), 1134–1144. https://doi.org/10.1044/2015_JSLHR-S-14-0243
- Anikin, A. (2020). A moan of pleasure should be breathy: The effect of voice quality on the meaning of human nonverbal vocalizations. *Phonetica*, 77(5), 327–349. <https://doi.org/10.1159/000504855>
- ANSI. (1973). Terminology, Psychoacoustical. S3. 20. *Terminology*, New York: American National Standards Institute.
- Arias, P., Rachman, L., Liuni, M., & Aucouturier, J.-J. (2021). Beyond correlation: Acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, 13(1), 12–24. <https://doi.org/10.1177/1754073920934544>
- Assmann, P. F., Dembling, S., & Nearey, T. M. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. *INTERSPEECH*.
- Assmann, P. F., & Katz, W. F. (2000). Time-varying spectral change in the vowels of children and adults. *The Journal of the Acoustical Society of America*, 108(4), 1856–1866. <https://doi.org/10.1121/1.1289363>
- Aubé, W., Angulo-Perkins, A., Peretz, I., Concha, L., & Armony, J. L. (2015). Fear across the senses: Brain responses to music, vocalizations and facial expressions. *Social Cognitive and Affective Neuroscience*, 10(3), 399–407. <https://doi.org/10.1093/scan/nsu067>
- Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain : A Journal of Neurology*, 125(Pt 2), 238–251. <https://doi.org/10.1093/brain/awf028>
- Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2), 53–57. <https://doi.org/10.1111/1467-8721.00013>
- Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion - acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, 6(4), 219–224.
- Bachorowski, J. A., & Owren, M. J. (2003). Sounds of emotion: Production and perception of affect-related vocal acoustics. *Annals of the New York Academy of Sciences*, 1000, 244–265. <https://doi.org/10.1196/annals.1280.012>

- Baird, A., Jørgensen, S. H., Parada-Cabaleiro, E., Cummings, N., Hantke, S., & Schüller, B. (2018). The perception of vocal traits in synthesized voices: Age, gender, and human likeness. *Journal of the Audio Engineering Society*, 66(4), 277–285. <https://doi.org/10.17743/jaes.2018.0023>
- Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummings, N., & Schüller, B. (2018). The perception and analysis of the likeability and human likeness of synthesized speech. *Interspeech 2018*, 2863–2867. <https://doi.org/10.21437/Interspeech.2018-1093>
- Balkwill, L.-L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception: An Interdisciplinary Journal*, 17(1), 43–64. <https://doi.org/10.2307/40285811>
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*, 70(3), 614–636. <https://doi.org/10.1037/0022-3514.70.3.614>
- Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PLoS One*, 10(9), e0136675. <https://doi.org/10.1371/journal.pone.0136675>
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J.-C., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1), 5–17.
- Barrett, S. E., & Rugg, M. D. (1989). Event-related potentials and the semantic matching of faces. *Neuropsychologia*, 27(7), 913–922. [https://doi.org/10.1016/0028-3932\(89\)90067-5](https://doi.org/10.1016/0028-3932(89)90067-5)
- Barrett, S. E., Rugg, M. D., & Perrett, D. I. (1988). Event-related potentials and the matching of familiar and unfamiliar faces. *Neuropsychologia*, 26(1), 105–117. [https://doi.org/10.1016/0028-3932\(88\)90034-6](https://doi.org/10.1016/0028-3932(88)90034-6)
- Başkent, D., Fuller, C. D., Galvin, J. J., Schepel, L., Gaudrain, E., & Free, R. H. (2018). Musician effect on perception of spectro-temporally degraded speech, vocal emotion, and music in young adolescents. *The Journal of the Acoustical Society of America*, 143(5), EL311. <https://doi.org/10.1121/1.5034489>
- Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *Br J Psychol*, 102(4), 711–725. <https://doi.org/10.1111/j.2044-8295.2011.02041.x>
- Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends Cogn Sci*, 8(3), 129–135. <https://doi.org/10.1016/j.tics.2004.01.008>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Besson, M., Schön, D., Moreno, S., Santos, A., & Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restorative Neurology and Neuroscience*, 25(3-4), 399–410.
- Bestelmeyer, P. E. G., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *J Neurosci*, 34(24), 8098–8105. <https://doi.org/10.1523/JNEUROSCI.4820-13.2014>
- Bestelmeyer, P. E. G., & Mühl, C. (2021). Individual differences in voice adaptability are specifically linked to voice perception skill. *Cognition*, 210, 104582. <https://doi.org/10.1016/j.cognition.2021.104582>
- Bestelmeyer, P. E. G., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition*, 117(2), 217–223. <https://doi.org/10.1016/j.cognition.2010.08.008>
- Bhatara, A., Tirovolas, A. K., Duan, L. M., Levy, B., & Levitin, D. J. (2011). Perception of emotional expression in musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3), 921–934. <https://doi.org/10.1037/a0021922>
- Bianco, R., Novembre, G., Keller, P. E., Villringer, A., & Sammler, D. (2018). Musical genre-dependent behavioural and EEG signatures of action planning: a comparison between

- classical and jazz pianists. *Neuroimage*, 169, 383–394. <https://doi.org/10.1016/j.neuroimage.2017.12.058>
- Bidelman, G. M. (2017). Amplified induced neural oscillatory activity predicts musicians' benefits in categorical speech perception. *Neuroscience*, 348, 107–113. <https://doi.org/10.1016/j.neuroscience.2017.02.015>
- Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. *Cognition*, 100(1), 100–130. <https://doi.org/10.1016/j.cognition.2005.11.007>
- Blonder, L. X., Pettigrew, L. C., & Kryscio, R. J. (2012). Emotion recognition and marital satisfaction in stroke. *Journal of clinical and experimental neuropsychology*, 34(6), 634–642. <https://doi.org/10.1080/13803395.2012.667069>
- Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences*, 98(20), 11818–11823. <https://doi.org/10.1073/pnas.191355898>
- Bodner, E., Aharoni, R., & Iancu, I. (2012). The effect of training with music on happiness recognition in social anxiety disorder. *Journal of Psychopathology and Behavioral Assessment*, 34(4), 458–466. <https://doi.org/10.1007/s10862-012-9304-7>
- Boersma, P. (2018). Praat: doing phonetics by computer [Computer program]: Version 6.0.46. <http://www.praat.org>.
- Bonde, L. O., Juel, K., & Ekholm, O. (2018). Associations between music and health-related outcomes in adult non-musicians, amateur musicians and professional musicians—results from a nationwide danish study. *Nordic Journal of Music Therapy*, 27(4), 262–282. <https://doi.org/10.1080/08098131.2018.1439086>
- Bonnell, A., Mottron, L., Peretz, I., Trudel, M., Gallun, E., & Bonnell, A. (2003). Enhanced pitch sensitivity in individuals with autism: A signal detection analysis. *Journal of Cognitive Neuroscience*, 15(2), 226–235. <https://doi.org/10.1162/089892903321208169>
- Breyer, B., & Bluemke, M. (2016). Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel). <https://doi.org/10.6102/zis242>
- Brück, C., Kreifelts, B., & Wildgruber, D. (2011). Emotional voices in context: A neurobiological model of multimodal affective information processing. *Phys Life Rev*, 8(4), 383–403. <https://doi.org/10.1016/j.plrev.2011.10.002>
- Bruckert, L., Bestelmeyer, P. E. G., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology*, 20(2), 116–120. <https://doi.org/10.1016/j.cub.2009.11.034>
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univ of California Press.
- Burgering, M. A., van Laarhoven, T., Baart, M., & Vroomen, J. (2020). Fluidity in the perception of auditory speech: Cross-modal recalibration of voice gender and vowel identity by a talking face. *Q J Exp Psychol (Hove)*, 73(6), 957–967. <https://doi.org/10.1177/1747021819900884>
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of german emotional speech. *Interspeech*, 5, 1517–1520.
- Burton, M. W., & Blumstein, S. E. (1995). Lexical effects on phonetic categorization: The role of stimulus naturalness and stimulus quality. *Journal of Experimental Psychology: Human Perception and Performance*, 21(5), 1230–1235. <https://doi.org/10.1037/0096-1523.21.5.1230>
- Cabral, J. P., Cowan, B. R., Zibrek, K., & McDonnell, R. (2017). The influence of synthetic voice on the evaluation of a virtual character. *Interspeech 2017*, 229–233. <https://doi.org/10.21437/Interspeech.2017-325>
- Calder, A. J., Rowland, D., Young, A. W., Nimmo-Smith, I., Keane, J., & Perrett, D. I. (2000). Caricaturing facial expressions. *Cognition*, 76(2), 105–146. [https://doi.org/10.1016/S0010-0277\(00\)00074-3](https://doi.org/10.1016/S0010-0277(00)00074-3)

- Carton, J. S., Kessler, E. A., & Pape, C. L. (1999). Nonverbal decoding skills and relationship well-being in adults. *Journal of Nonverbal Behavior*, 23(1), 91–100. <https://doi.org/10.1023/A:1021339410262>
- Chari, D. A., Barrett, K. C., Patel, A. D., Colgrove, T. R., Jiradejvong, P., Jacobs, L. Y., & Limb, C. J. (2020). Impact of auditory-motor musical training on melodic pattern recognition in cochlear implant users. *Otology & Neurotology : Official Publication of the American Otological Society, American Neurotology Society and European Academy of Otology and Neurotology*, 41(4), e422–e431. <https://doi.org/10.1097/MAO.0000000000002525>
- Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters*, 405(3), 164–167. <https://doi.org/10.1016/j.neulet.2006.06.053>
- Chartrand, J.-P., Peretz, I., & Belin, P. (2008). Auditory recognition expertise and domain specificity. *Brain Research*, 1220, 191–198. <https://doi.org/10.1016/j.brainres.2008.01.014>
- Chen, X., Yang, J., Gan, S., & Yang, Y. (2012). The contribution of sound intensity in vocal emotion perception: Behavioral and electrophysiological evidence. *PLoS One*, 7(1), e30278. <https://doi.org/10.1371/journal.pone.0030278>
- Cheung, Y. L., Zhang, C., & Zhang, Y. (2020). Emotion processing in congenital amusia: The deficits do not generalize to written emotion words. *Clinical Linguistics & Phonetics*, 1–16. <https://doi.org/10.1080/02699206.2020.1719209>
- Christensen, J. A., Sis, J., Kulkarni, A. M., & Chatterjee, M. (2019). Effects of age and hearing loss on the recognition of emotions in speech. *Ear & Hearing*, 40(5), 1069–1083. <https://doi.org/10.1097/AUD.0000000000000694>
- Christensen, R. H. B. (2015). Ordinal: Regression models for ordinal data. (*R package [Computer software]*).
- Clark, C. N., Downey, L. E., & Warren, J. D. (2015). Brain disorders and the biological role of music. *Social Cognitive and Affective Neuroscience*, 10(3), 444–452. <https://doi.org/10.1093/scan/nsu079>
- Correia, A. I., Castro, S. L., MacGregor, C., Müllensiefen, D., Schellenberg, E. G., & Lima, C. F. (2022). Enhanced recognition of vocal emotions in individuals with naturally good musical abilities. *Emotion*, 22(5), 894–906. <https://doi.org/10.1037/emo0000770>
- Corrigall, K. A., Schellenberg, E. G., & Misura, N. M. (2013). Music training, cognition, and personality. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00222>
- Cotral-Labor-GmbH. (2013). Labor Cotral GmbH, Computer Software, Version 1.02B.
- Coughlin-Woods, S., Lehman, M. E., & Cooke, P. A. (2005). Ratings of speech naturalness of children ages 8-16 years. *Perceptual and Motor Skills*, 100(2), 295–304. <https://doi.org/10.2466/pms.100.2.295-304>
- Crookes, K., Ewing, L., Gildenhuys, J.-D., Kloth, N., Hayward, W. G., Oxner, M., Pond, S., & Rhodes, G. (2015). How well do computer-generated faces tap face expertise? *PLoS One*, 10(11), e0141353. <https://doi.org/10.1371/journal.pone.0141353>
- Crumpton, J., & Bethel, C. L. (2016). A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics*, 8(2), 271–285. <https://doi.org/10.1007/s12369-015-0329-4>
- Cui, A., & Kuang, J. (2019). The effects of musicality and language background on cue integration in pitch perception. *The Journal of the Acoustical Society of America*, 146(6), 4086. <https://doi.org/10.1121/1.5134442>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, 25(1), 7–29.
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. John Murray, London.
- Darwin, C. (1872). *The expression of the emotions in man and animals*. Oxford University Press, USA.
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>

- Denham, S. L., & Winkler, I. (2020). Predictive coding in auditory perception: Challenges and unresolved questions. *European Journal of Neuroscience*, 51(5), 1151–1160. <https://doi.org/10.1111/ejn.13802>
- Di Yang, Tao, H., Ge, H., Li, Z., Hu, Y., & Meng, J. (2022). Altered processing of social emotions in individuals with autistic traits. *Frontiers in Psychology*, 13, 746192. <https://doi.org/10.3389/fpsyg.2022.746192>
- Dibben, N., Coutinho, E., Vilar, J. A., & Estévez-Pérez, G. (2018). Do individual differences influence moment-by-moment reports of emotion perceived in music and speech prosody? *Frontiers in Behavioral Neuroscience*, 12, 184. <https://doi.org/10.3389/fnbeh.2018.00184>
- Dmitrieva, E. S., Gel'man, V. Y., Zaitseva, K. A., & Am Orlov. (2006). Ontogenetic features of the psychophysiological mechanisms of perception of the emotional component of speech in musically gifted children. *Neuroscience and Behavioral Physiology*, 36(1), 53.
- Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (te) speakers. *J Speech Lang Hear Res*, 45(6), 1088–1096. [https://doi.org/10.1044/1092-4388\(2002/087\)](https://doi.org/10.1044/1092-4388(2002/087))
- Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99(3), 550–553. <https://doi.org/10.1037/0033-295X.99.3.550>
- Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B., & Taub, E. (1995). Increased cortical representation of the fingers of the left hand in string players. *Science*, 270(5234), 305–307.
- Elmer, S., Dittinger, E., & Besson, M. (2018). One step beyond – musical expertise and word learning. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 208–234). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198743187.013.10>
- Escoffier, N., Zhong, J., Schirmer, A., & Qiu, A. (2013). Emotional expressions in voice and music: Same code, same effect? *Human brain mapping*, 34(8), 1796–1810. <https://doi.org/10.1002/hbm.22029>
- Fant, G. (1970). *Acoustic theory of speech production: With calculations based on x-ray studies of russian articulations* (Vol. 2). Walter de Gruyter.
- Farmer, E., Jicol, C., & Petrini, K. (2020). Musicianship enhances perception but not feeling of emotion from others' social interaction through speech prosody. *Music Perception: An Interdisciplinary Journal*, 37(4), 323–338. <https://doi.org/10.1525/MP.2020.37.4.323>
- Fitch, W. T. (2013). Musical protolanguage: Darwin's theory of language evolution revisited. *Birdsong, Speech, and Language: Exploring the Evolution of Mind and Brain*, 489, 503.
- Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology*, 8, 1180. <https://doi.org/10.3389/fpsyg.2017.01180>
- Freitag, C. M., Retz-Junginger, P., Retz, W., Seitz, C., Palmason, H., Meyer, J., Rösler, M., & von Gontard, A. (2007). Evaluation der deutschen Version des Autismus-Spektrum-Quotienten (AQ) - die Kurzversion AQ-k. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 36(4), 280–289. <https://doi.org/10.1026/1616-3443.36.4.280>
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology. General*, 141(1), 2–18. <https://doi.org/10.1037/a0024338>
- Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A. D., & Koelsch, S. (2009). Universal recognition of three basic emotions in music. *Current Biology*, 19(7), 573–576.
- Frühholz, S., Klaas, H. S., Patel, S., & Grandjean, D. (2015). Talking in fury: The cortico-subcortical network underlying angry vocalizations. *Cerebral Cortex*, 25(9), 2752–2762. <https://doi.org/10.1093/cercor/bhu074>

- Frühholz, S., & Schweinberger, S. R. (2021). Nonverbal auditory communication - evidence for integrated neural systems for voice signal production and perception. *Progress in Neurobiology*, 199, 101948. <https://doi.org/10.1016/j.pneurobio.2020.101948>
- Frühholz, S., Trost, W., & Grandjean, D. (2014). The role of the medial temporal limbic system in processing emotions in voice and music. *Progress in Neurobiology*, 123, 1–17. <https://doi.org/10.1016/j.pneurobio.2014.09.003>
- Frühholz, S., Trost, W., & Kotz, S. A. (2016). The sound of emotions- towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews*, 68, 96–110. <https://doi.org/10.1016/j.neubiorev.2016.05.002>
- Fuller, C. D., Galvin, J. J., Maat, B., Başkent, D., & Free, R. H. (2018). Comparison of two music training approaches on music and speech perception in cochlear implant users. *Trends in Hearing*, 22, 2331216518765379. <https://doi.org/10.1177/2331216518765379>
- Fuller, C. D., Galvin, J. J., Maat, B., Free, R. H., & Başkent, D. (2014). The musician effect: Does it persist under degraded pitch conditions of cochlear implant simulations? *Frontiers in Neuroscience*, 8, 179. <https://doi.org/10.3389/fnins.2014.00179>
- Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., Surguladze, S., Benedetti, F., Abbamonte, M., Gasparotti, R., Barale, F., Perez, J., McGuire, P., & Politi, P. (2009). Functional atlas of emotional faces processing: A voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of Psychiatry and Neuroscience*, 34(6), 418–432. <https://www.jpn.ca/content/34/6/418.short>
- Globerson, E., Amir, N., Golan, O., Kishon-Rabin, L., & Lavidor, M. (2013). Psychoacoustic abilities as predictors of vocal emotion recognition. *Attention, Perception & Psychophysics*, 75(8), 1799–1810. <https://doi.org/10.3758/s13414-013-0518-x>
- Globerson, E., Amir, N., Kishon-Rabin, L., & Golan, O. (2015). Prosody recognition in adults with high-functioning autism spectrum disorders: From psychoacoustics to cognition. *Autism Research*, 8(2), 153–163. <https://doi.org/10.1002/aur.1432>
- Gobl, C. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2), 189–212. [https://doi.org/10.1016/S0167-6393\(02\)00082-1](https://doi.org/10.1016/S0167-6393(02)00082-1)
- Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, 24(4), 1494–1509. <https://doi.org/10.1016/j.chb.2007.05.007>
- Good, A., Gordon, K. A., Papsin, B. C., Nespoli, G., Hopyan, T., Peretz, I., & Russo, F. A. (2017). Benefits of music training for perception of emotional speech prosody in deaf children with cochlear implants. *Ear & Hearing*, 38(4), 455.
- Grandjean, D. (2021). Brain networks of emotional prosody processing. *Emotion Review*, 13(1), 34–43. <https://doi.org/10.1177/1754073919898522>
- Greenspon, E. B., & Montanaro, V. (2023). Singing ability is related to vocal emotion recognition: Evidence for shared sensorimotor processing across speech and music. *Attention, Perception & Psychophysics*, 85(1), 234–243. <https://doi.org/10.3758/s13414-022-02613-0>
- Grichkovtsova, I., Morel, M., & Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3), 414–429. <https://doi.org/10.1016/j.specom.2011.10.005>
- Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J Cogn Neurosci*, 29(4), 677–697. https://doi.org/10.1162/jocn_a_01068
- Hajarolasvadi, N., Ramirez, M. A., Beccaro, W., & Demirel, H. (2020). Generative adversarial networks in human emotion synthesis: A review. *IEEE Access*, 8, 218499–218529. <https://doi.org/10.1109/ACCESS.2020.3042328>
- Hajcak, G., & Foti, D. (2020). Significance?... Significance! Empirical, methodological, and theoretical connections between the late positive potential and P300 as neural responses

- to stimulus significance: An integrative review. *Psychophysiology*, 57(7). <https://doi.org/10.1111/psyp.13570>
- Hallam, S. (2017). The impact of making music on aural perception and language skills: A research synthesis. *London Review of Education*, 15(3), 388–406. <https://doi.org/10.18546/Lre.15.3.05>
- Heaton, P., Hermelin, B., & Pring, L. (1998). Autism and pitch processing: A precursor for savant musical ability? *Music Perception: An Interdisciplinary Journal*, 15(3), 291–305. <https://doi.org/10.2307/40285769>
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57(2), 243. <https://doi.org/10.2307/1416950>
- Herholz, S. C., & Zatorre, R. J. (2012). Musical training as a framework for brain plasticity: Behavior, function, and structure. *Neuron*, 76(3), 486–502. <https://doi.org/10.1016/j.neuron.2012.10.011>
- Hoekstra, R. A., Bartels, M., Cath, D. C., & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the Autism-Spectrum Quotient (AQ): A study in dutch population and patient groups. *J Autism Dev Disord*, 38(8), 1555–1566. <https://doi.org/10.1007/s10803-008-0538-x>
- Hortensius, R., Hekele, F., & Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4), 852–864. <https://doi.org/10.1109/TCDS.2018.2826921>
- Hubbard, D. J., & Assmann, P. F. (2013). Perceptual adaptation to gender and expressive properties in speech: The role of fundamental frequency. *J Acoust Soc Am*, 133(4), 2367–2376. <https://doi.org/10.1121/1.4792145>
- Hunter, P. G., & Schellenberg, E. G. (2010). Music and emotion. In M. Riess Jones, R. R. Fay & A. N. Popper (Eds.), *Music Perception* (pp. 129–164, Vol. 36). Springer New York. https://doi.org/10.1007/978-1-4419-6114-3_5
- Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception: An Interdisciplinary Journal*, 23(4), 319–330. <https://doi.org/10.1525/mp.2006.23.4.319>
- Ilie, G., & Thompson, W. F. (2011). Experiential and cognitive changes following seven minutes exposure to music and speech. *Music Perception: An Interdisciplinary Journal*, 28(3), 247–264. <https://doi.org/10.1525/Mp.2011.28.3.247>
- Ilves, M., & Surakka, V. (2013). Subjective responses to synthesised speech with lexical emotional content: The effect of the naturalness of the synthetic voice. *Behaviour & Information Technology*, 32(2), 117–131. <https://doi.org/10.1080/0144929X.2012.702285>
- Ilves, M., Surakka, V., & Vanhala, T. (2011). The effects of emotionally worded synthesized speech on the ratings of emotions and voice quality, 588–598. https://doi.org/10.1007/978-3-642-24600-5_62
- Jiam, N. T., Caldwell, M., Deroche, M. L., Chatterjee, M., & Limb, C. J. (2017). Voice emotion perception and production in cochlear implant users. *Hear Res*, 352, 30–39. <https://doi.org/10.1016/j.heares.2017.01.006>
- Jiang, X., Paulmann, S., Robin, J., & Pell, M. D. (2015). More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures. *Journal of Experimental Psychology: Human Perception and Performance*, 41(3), 597–612. <https://doi.org/10.1037/xhp0000043>
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychol Bull*, 129(5), 770–814. <https://doi.org/10.1037/0033-2909.129.5.770>
- Kachel, S., Radtke, A., Skuk, V. G., Zäske, R., Simpson, A. P., & Steffens, M. C. (2018). Investigating the common set of acoustic parameters in sexual orientation groups: A voice

- averaging approach. *PLoS One*, 13(12), e0208686. <https://doi.org/10.1371/journal.pone.0208686>
- Kaganovich, N., Kim, J., Herring, C., Schumaker, J., Macpherson, M., & Weber-Fox, C. (2013). Musicians show general enhancement of complex sound encoding and better inhibition of irrelevant auditory change in music: An erp study. *The European Journal of Neuroscience*, 37(8), 1295–1307. <https://doi.org/10.1111/ejn.12110>
- Kätsyri, J., Förger, K., Mäkräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, 6, 390. <https://doi.org/10.3389/fpsyg.2015.00390>
- Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4), 187–207. [https://doi.org/10.1016/S0167-6393\(98\)00085-5](https://doi.org/10.1016/S0167-6393(98)00085-5)
- Kawahara, H., Morise, M., & Skuk, V. G. (2013). Temporally variable multi-aspect n-way morphing based on interference-free speech representations. *IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 3933–3936.
- Kawahara, H., & Skuk, V. G. (2018). Voice morphing. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 684–706). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198743187.013.31>
- Kayser, J. (2009). Current source density (CSD) interpolation using spherical splines-csd toolbox. *New York State Psychiatric Institute*.
- Kayser, J., & Tenke, C. E. (2006). Principal components analysis of laplacian waveforms as a generic method for identifying erp generator patterns: I. evaluation with auditory oddball tasks. *Clinical Neurophysiology*, 117(2), 348–368. <https://doi.org/10.1016/j.clinph.2005.08.034>
- Kayser, J., & Tenke, C. E. (2015). On the benefits of using surface laplacian (current source density) methodology in electrophysiology. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, 97(3), 171–173. <https://doi.org/10.1016/j.ijpsycho.2015.06.001>
- Kim, S. (2015). Ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22(6), 665–674. <https://doi.org/10.5351/CSAM.2015.22.6.665>
- Klopfenstein, M., Bernard, K., & Heyman, C. (2020). The study of speech naturalness in communication disorders: A systematic review of the literature. *Clinical Linguistics & Phonetics*, 34(4), 327–338. <https://doi.org/10.1080/02699206.2019.1652692>
- Kloth, N., Rhodes, G., & Schweinberger, S. R. (2017). Watching the brain recalibrate: Neural correlates of renormalization during face adaptation. *Neuroimage*, 155, 1–9. <https://doi.org/10.1016/j.neuroimage.2017.04.049>
- Koelsch, S., Schröger, E., & Tervaniemi, M. (1999). Superior pre-attentive auditory processing in musicians. *Neuroreport*, 10(6), 1309–1313. <https://doi.org/10.1097/00001756-199904260-00029>
- Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, 11(8), 599–605. <https://doi.org/10.1038/nrn2882>
- Kraus, N., Hornickel, J., Strait, D. L., Slater, J., & Thompson, E. (2014). Engagement in community music classes sparks neuroplasticity and language development in children from disadvantaged backgrounds. *Frontiers in Psychology*, 5, 1403. <https://doi.org/10.3389/fpsyg.2014.01403>

- Kraus, N., & White-Schwoch, T. (2017). Neurobiology of everyday communication: What have we learned from music? *The Neuroscientist : a Review Journal bringing Neurobiology, Neurology and Psychiatry*, 23(3), 287–298. <https://doi.org/10.1177/1073858416653593>
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (erp). *Annu Rev Psychol*, 62, 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205. <https://doi.org/10.1126/science.7350657>
- Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America*, 78(2), 435–444. <https://doi.org/10.1121/1.392466>
- Ladefoged, P. (1996). *Elements of acoustic phonetics*. University of Chicago Press.
- Lagrois, M.-É., & Peretz, I. (2019). The co-occurrence of pitch and rhythm disorders in congenital amusia. *Cortex*, 113, 229–238. <https://doi.org/10.1016/j.cortex.2018.11.036>
- Lakens, D., & Caldwell, A. R. (2019). *Simulation-based power-analysis for factorial anova designs*. <https://doi.org/10.31234/osf.io/baxsf>
- Lappe, C., Herholz, S. C., Trainor, L. J., & Pantev, C. (2008). Cortical plasticity induced by short-term unimodal and multimodal musical training. *Journal of Neuroscience*, 28(39), 9632–9639. <https://doi.org/10.1523/JNEUROSCI.2254-08.2008>
- Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *J Pers Soc Psychol*, 111(5), 686–705. <https://doi.org/10.1037/pspi0000066>
- Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1997). On the perception of emotions in speech: The role of voice quality. *Logopedics Phoniatrics Vocology*, 22(4), 157–168. <https://doi.org/10.3109/14015439709075330>
- Lausen, A., & Hammerschmidt, K. (2020). Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, 7(1). <https://doi.org/10.1057/s41599-020-0499-z>
- Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 26(1), 90–102. <https://doi.org/10.3758/s13423-018-1497-7>
- Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLoS One*, 7(12), e52508. <https://doi.org/10.1371/journal.pone.0052508>
- Lehmann, A., & Paquette, S. (2015). Cross-domain processing of musical and vocal emotions in cochlear implant users. *Frontiers in Neuroscience*, 9, 343. <https://doi.org/10.3389/fnins.2015.00343>
- Lima, C. F., Anikin, A., Monteiro, A. C., Scott, S. K., & Castro, S. L. (2019). Automaticity in the recognition of nonverbal emotional vocalizations. *Emotion*, 19(2), 219–233. <https://doi.org/10.1037/emo0000429>
- Lima, C. F., Brancatisano, O., Fancourt, A., Müllensiefen, D., Scott, S. K., Warren, J. D., & Stewart, L. (2016). Impaired socio-emotional processing in a developmental music disorder. *Scientific Reports*, 6, 34911. <https://doi.org/10.1038/srep34911>
- Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, 11(5), 1021–1031. <https://doi.org/10.1037/a0024521>
- Livingstone, S. R., & Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north

- american english. *PLoS One*, 13(5), e0196391. <https://doi.org/10.1371/journal.pone.0196391>
- Lolli, S. L., Lewenstein, A. D., Basurto, J., Winnik, S., & Loui, P. (2015). Sound frequency affects speech emotion perception: Results from congenital amusia. *Frontiers in Psychology*, 6, 1340. <https://doi.org/10.3389/fpsyg.2015.01340>
- Lütkenhöner, B., Seither-Preisler, A., & Seither, S. (2006). Piano tones evoke stronger magnetic fields than pure tones or noise, both in musicians and non-musicians. *Neuroimage*, 30(3), 927–937. <https://doi.org/10.1016/j.neuroimage.2005.10.034>
- Mackey, L. S., Finn, P., & Ingham, R. J. (1997). Effect of speech dialect on speech naturalness ratings: A systematic replication of martin, haroldson, and triden (1984). *Journal of Speech, Language, and Hearing Research*, 40(2), 349–360. <https://doi.org/10.1044/jslhr.4002.349>
- Mankel, K., & Bidelman, G. M. (2018). Inherent auditory skills rather than formal music training shape the neural encoding of speech. *Proceedings of the National Academy of Sciences*, 115(51), 13129–13134. <https://doi.org/10.1073/pnas.1811793115>
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- Martin, R. R., Haroldson, S. K., & Triden, K. A. (1984). Stuttering and speech naturalness. *The Journal of Speech and Hearing Disorders*, 49(1), 53–58. <https://doi.org/10.1044/jshd.4901.53>
- Martins, I., Lima, C. F., & Pinheiro, A. P. (2022). Enhanced salience of musical sounds in singers and instrumentalists. *Cogn Affect Behav Neurosci*. <https://doi.org/10.3758/s13415-022-01007-x>
- Martins, M., Pinheiro, A. P., & Lima, C. F. (2021). Does music training improve emotion recognition abilities? A Critical Review. *Emotion Review*, 13(3), 199–210. <https://doi.org/10.1177/17540739211022035>
- MATLAB. (2020). *Version 9.8.0 (R2020a)*. The MathWorks Inc.
- Mayo, C., Clark, R. A. J., & King, S. (2011). Listeners’ weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. *Speech Communication*, 53(3), 311–326. <https://doi.org/10.1016/j.specom.2010.10.003>
- McAler, P., Todorov, A., & Belin, P. (2014). How do you say ‘Hello’? Personality impressions from brief novel voices. *PLoS One*, 9(3), e90779. <https://doi.org/10.1371/journal.pone.0090779>
- McGinn, C., & Torre, I. (2019). Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 211–221. <https://doi.org/10.1109/HRI.2019.8673305>
- Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O’Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, 366(6468). <https://doi.org/10.1126/science.aax0868>
- Meister, H., Fuersen, K., Streicher, B., Lang-Roth, R., & Walger, M. (2020). Letter to the editor concerning Skuk et al., “parameter-specific morphing reveals contributions of timbre and fundamental frequency cues to the perception of voice gender and age in cochlear implant users”. *Journal of Speech, Language, and Hearing Research*, 63(12), 4325–4326. https://doi.org/10.1044/2020_JSLHR-20-00563
- Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *J Speech Lang Hear Res*, 48(4), 766–779. [https://doi.org/10.1044/1092-4388\(2005/053\)](https://doi.org/10.1044/1092-4388(2005/053))
- Merrett, D. L., Peretz, I., & Wilson, S. J. (2013). Moderating variables of music training-induced neuroplasticity: A review and discussion. *Frontiers in Psychology*, 4, 606. <https://doi.org/10.3389/fpsyg.2013.00606>

- Mill, A., Allik, J., Realo, A., & Valk, R. (2009). Age-related differences in emotion recognition ability: A cross-sectional study. *Emotion*, 9(5), 619. <https://doi.org/10.1037/a0016562>
- Mitchell, R. L. C., & Kingston, R. A. (2014). Age-related decline in emotional prosody discrimination: Acoustic correlates. *Experimental Psychology*, 61(3), 215–223. <https://doi.org/10.1027/1618-3169/a000241>
- Mitchell, W. J., Szerszen, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & Macdorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, 2(1), 10–12. <https://doi.org/10.1068/i0415>
- Mithen, S., Morley, I., Wray, A., Tallerman, M., & Gamble, C. (2006). The singing neanderthals: The origins of music, language, mind and body. *Cambridge Archaeological Journal*, 16(1), 97–112. <https://doi.org/10.1017/S0959774306000060>
- Mori, M., Macdorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics & Automation Magazine*, 19(2), 98–100. <https://doi.org/10.1109/mra.2012.2192811>
- Morningstar, M., Gilbert, A. C., Burdo, J., Leis, M., & Dirks, M. A. (2021). Recognition of vocal socioemotional expressions at varying levels of emotional intensity. *Emotion*, 21(7), 1570–1575. <https://doi.org/10.1037/emo0001024>
- Morrison, S. J., & Demorest, S. M. (2009). Cultural constraints on music perception and cognition. *Progress in Brain Research*, 178, 67–77. [https://doi.org/10.1016/S0079-6123\(09\)17805-6](https://doi.org/10.1016/S0079-6123(09)17805-6)
- Mualem, O., & Lavidor, M. (2015). Music education intervention improves vocal emotion recognition. *International Journal of Music Education*, 33(4), 413–425. <https://doi.org/10.1177/0255761415584292>
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS One*, 9(2), e89642. <https://doi.org/10.1371/journal.pone.0101091>
- Naranjo, C., Kornreich, C., Campanella, S., Noël, X., Vandriette, Y., Gillain, B., de Longueville, X., Delatte, B., Verbanck, P., & Constant, E. (2011). Major depression is associated with impaired processing of emotion in music as well as in facial and vocal stimuli. *Journal of Affective Disorders*, 128(3), 243–251. <https://doi.org/10.1016/j.jad.2010.06.039>
- Nashkoff, K. (2007). *The relationship between pitch discrimination skills and speech prosody decoding skills* [Doctoral dissertation, Walden University].
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. <https://doi.org/10.1145/191666.191703>
- Neves, L., Martins, M., Correia, A. I., Castro, S. L., & Lima, C. F. (2021). Associations between vocal emotion recognition and socio-emotional adjustment in children. *Royal Society Open Science*, 8(11), 211412. <https://doi.org/10.1098/rsos.211412>
- Nilsson, Å., & Sundberg, J. (1985). Differences in ability of musicians and nonmusicians to judge emotional state from the fundamental frequency of voice samples. *Music Perception: An Interdisciplinary Journal*, 2(4), 507–516. <https://doi.org/10.2307/40285316>
- Nolden, S., Rigoulot, S., Jolicoeur, P., & Armony, J. L. (2017). Effects of musical expertise on oscillatory brain activity in response to emotional sounds. *Neuropsychologia*, 103, 96–105. <https://doi.org/10.1016/j.neuropsychologia.2017.07.014>
- Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1997). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology*, 2(1), 7–19.
- Nussbaum, C., Pöhlmann, M., Kreysa, H., & Schweinberger, S. R. (2023). *Perceived naturalness of emotional voice morphs [in revision]*.
- Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2022). Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates. *Social Cognitive and Affective Neuroscience*, 17(12), 1145–1154. <https://doi.org/10.1093/scan/nsac033>

- Nussbaum, C., & Schweinberger, S. R. (2021). Links between musicality and vocal emotion perception. *Emotion Review*, 13(3), 211–224. <https://doi.org/10.1177/17540739211022803>
- Nussbaum, C., von Eiff, C. I., Skuk, V. G., & Schweinberger, S. R. (2022). Vocal emotion adaptation aftereffects within and across speaker genders: Roles of timbre and fundamental frequency. *Cognition*, 219, 104967. <https://doi.org/10.1016/j.cognition.2021.104967>
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). Fieldtrip: Open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011, 156869. <https://doi.org/10.1155/2011/156869>
- Palomar-García, M.-Á., Zatorre, R. J., Ventura-Campos, N., Bueichekú, E., & Ávila, C. (2017). Modulation of functional connectivity in auditory-motor networks in musicians compared with nonmusicians. *Cerebral Cortex*, 27(5), 2768–2778. <https://doi.org/10.1093/cercor/bhw120>
- Pan, Z., Liu, X., Luo, Y., & Chen, X. (2017). Emotional intensity modulates the integration of bimodal angry expressions: ERP evidence. *Frontiers in Neuroscience*, 11, 349. <https://doi.org/10.3389/fnins.2017.00349>
- Pantev, C., & Herholz, S. C. (2011). Plasticity of the human auditory cortex related to musical training. *Neuroscience and Biobehavioral Reviews*, 35(10), 2140–2154. <https://doi.org/10.1016/j.neubiorev.2011.06.010>
- Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E., & Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature*, 392(6678), 811–814. <https://doi.org/10.1038/33918>
- Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., & Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *Neuroreport*, 12(1), 169–174.
- Paquette, S., Ahmed, G. D., Goffi-Gomez, M. V., Hoshino, A. C. H., Peretz, I., & Lehmann, A. (2018). Musical and vocal emotion perception for cochlear implants users. *Hearing Research*, 370, 272–282. <https://doi.org/10.1016/j.heares.2018.08.009>
- Park, M., Gutyrchik, E., Welker, L., Carl, P., Pöppel, E., Zaytseva, Y., Meindl, T., Blautzik, J., Reiser, M., & Bao, Y. (2015). Sadness is unique: Neural processing of emotions in speech prosody in musicians and non-musicians. *Frontiers in Human Neuroscience*, 8, 1049. <https://doi.org/10.3389/fnhum.2014.01049>
- Parsons, C. E., Young, K. S., Jegindø, E.-M. E., Vuust, P., Stein, A., & Kringelbach, M. L. (2014). Music training and empathy positively impact adults' sensitivity to infant distress. *Frontiers in Psychology*, 5, 1440. <https://doi.org/10.3389/fpsyg.2014.01440>
- Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA Hypothesis. *Front Psychol*, 2, 142. <https://doi.org/10.3389/fpsyg.2011.00142>
- Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, 87(1), 93–98. <https://doi.org/10.1016/j.biopsycho.2011.02.010>
- Paulmann, S., Bleichner, M., & Kotz, S. A. (2013). Valence, arousal, and task effects in emotional prosody processing. *Front Psychol*, 4, 345. <https://doi.org/10.3389/fpsyg.2013.00345>
- Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain Lang*, 105(1), 59–69. <https://doi.org/10.1016/j.bandl.2007.11.005>
- Paulmann, S., & Kotz, S. A. (2018). The electrophysiology and time course of processing vocal emotion expressions. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 458–472). Oxford University Press. <https://doi.org/10.1093/oxfordhob/9780198743187.013.20>
- Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: How much is enough? *Cogn Affect Behav Neurosci*, 10(2), 230–242. <https://doi.org/10.3758/CABN.10.2.230>

- Pell, M. D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., & Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biological Psychology*, 111, 14–25. <https://doi.org/10.1016/j.biopsycho.2015.08.008>
- Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders: The montreal battery of evaluation of amusia. *Annals of the New York Academy of Sciences*, 999(1), 58–75. <https://doi.org/10.1196/annals.1284.006>
- Peretz, I., Vuvan, D., Lagrois, M.-É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1664), 20140090. <https://doi.org/10.1098/rstb.2014.0090>
- Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, 3, 23. <https://doi.org/10.3389/fpsyg.2012.00023>
- Petersen, B., Mortensen, M. V., Hansen, M., & Vuust, P. (2012). Singing in the key of life: A study on effects of musical ear training after cochlear implantation. *Psychomusicology: Music, Mind, and Brain*, 22(2), 134–151. <https://doi.org/10.1037/a0031140>
- Phillips, L. H., Scott, C., Henry, J. D., Mowat, D., & Bell, J. S. (2010). Emotion perception in Alzheimer's disease and mood disorder in old age. *Psychology and aging*, 25(1), 38. <https://doi.org/10.1037/a0017369>
- Piazza, E. A., Theunissen, F. E., Wessel, D., & Whitney, D. (2018). Rapid adaptation to the timbre of natural sounds. *Sci Rep*, 8(1), 13826. <https://doi.org/10.1038/s41598-018-32018-9>
- Pinheiro, A. P., Vasconcelos, M., Dias, M., Arrais, N., & Gonçalves, Ó. F. (2015). The music of language: An ERP investigation of the effects of musical training on emotional prosody processing. *Brain Lang*, 140, 24–34. <https://doi.org/10.1016/j.bandl.2014.10.009>
- Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science*, 25(5), 679–693. https://doi.org/10.1207/s15516709cog2505_3
- Pralus, A., Fornoni, L., Bouet, R., Gomot, M., Bhatara, A., Tillmann, B., & Caclin, A. (2019). Emotional prosody in congenital amusia: Impaired and spared processes. *Neuropsychologia*, 134, 107234. <https://doi.org/10.1016/j.neuropsychologia.2019.107234>
- Psychology Software Tools, Inc. (2016). *E-prime 3.0*. <https://support.pstnet.com/>.
- R Core Team. (2020). R: A language and environment for statistical computing. <https://www.R-project.org/>
- Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2018). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000481>
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1), 203–212. <https://doi.org/10.1016/j.jrp.2006.02.001>
- Rigoulot, S., Pell, M. D., & Armony, J. L. (2015). Time course of the influence of musical expertise on the processing of vocal and musical sounds. *Neuroscience*, 290, 175–184. <https://doi.org/10.1016/j.neuroscience.2015.01.033>
- Rogenmoser, L., Kernbach, J., Schlaug, G., & Gaser, C. (2018). Keeping brains young with making music. *Brain Structure & Function*, 223(1), 297–305. <https://doi.org/10.1007/s00429-017-1491-2>
- Rothermund, K., & Eder, A. (2011). Emotion. In *Allgemeine Psychologie: Motivation und Emotion* (pp. 165–204). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93420-4_5
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. <https://doi.org/10.1037/h0077714>
- Sauter, D. A. (2017). The nonverbal communication of positive emotions: An emotion family approach. *Emotion Review*, 9(3), 222–234. <https://doi.org/10.1177/1754073916667236>
- Schäfer, T., Sedlmeier, P., Städtler, C., & Huron, D. (2013). The psychological functions of music listening. *Frontiers in Psychology*, 4, 511. <https://doi.org/10.3389/fpsyg.2013.00511>

- Scheiner, E., & Fischer, J. (2011). Emotion expression: The evolutionary heritage in the human voice. In W. Welsch, W. J. Singer & A. Wunder (Eds.), *Interdisciplinary Anthropology* (pp. 105–129). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11668-1_5
- Schelinski, S., Roswandowitz, C., & von Kriegstein, K. (2017). Voice identity processing in autism spectrum disorder. *Autism Res*, 10(1), 155–168. <https://doi.org/10.1002/aur.1639>
- Schellenberg, E. G. (2001). Music and nonmusical abilities. *Annals of the New York Academy of Sciences*, 930(1), 355–371. <https://doi.org/10.1111/j.1749-6632.2001.tb05744.x>
- Schellenberg, E. G. (2016). Music training and nonmusical abilities. *The Oxford Handbook of Music Psychology*, 2, 415–429.
- Schellenberg, E. G., & Mankarious, M. (2012). Music training and emotion comprehension in childhood. *Emotion*, 12(5), 887–891. <https://doi.org/10.1037/a0027971>
- Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychol Bull*, 99(2), 143–165. <https://doi.org/10.1037/0033-2909.99.2.143>
- Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, 9(3), 235–248.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2), 227–256. [https://doi.org/10.1016/S0167-6393\(02\)00084-5](https://doi.org/10.1016/S0167-6393(02)00084-5)
- Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, 27(1), 40–58. <https://doi.org/10.1016/j.csl.2011.11.003>
- Scherer, K. R. (2018). Acoustic patterning of emotion vocalizations. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 60–92). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198743187.013.4>
- Scherer, K. R., Banse, R., & Wallbott, H. G. (2016). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32(1), 76–92. <https://doi.org/10.1177/0022022101032001009>
- Schindler, S., & Bublatzky, F. (2020). Attention and emotion: An integrative review of emotional face processing as a function of attention. *Cortex*, 130, 362–386. <https://doi.org/10.1016/j.cortex.2020.06.010>
- Schindler, S., Zell, E., Botsch, M., & Kissler, J. (2017). Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Scientific Reports*, 7, 45003. <https://doi.org/10.1038/srep45003>
- Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends Cogn Sci*, 21(3), 216–228. <https://doi.org/10.1016/j.tics.2017.01.001>
- Schirmer, A., Chen, C.-B., Ching, A., Tan, L., & Hong, R. Y. (2013). Vocal emotions influence verbal memory: Neural correlates and interindividual differences. *Cognitive, Affective, & Behavioral Neuroscience*, 13(1), 80–93. <https://doi.org/10.3758/s13415-012-0132-8>
- Schirmer, A., Croy, I., & Schweinberger, S. R. (2022). Social touch - a tool rather than a signal. *Current Opinion in Behavioral Sciences*, 44, 101100. <https://doi.org/10.1016/j.cobeha.2021.101100>
- Schirmer, A., & Escoffier, N. (2010). Emotional MMN: Anxiety and heart rate correlate with the erp signature for auditory change detection. *Clinical Neurophysiology*, 121(1), 53–59. <https://doi.org/10.1016/j.clinph.2009.09.029>
- Schirmer, A., Fox, P. M., & Grandjean, D. (2012). On the spatial organization of sound processing in the human temporal lobe: A meta-analysis. *Neuroimage*, 63(1), 137–147. <https://doi.org/10.1016/j.neuroimage.2012.06.025>
- Schirmer, A., & Gunter, T. C. (2017). Temporal signatures of processing voiceness and emotion in sound. *Social Cognitive and Affective Neuroscience*, 12(6), 902–909. <https://doi.org/10.1093/scan/nsx020>
- Schirmer, A., & Kotz, S. A. (2003). ERP evidence for a sex-specific stroop effect in emotional speech. *Journal of Cognitive Neuroscience*, 15(8), 1135–1148.

- Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends Cogn Sci*, 10(1), 24–30. <https://doi.org/10.1016/j.tics.2005.11.009>
- Schirmer, A., Kotz, S. A., & Friederici, A. D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research*, 14(2), 228–233. [https://doi.org/10.1016/S0926-6410\(02\)00108-8](https://doi.org/10.1016/S0926-6410(02)00108-8)
- Schirmer, A., Kotz, S. A., & Friederici, A. D. (2005). On the role of attention for the processing of emotions in speech: Sex differences revisited. *Cognitive Brain Research*, 24(3), 442–452. <https://doi.org/10.1016/j.cogbrainres.2005.02.022>
- Schirmer, A., Striano, T., & Friederici, A. D. (2005). Sex differences in the preattentive processing of vocal emotional expressions. *Neuroreport*, 16(6), 635–639.
- Schneider, P., Sluming, V., Roberts, N., Bleeck, S., & Rupp, A. (2005). Structural, functional, and perceptual differences in heschl's gyrus and musical instrument preference. *Annals of the New York Academy of Sciences*, 1060, 387–394. <https://doi.org/10.1196/annals.1360.033>
- Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., Dosch, H. G., Bleeck, S., Stippich, C., & Rupp, A. (2005). Structural and functional asymmetry of lateral heschl's gyrus reflects pitch perception preference. *Nature Neuroscience*, 8(9), 1241–1247. <https://doi.org/10.1038/nm1530>
- Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology*, 41(3), 341–349. <https://doi.org/10.1111/1469-8986.00172.x>
- Schorr, E. A., Roth, F. P., & Fox, N. A. (2009). Quality of life for children with cochlear implants: Perceived benefits and problems and the perception of single words and emotional sounds. *Journal of Speech, Language, and Hearing Research*. [https://doi.org/10.1044/1092-4388\(2008/07-0213\)](https://doi.org/10.1044/1092-4388(2008/07-0213))
- Schupp, H. T., Öhman, A., Junghöfer, M., Weike, A. I., Stockburger, J., & Hamm, A. O. (2004). The facilitated processing of threatening faces: An ERP analysis. *Emotion*, 4(2), 189–200. <https://doi.org/10.1037/1528-3542.4.2.189>
- Schutz, M. (2017). Acoustic constraints and musical consequences: Exploring composers' use of cues for musical emotion. *Frontiers in Psychology*, 8, 1402. <https://doi.org/10.3389/fpsyg.2017.01402>
- Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., Robertson, D. M., Simpson, A. P., & Zäske, R. (2008). Auditory adaptation in voice perception. *Curr Biol*, 18(9), 684–688. <https://doi.org/10.1016/j.cub.2008.04.015>
- Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *Wiley Interdiscip Rev Cogn Sci*, 5(1), 15–25. <https://doi.org/10.1002/wcs.1261>
- Schweinberger, S. R., Pohl, M., & Winkler, P. (2020). Autistic traits, personality, and evaluations of humanoid robots by young and older adults. *Computers in Human Behavior*, 106, 106256. <https://doi.org/10.1016/j.chb.2020.106256>
- Shahin, A. J., Bosnyak, D. J., Trainor, L. J., & Roberts, L. E. (2003). Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. *The Journal of Neuroscience*, 23(13), 5545–5552. <https://doi.org/10.1523/JNEUROSCI.23-13-05545.2003>
- Shahin, A. J., Roberts, L. E., Chau, W., Trainor, L. J., & Miller, L. M. (2008). Music training leads to the development of timbre-specific gamma band activity. *Neuroimage*, 41(1), 113–122. <https://doi.org/10.1016/j.neuroimage.2008.01.067>
- Shahin, A. J., Roberts, L. E., Pantev, C., Trainor, L. J., & Ross, B. (2005). Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. *Neuroreport*, 16(16), 1781–1785. <https://doi.org/10.1097/01.wnr.0000185017.29316.63>
- Shariff, A. F., & Tracy, J. L. (2011). What are emotion expressions for? *Current Directions in Psychological Science*, 20(6), 395–399. <https://doi.org/10.1177/0963721411424739>

- Sheehan, K. A., McArthur, G. M., & Bishop, D. V. M. (2005). Is discrimination training necessary to cause changes in the p2 auditory event-related brain potential to speech sounds? *Cognitive Brain Research*, 25(2), 547–553. <https://doi.org/10.1016/j.cogbrainres.2005.08.007>
- Singh, P. G., & Hirsh, I. J. (1992). Influence of spectral locus and F0 changes on the pitch and timbre of complex tones. *The Journal of the Acoustical Society of America*, 92(5), 2650–2661. <https://doi.org/10.1121/1.404381>
- Sivathanan, S., Philibert-Lignières, G., & Quintin, E.-M. (2022). Individual differences in autism traits, personality, and emotional responsiveness to music in the general population. *Musicae Scientiae*, 26(3), 538–557. <https://doi.org/10.1177/1029864920988160>
- Skuk, V. G., Dammann, L. M., & Schweinberger, S. R. (2015). Role of timbre and fundamental frequency in voice gender adaptation. *J Acoust Soc Am*, 138(2), 1180–1193. <https://doi.org/10.1121/1.4927696>
- Skuk, V. G., Kirchen, L., Oberhoffner, T., Guntinas-Lichius, O., Dobel, C., & Schweinberger, S. R. (2020). Parameter-specific morphing reveals contributions of timbre and fundamental frequency cues to the perception of voice gender and age in cochlear implant users. *Journal of Speech, Language, and Hearing Research*, 63(9), 3155–3175. https://doi.org/10.1044/2020_JSLHR-20-00026
- Skuk, V. G., Palermo, R., Broemer, L., & Schweinberger, S. R. (2019). Autistic traits are linked to individual differences in familiar voice identification. *Journal of Autism and Developmental Disorders*, 49(7), 2747–2767. <https://doi.org/10.1007/s10803-017-3039-y>
- Skuk, V. G., & Schweinberger, S. R. (2013). Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices. *PLoS One*, 8(11), e81691. <https://doi.org/10.1371/journal.pone.0081691>
- Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *J Speech Lang Hear Res*, 57(1), 285–296. [https://doi.org/10.1044/1092-4388\(2013\)12-0314](https://doi.org/10.1044/1092-4388(2013)12-0314)
- Sorati, M., & Behne, D. M. (2019). Musical expertise affects audiovisual speech perception: Findings from event-related potentials and inter-trial phase coherence. *Frontiers in Psychology*, 10, 2562. <https://doi.org/10.3389/fpsyg.2019.02562>
- Sormaz, M., Young, A. W., & Andrews, T. J. (2016). Contributions of feature shapes and surface cues to the recognition of facial expressions. *Vision Res*, 127, 1–10. <https://doi.org/10.1016/j.visres.2016.07.002>
- Spackman, M. P., Brown, B. L., & Otto, S. (2009). Do emotions have distinct vocal profiles? A study of idiographic patterns of expression. *Cognition and Emotion*, 23(8), 1565–1588. <https://doi.org/10.1080/02699930802536268>
- Spatola, N., & Wudarczyk, O. A. (2021). Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism. *Computers in Human Behavior*, 124, 106934. <https://doi.org/10.1016/j.chb.2021.106934>
- Steinbeis, N., & Koelsch, S. (2011). Affective priming effects of musical sounds on the processing of word meaning. *J Cogn Neurosci*, 23(3), 604–621. <https://doi.org/10.1162/jocn.2009.21383>
- Stewart, L., von Kriegstein, K., Warren, J. D., & Griffiths, T. D. (2006). Music and the brain: Disorders of musical listening. *Brain : A Journal of Neurology*, 129(10), 2533–2553. <https://doi.org/10.1093/brain/awl171>
- Stoet, G. (2010). Psytoolkit: A software package for programming psychological experiments using linux. *Behavior Research Methods*, 42(4), 1096–1104. <https://doi.org/10.3758/BRM.42.4.1096>
- Stoet, G. (2017). Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24–31. <https://doi.org/10.1177/0098628316677643>

- Strait, D. L., Kraus, N., Skoe, E., & Ashley, R. (2009). Musical experience and neural efficiency: Effects of training on subcortical processing of vocal expressions of emotion. *The European Journal of Neuroscience*, 29(3), 661–668. <https://doi.org/10.1111/j.1460-9568.2009.06617.x>
- Sun, L., Thompson, W. F., Liu, F., Zhou, L., & Jiang, C. (2020). The human brain processes hierarchical structures of meter and harmony differently: Evidence from musicians and nonmusicians. *Psychophysiology*, 57(9), e13598. <https://doi.org/10.1111/psyp.13598>
- Tartter, V. C., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America*, 96(4), 2101–2107. <https://doi.org/10.1121/1.410151>
- Tervaniemi, M. (2009). Musicians-same or different? *Ann N Y Acad Sci*, 1169, 151–156. <https://doi.org/10.1111/j.1749-6632.2009.04591.x>
- Thompson, W. F., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica*, 2006(158). <https://doi.org/10.1515/SEM.2006.017>
- Thompson, W. F., Balkwill, L.-L., & Laura-Lee. (2010). Cross-cultural similarities and differences. In Patrik N. Juslin & John Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press.
- Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proc Natl Acad Sci U S A*, 109(46), 19027–19032. <https://doi.org/10.1073/pnas.1210344109>
- Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion*, 4(1), 46–64. <https://doi.org/10.1037/1528-3542.4.1.46>
- Trainor, L. J., Shahin, A. J., & Roberts, L. E. (2009). Understanding the benefits of musical training: Effects on oscillatory brain activity. *Ann N Y Acad Sci*, 1169, 133–142. <https://doi.org/10.1111/j.1749-6632.2009.04589.x>
- Trimmer, C. G., & Cuddy, L. L. (2008). Emotional intelligence, not music training, predicts recognition of emotional speech prosody. *Emotion*, 8(6), 838–849. <https://doi.org/10.1037/a0014080>
- Tursunov, A., Kwon, S., & Pang, H.-S. (2019). Discriminating emotions in the valence dimension from speech using timbre features. *Applied Sciences*, 9(12), 2470. <https://doi.org/10.3390/app9122470>
- Twaite, J. (2016). *Examining relationships between basic emotion perception and musical training in the prosodic, facial, and lexical channels of communication and in music* [Doctoral dissertation].
- Vincenzi, M., Correia, A. I., Vanzella, P., Pinheiro, A. P., Lima, C. F., & Schellenberg, E. G. (2022). Associations between music training and cognitive abilities: The special case of professional musicians. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10.1037/aca0000481>
- Vojtech, J. M., Noordzij, J. P., Cler, G. J., & Stepp, C. E. (2019). The effects of modulating fundamental frequency and speech rate on the intelligibility, communication efficiency, and perceived naturalness of synthetic speech. *American Journal of Speech-language Pathology*, 28(2S), 875–886. https://doi.org/10.1044/2019_AJSLP-MS18-18-0052
- von Eiff, C. I., Skuk, V. G., Zäske, R., Nussbaum, C., Frühholz, S., Feuer, U., Guntinas-Lichius, O., & Schweinberger, S. R. (2022). Parameter-specific morphing reveals contributions of timbre to the perception of vocal emotions in cochlear implant users. *Ear & Hearing*, 43(4), 1178–1188. <https://doi.org/10.1097/AUD.0000000000001181>
- Waaramaa, T., Kukkonen, T., Mykkänen, S., & Geneid, A. (2018). Vocal emotion identification by children using cochlear implants, relations to voice quality, and musical interests. *Journal of Speech, Language, and Hearing Research*, 61(4), 973–985. https://doi.org/10.1044/2017_JSLHR-H-17-0054

- Waaramaa, T., Laukkanen, A.-M., Alku, P., & Väyrynen, E. (2008). Monopitched expression of emotions in different vowels. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, 60(5), 249–255. <https://doi.org/10.1159/000151762>
- Waaramaa, T., & Leisiö, T. (2013). Perception of emotionally loaded vocal expressions and its connection to responses to music. A cross-cultural investigation: Estonia, Finland, Sweden, Russia, and the USA. *Frontiers in Psychology*, 4, 344. <https://doi.org/10.3389/fpsyg.2013.00344>
- Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, 17(1), 3–28. <https://doi.org/10.1007/BF00987006>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, 6(4), 647–653. <https://doi.org/10.3758/BF03212974>
- Weijkamp, J., & Sadakata, M. (2017). Attention to affective audio-visual information: Comparison between musicians and non-musicians. *Psychology of Music*, 45(2), 204–215. <https://doi.org/10.1177/0305735616654216>
- Wenhart, T., & Altenmüller, E. (2019). A tendency towards details? inconsistent results on auditory and visual local-to-global processing in absolute pitch musicians. *Frontiers in Psychology*, 10, 31. <https://doi.org/10.3389/fpsyg.2019.00031>
- Wenhart, T., Bethlehem, R. A. I., Baron-Cohen, S., & Altenmüller, E. (2019). Autistic traits, resting-state connectivity, and absolute pitch in professional musicians: Shared and distinct neural features. *Molecular Autism*, 10, 20. <https://doi.org/10.1186/s13229-019-0272-6>
- Whiting, C. M., Kotz, S. A., Gross, J., Giordano, B. L., & Belin, P. (2020). The perception of caricatured emotion in voice. *Cognition*, 200, 104249. <https://doi.org/10.1016/j.cognition.2020.104249>
- Wiese, H., Komes, J., Tutenberg, S., Leidinger, J., & Schweinberger, S. R. (2017). Age-related differences in face recognition: Neural correlates of repetition and semantic priming in young and older adults. *J Exp Psychol Learn Mem Cogn*, 43(8), 1254–1273. <https://doi.org/10.1037/xlm0000380>
- Yamasaki, R., Montagnoli, A., Murano, E. Z., Gebrim, E., Hachiya, A., Lopes da Silva, J. V., Behlau, M., & Tsuji, D. (2017). Perturbation measurements on the degree of naturalness of synthesized vowels. *Journal of Voice*, 31(3), 389.e1–389.e8. <https://doi.org/10.1016/j.jvoice.2016.09.020>
- Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2013). Voice quality in affect cueing: Does loudness matter? *Frontiers in Psychology*, 4, 335. <https://doi.org/10.3389/fpsyg.2013.00335>
- Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2018). Cross-language differences in how voice quality and F0 contours map to affect. *The Journal of the Acoustical Society of America*, 144(5), 2730. <https://doi.org/10.1121/1.5066448>
- Yorkston, K. M., Beukelman, D. R., Strand, E. A., & Hakel, M. (1999). *Management of motor speech disorders in children and adults*. Pro-ed Austin, TX.
- Yorkston, K. M., Hammen, V. L., Beukelman, D. R., & Traynor, C. D. (1990). The effect of rate control on the intelligibility and naturalness of dysarthric speech. *The Journal of Speech and Hearing Disorders*, 55(3), 550–560. <https://doi.org/10.1044/jshd.5503.550>
- Young, A. W., & Bruce, V. (2011). Understanding person perception. *Br J Psychol*, 102(4), 959–974. <https://doi.org/10.1111/j.2044-8295.2011.02045.x>
- Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends Cogn Sci*, 24(5), 398–410. <https://doi.org/10.1016/j.tics.2020.02.001>

- Young, K. S., Parsons, C. E., Stein, A., & Kringelbach, M. L. (2012). Interpreting infant vocal distress: The ameliorative effect of musical training in depression. *Emotion*, 12(6), 1200–1205. <https://doi.org/10.1037/a0028705>
- Zäske, R., & Schweinberger, S. R. (2011). You are only as old as you sound: Auditory aftereffects in vocal age perception. *Hear Res*, 282(1-2), 283–288. <https://doi.org/10.1016/j.heares.2011.06.008>
- Zäske, R., Schweinberger, S. R., Kaufmann, J. M., & Kawahara, H. (2009). In the ear of the beholder: Neural correlates of adaptation to voice gender. *Eur J Neurosci*, 30(3), 527–534. <https://doi.org/10.1111/j.1460-9568.2009.06839.x>
- Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hear Res*, 268(1-2), 38–45. <https://doi.org/10.1016/j.heares.2010.04.011>
- Zäske, R., Skuk, V. G., Kaufmann, J. M., & Schweinberger, S. R. (2013). Perceiving vocal age and gender: An adaptation approach. *Acta Psychologica*, 144(3), 583–593. <https://doi.org/10.1016/j.actpsy.2013.09.009>
- Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the short-proms and the mini-proms. *Annals of the New York Academy of Sciences*, 1400(1), 33–45. <https://doi.org/10.1111/nyas.13410>
- Zhang, Y., Geng, T., & Zhang, J. (2018). Emotional prosody perception in mandarin-speaking congenital amusics. *Interspeech*, 2196–2200. <https://doi.org/10.21437/Interspeech.2018-91>

E. Danksagung

Es gibt viele Menschen, die mir die erfolgreiche Arbeit an dieser Dissertation ermöglicht haben und bei denen ich mich an dieser Stelle ganz herzlich bedanken möchte. Zuallererst bedanke ich mich bei Stefan Schweinberger und Annett Schirmer für ihre exzellente Betreuung während der vergangenen vier Jahre. Ich habe von euch beiden unglaublich viel gelernt! Stefan, deine Tür war immer offen und ich konnte mit jeder Frage auf dich zukommen und deinen Rat einholen. Du hast mir sehr viele Freiheiten bei der thematischen Ausrichtung meiner Arbeit ermöglicht und meine Forschungsinteressen unterstützt. Du hast mir dabei nicht nur beigebracht, was gute Forschung ausmacht, sondern auch, wie man sie gut kommuniziert. Vor allem aber habe ich von dir neben der fachlichen Unterstützung auch Ermutigung in den schweren Phasen erfahren. Dadurch habe ich gelernt, wieder an mich selbst zu glauben und dafür bin ich dir ausgesprochen dankbar. Annett, ich danke dir für die enge Zusammenarbeit trotz der großen räumlichen Distanz! Du hast dir für meine Arbeit unglaublich viel Zeit genommen - sei es bei der Erschließung neuer Analysemethoden, Feedback zu meinen Texten, oder das detaillierte Ausdiskutieren aller meiner Fragen - was manchmal sicher ein wenig Geduld mit mir erfordert hat. Du hast mich fachlich gefordert, mir aber gleichzeitig den Rücken gestärkt und mich so zu einer besseren Wissenschaftlerin gemacht. Dafür bin ich dir ausgesprochen dankbar.

Ich bedanke mich bei unserem tollen Team. Helene, danke für dein offenes Ohr, wann immer ich es brauchte, und deine unverzichtbare Hilfe mit den vielen administrativen Aufgaben, wovon ich dir viele gern erspart hätte. Manuel, ich danke dir zum einen für das detaillierte Gegenlesen meiner Texte, und zum anderen für deinen einzigartigen Humor, mit dem du mich in den letzten Monaten immer wieder aufgebaut hast. Celina, du bist eine ganz wundervolle Bürokollegin. Ich bin dankbar für unsere erfolgreiche Zusammenarbeit, für die vielen geteilten Keksrollen und natürlich für deine Hilfe beim Korrekturlesen. Linda, ich danke dir für deine hilfreichen Rückmeldungen zu meiner Arbeit, inklusive der beachtlichen Kollektion thematisch abgestimmter Memes. Du bringst mich immer wieder zum Lachen und das schätze ich an dir. Ich bedanke mich bei unseren technischen Assistentinnen Kathrin und Bettina für eure Unterstützung mit der EEG-Datenerhebung. Ich hätte das ohne euch nicht geschafft! Ich danke euch auch für euren achtsamen Blick auf unsere Gruppe, mit euren vielen kleinen Handgriffen und Erfahrungen, womit ihr uns tagtäglich reibungslose Abläufe ermöglicht.

Ich danke Jürgen, der bei mir vor vielen Jahren die Begeisterung für die Erforschung der menschlichen Wahrnehmung geweckt hat (ja, das ist tatsächlich fast zehn Jahre her). Ich danke Verena, dass sie ihr umfangreiches Wissen zu Stimmenwahrnehmung, Voice Morphing und R-Scripten mit mir geteilt hat. Ich danke Romi, von der ich mir das *Oxford Handbook of Voice*

Perception so oft geliehen habe, dass sie es mir schließlich ganz überlassen hat. Ich danke Andrea für die produktive Zusammenarbeit an der Ratingstudie. Desweiteren danke ich Samaneh, Ayaka, Ulrike, Julian und allen weiteren Mitgliedern des PhD-Clubs für die wunderbare Kollegialität, eure Kreativität und die inspirierenden Eindrücke aus euren Projekten. Ich danke Johannes und Laura, die diese Forschung durch ihre Abschlussarbeiten bereichert haben. Ich danke Adrian Simpson, Christian Dobel, Sven Kachel und César Lima für wertvolle Rückmeldungen zu meiner Arbeit. Ich danke außerdem den hilfsbereiten Sekretariaten an der Hochschule für Musik in Weimar, der Musik- und Kunstschule Jena sowie in den Dekanaten der Friedrich-Schiller-Universität Jena, die mich bei der Rekrutierung der Teilnehmenden tatkräftig unterstützt haben, und selbstverständlich allen StudienteilnehmerInnen. Ich danke der Studienstiftung des Deutschen Volkes für die Finanzierung meiner Promotion.

Ich möchte mich auch für die vielfältige Unterstützung durch meine Freunde bedanken. Ohne euch hätte ich im Lockdown vermutlich den Verstand verloren. Ich danke euch – Tonia, Till, Susan, Sven und Sebastian – dafür, dass man mit euch einfach alles machen kann: Musizieren, Tischtennis spielen, Bouldern gehen, Film schauen, Corona kriegen, Urlaub machen - aber auch Probleme und Sorgen teilen. Ihr seid für mich besonders in den letzten Monaten eine unverzichtbare Stütze gewesen. Ich danke meinen Kommilitonen, darunter Jule und Alex, mit denen ich fünf tolle Jahre studiert habe. Natürlich haben noch viel mehr Personen zum Erfolg dieser Arbeit beigetragen, die hier nicht namentlich genannt sind. Allen, die sich dabei angesprochen fühlen, gilt mein Dank.

Ein besonderer Dank gilt meiner Familie. Danke an meine Eltern, Rhena und Hans, dass ihr immer an mich glaubt und mich auf meinem Weg unterstützt, wo ihr könnt, auch wenn das nicht immer leicht ist (besonders bei Biber-bedingten Bruchlandungen...). Dasselbe gilt für meine Schwester Phia. Ich glaube, die letzten Monate hätten wir ohne einander nicht geschafft. Ich bin so froh, dass es dich gibt und ich bin unglaublich stolz auf dich! Ich danke auch meinem Bruder Hans, einem der klügsten Köpfe, den ich kenne. Bei deiner Verteidigung vor sieben Jahren hast du zu mir gesagt: "So, jetzt bist du dran." Bitte sehr! Ich danke natürlich auch allen anderen Mitgliedern meiner Familie, inklusive der angeheirateten - ihr seid der Jackpott! Schlussendlich möchte ich mich bei meinem Mann Frank bedanken. Du stehst mir in allen Lebenslagen zur Seite, durch alle Höhen und Tiefen und durch alle Erfolge und Rückschläge. Du bist die wichtigste Stütze in meinem Leben - und außerdem machst du ausgezeichnetes Frühstück! Ich liebe dich. Danke für alles!

F. Ehrenwörtliche Erklärung

Ich, Christine Nussbaum, bestätige hiermit, dass mir die geltende Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften bekannt ist. Ich habe die Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben. Ich habe angegeben, welche Personen mich bei der Auswahl und Auswertung des Materials sowie bei der Herstellung von Manuskripten unterstützten und welche Personen an den bereits publizierten Teilen dieser Dissertationen als Ko-AutorInnen mitgewirkt haben. Ich bestätige weiterhin, dass die Hilfe eines kommerziellen Promotionsvermittlers nicht in Anspruch genommen wurde und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Diese Dissertation wurde von mir weder als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht, noch habe ich die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht.

Ort, Datum

Unterschrift