Tampere University

TUA HAKANPÄÄ

Emotion expression in the singing voice

Testing a parameter modulation technique for improving communication of emotions through voice qualities

Tampere University Dissertations 558

Tampere University Dissertations 558

TUA HAKANPÄÄ

Emotion expression in the singing voice

Testing a parameter modulation technique for improving communication of emotions through voice qualities

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Social Sciences
of Tampere University,
for public discussion in the Lecture room K103
of the Linna Building, Kalevantie 5, Tampere,
on April 1st 2022, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Social Sciences
Finland

Responsible supervisor or/and Custos	Professor Anne-Maria Laukkanen Tampere University Finland	
Supervisor	Associate Professor Teija Waaramaa Tampere University and University of Vaasa Finland	
Pre-examiners	Assistant Professor Filippa M.B. Lã Universidad Nacional de Educación a Distancia Spain	Professor Juha Ojala Sibelius Academy, University of the Arts Helsinki Finland
Opponent	Professor Allan Vurma Estonian Academy of Music and Theatre Estonia	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-2304-2 (print)
ISBN 978-952-03-2305-9 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-2305-9

PunaMusta Oy – Yliopistopaino
Joensuu 2022

ABSTRACT

This study examines emotional expression in singing and its teachability using a novel parameter modulation technique. The work is an experimental comparative study using listener evaluations, acoustic analyses, and statistical deduction to assess the emotional expressiveness of the singing voice from short vocal samples and sung phrases. The investigation consists of three sub-studies, the first of which explores the auditory recognition of emotion from samples sung with Classical and non-Classical singing techniques at three different pitches. The second study compares the qualitative features of emotional expression in Classical and non-Classical singing techniques by means of acoustic analysis. The third sub-study focuses on teaching the parameter modulation technique to acting students. It compares the clarity of emotional expression between the instructional and control groups before and after the training intervention. The measures of emotional expression clarity in this study are considered to be the auditory recognition of emotional expression and the qualitative variation of the voice between different emotional expressions.

The study involved 29 (Study I) and 32 (Study III) listeners of sound samples, 11 female singers (six with Classical singing technique training and five with popular music singing technique training) (Studies I & II), two male singers (one with Classical singing technique training and one with popular singing technique training) (Study I), and six + six acting students who gave song samples, one group of whom participated in the parameter modulation training while the other group received standard singing training (Study III). Listeners were to classify samples into neutral expressions and expressions of joy, tenderness, sadness, and anger from short vowel samples and phrases. The emotions were chosen because of their opposite positioning on the valence-activation scale. Singers sang spontaneous emotional expression into short melodies (16 bars in Studies I & II & 8 bars in Study III) from which sound samples were cut. Samples of the sung [a:] vowel were analyzed with the Praat sound analysis program. The samples were analyzed for fundamental frequency (*fo*), sound pressure level (SPL), formant frequencies (F1-F5), harmonics-to-noise ratio (HNR), energy ratio between upper and lower frequencies of the spectrum (Alpha ratio), irregular cycle-to-cycle variation of fundamental frequency

(Jitter rap & ppq5), irregular cycle-to-cycle variation of amplitude (apq3 & apq5), vibratos (*fo* rate and extent & rate and extent of amplitude), and amplitude contour: attack, sustain, release.

The results of the study showed that emotional expression can be identified from the singing voice when singers express emotion spontaneously (Studies I & III). In this study, the identification of emotional expression became easier after singers received instructions on the use of the parameter modulation technique (Study III). Emotional expression was better identified from song samples sung in a non-Classical style (Study I). Pitch, SPL, emotional valence (positive / negative), and activation level (high / low) had an effect on emotional recognition (Studies I-III). SPL, Alpha ratio, and HNR values increased in expressions of high activity emotions (anger and joy) and decreased in expressions of low activity emotions (sadness and tenderness), suggesting increased muscle activity and tighter vocal fold adduction in high energy emotions (Studies II & III). Formants packed in high-energy emotions and scattered in low-energy emotions, suggesting a modification of the vocal tract for the expression of different emotions (Studies II & III). Jitter and shimmer were more prevalent in low-energy emotions, suggesting lower muscle activity (Study II). Fundamental frequency vibrato was slower in Classically trained singers (Study II), whereas in non-Classical singers, amplitude vibrato was statistically significant in differentiating emotions (Study II). Vocal offsets were statistically significant in terms of emotional expression in singers singing with a non-Classical singing technique (Study II).

The main question of the study was whether it is possible to integrate vocological research data on the acoustic parameters of emotional expression into practical singing exercises and thus enhance emotional expression in the singing voice. In the study, we used a seven-week training program focusing on parameter modulation techniques that taught the use of different sound qualities to a group of acting students. A similar group of acting students who did not receive special training served as a control group. The test group increased the use of different sound qualities as a means of emotional expression after training. This result was confirmed by acoustic analyses and improved recognition of emotions by the listeners. The control group did not show such an effect. After training, the test team appeared to use F1, SPL, HNR, and alpha ratio for emotional expression more systematically.

The study showed that the sound pressure level and the way energy is distributed in the sound spectrum were the two most typical indicators of the emotional characteristics of sound. The study finds that training in different sound qualities can help with the expression of emotions in the singing voice.

TIIVISTELMÄ

Tässä tutkimuksessa tarkastellaan tunneilmaisua lauluäänessä ja sen opetettavuutta parametrimodulaatiotekniikan avulla. Tutkimus on kokeellinen vertailututkimus, jossa käytetään kuuntelijoiden arvioita, akustisia analyysejä ja tilastollista päättelyä lauluäänen tunneilmaisevuuden arvioimiseksi lyhyistä vokaalinäytteistä ja lauletuista fraaseista. Tutkimus koostuu kolmesta osatutkimuksesta, joista ensimmäisessä selvitetään kuulonvaraista tunteen tunnistamista klassisella ja ei-klassisella laulutekniikalla lauletuista näytteistä kolmelta eri sävelkorkeudelta laulettuna. Toinen tutkimus vertailee tunneilmaisun akustisia parametreja klassisessa ja ei-klassisessa laulutekniikassa akustisen analyysin keinoin. Kolmas osatutkimus keskittyy parametrimodulaatio-tekniikan opettamiseen näyttelijäntyön opiskelijoille. Siinä vertaillaan tunneilmaisun selkeyttä opetusta saavan ja verrokkiryhmän välillä ennen ja jälkeen koulutusintervention. Tunneilmaisun selkeyden mittareina tässä tutkimuksessa pidetään tunneilmaisun kuulonvaraista tunnistamista ja ääniparametrien vaihtelua eri tunneilmaisujen välillä.

Tutkimukseen osallistui 29 (tutkimus I) ja 32 (tutkimus III) ääninäytteiden kuuntelijaa, 11 naislaulajaa (6 klassisen laulutekniikan koulutuksen ja 5 populaarimusiikin laulutekniikan koulutuksen saaneita) (tutkimukset I & II), 2 mieslaulajaa (1 klassisen laulutekniikan koulutuksen ja 1 populaarimusiikin laulutekniikan koulutuksen saanut)(tutkimus I) sekä 6 + 6 laulunäytteitä antanutta näyttelijäopiskelijaa, joista toinen ryhmä osallistui parametrimodulaatiokoulutukseen ja toinen ryhmä sai tavanomaista laulukoulutusta (tutkimus III). Kuuntelijat tunnistivat neutraaleja ilmaisuja ja ilon, lempeyden, surun ja vihan tunteiden ilmaisua lyhyistä vokaalinäytteistä ja fraaseista. Laulajat ilmaisivat tunteita lyhyisiin (16-tahtia tutkimuksissa I & II & 8-tahtia tutkimuksessa III) melodioihin, joista ääninäytteet leikattiin. Pitkät [a:]- vokaalinäytteet analysoitiin Praat-äänenanalyysiohjelmalla. Äänestä mitattiin perustaajuus (*fo*), äänenpainetaso (SPL), formanttitaajuudet (F1-F5), hälyn suhde periodiseen ääneen (HNR), energian suhde spektrin ylempien ja alempien taajuuksien välillä (Alpha ratio), epäsäännöllinen syklinen variaatio perustaajuudessa (Jitter rap & ppq5), epäsäännöllinen syklinen variaatio amplitudissa (Shimmer apq3 ja apq5), vibratot (fo -vaihtelun taajuus ja laajuus & amplitudivaihtelun taajuus ja laajuus) sekä amplitudikontuurin muoto (äänen

voimakkuuskäyrän muoto vokaalin aikana): isku, pidätys ja haipuminen (attack, sustain, release).

Tutkimuksen tulokset osoittivat, että tunneilmaisu on mahdollista tunnistaa lauluäänestä, kun laulajat ilmaisevat tunnetta (tutkimukset I & III). Tässä tutkimuksessa tunneilmaisun tunnistaminen tuli helpommaksi sen jälkeen, kun laulajat saivat ohjeet parametrimodulaatiotekniikan käytöstä (tutkimus III). Tunneilmaisu tunnistettiin paremmin ei-klassisella tyylillä lauletuista laulunäytteistä (tutkimus I). Äänenkorkeudella, äänenpaineen tasolla sekä tunteen valenssilla (positiivinen/negatiivinen) ja aktivaatiotasolla (korkea/matala) oli vaikutusta tunteen tunnistamiseen (tutkimukset I-III). SPL, Alpha ratio ja HNR arvot kohosivat korkean aktiviteetin tunteissa (viha ja ilo) ja laskivat matalan aktiviteetin tunteissa (suru ja lempeys), mikä viittaa suurempaan lihasaktiivisuuteen ja tiukempaan äänihuulisulkuun korkean energian tunteissa (tutkimukset II & III). Formantit pakkautuivat korkean energian tunteissa ja sirottuivat matalan energian tunteissa, joka viittaa ääntöväylän muokkaukseen tunneilmaisussa (tutkimukset II & III). Jitteriä ja shimmeriä esiintyi enemmän matalan energian tunteissa, joka viittaa matalampaan lihasaktiivisuuteen (tutkimus II), fo-vibrato oli hitaampaa klassisesti koulutetuilla laulajilla (tutkimus II), kun taas ei-klassisilla laulajilla amplitudivibrato erotteli tunteita (tutkimus II). Äänen lopetukset olivat tilastollisesti merkitseviä tunneilmaisun kannalta ei-klassisella laulutekniikalla laulavilla laulajilla (tutkimus II).

Tutkimuksen pääkysymys oli, onko mahdollista integroida vokologista tutkimustietoa tunneilmaisun akustisista parametreista laulunopetukseen ja sillä tavoin tehostaa tunneilmaisua lauluäänessä. Käytimme tutkimuksessa parametrimodulaatiotekniikkaan keskittyvää seitsemän viikon harjoitusohjelmaa, jossa opetettiin erilaisten äänenlaatujen käyttöä näyttelijäopiskelijaryhmälle. Samanlainen näyttelijäopiskelijaryhmä, joka ei saanut erityiskoulutusta, toimi kontrolliryhmänä. Testiryhmä lisäsi erilaisten äänenlaatujen käyttöä tunneilmaisun välineenä koulutuksen jälkeen. Tämä tulos vahvistettiin kuuntelijoiden arvioinneilla ja akustisilla analyyseillä. Tällaista vaikutusta ei näkynyt kontrolliryhmällä. Koulutuksen jälkeen testiryhmä näytti käyttävän taktisesti systemaattisemmin ensimmäistä formanttitaajuutta, äänenpainetasoa, hälyn määrää äänessä ja Alpha ratiota tunneilmaisuun.

Tutkimus osoitti, että äänenpainetaso ja tapa, jolla energia jakautuu äänispektrissä, olivat kaksi tyypillisintä äänen tunnepiirteiden indikaattoria. Tutkimuksessa todetaan, että erilaisten äänenlaatujen kouluttaminen voi auttaa ilmaisemaan tunteita lauluäänessä.

CONTENTS

List of Tables

ABBREVIATIONS

Alpha ratio	Difference in sound pressure level between the ranges 1500-5000 Hz and 50-1500 Hz (a measure of spectral slope)
ANSI	the American national standards institute
bpm	Beats per minute
dB	Decibel
CCM	Contemporary Commercial Music
cm H_2O	Centimeters of Water Column (manometric (pressure) unit)
CPM	Component process model (of emotion)
CQ_{EGG}	Electroglottographic contact quotient
ERV	Expiratory reserve volume
F1-F5	Formant frequencies
fo	Fundamental frequency
f_{R1}-f_{R5}	Vocal tract resonances
FRC	Functional residual capacity
H1-H2	Lowest harmonic partials
HNR	Harmonics-to-noise ratio
Hz	Hertz
IC	Inspiratory capacity
IRV	Inspiratory reserve volume
Jitter	Period-to-period variation in *fo*
kHz	Kilohertz (1000 Hz)
MFCC	Mel-frequency cepstral coefficient
MFDR	Maximum flow declination rate
MRI	Magnetic resonance imaging
NS	Nervous system
Pa	Pascal
P_m	Intraoral pressure
P_{sub}	Subglottal pressure
PTP	Phonation threshold pressure
RV	Residual volume
SPL	Sound pressure level
TLC	Total lung capacity
TV	Tidal volume
VC	Vital capacity

GLOSSARY

Activity/arousal	A state that is characterized by increased or decreased physiological activity in the body (the physiological and/or subjective intensity of emotion)
Affect	An umbrella term that covers all evaluative or valenced states (emotion, mood, preference)
Attack	The phase when the amplitude of the voice sample rises and reaches its peak
Emotion	Short and intensive affective reaction that typically involves a number of more or less synchronized sub-components such as subjective feeling, physiological arousal, expression, action tendency, and regulation. Emotions focus on specific objects and typically last from minutes to a few hours
Feeling	A subjective experience of emotions and moods (typically measured via self-report)
Formant	Peak of enhanced spectral energy in the output spectrum
Mood	Affective state that is lower in intensity than emotion, does not have a clear object, and lasts considerably longer than emotions
Musical emotion	An emotion that is somehow induced by music, without any further implication about the precise nature of these emotions
Off-set	A decrease of the signal amplitude until silent
On-set	Time interval between the release of a plosive and the beginning of vocal fold vibration associated with the subsequent vowel
Release	A decrease of the signal amplitude until silent
Resonance	Reinforced oscillation at the natural frequency of the vocal tract
Resonant voice	Refers to the interaction effects between the vocal tract and vocal fold vibration
Shimmer	Irregular variation of the period amplitude
SPL	Sound pressure level
Sustain	Constant amplitude phase in an amplitude envelope
Timbre	Perceived sound quality of a musical tone
Valence	The affective quality referring to the intrinsic attractiveness or averseness of an event, object, or situation
Voice color	Sound energy distribution along the frequencies resulting in a dark-bright perception of vocal timbre
Voice quality	The characteristic auditory coloring of an individual voice, which emerges as the conjoined function of the voice source and the vocal tract

ORIGINAL PUBLICATIONS

Publication I Hakanpää, T., Waaramaa, T., & Laukkanen, A. (2019). Emotion Recognition From Singing Voices Using Contemporary Commercial Music and Classical Styles. *Journal of Voice*, *33*(4), 501–509. https://doi.org/10.1016/j.jvoice.2018.01.012

Publication II Hakanpää, T., Waaramaa, T., & Laukkanen, A. (2021). Comparing Contemporary Commercial and Classical Styles: Emotion Expression in Singing. *Journal of Voice*, 35(4)570-580 https://doi.org/10.1016/j.jvoice.2019.10.002

Publication III Hakanpää, T., Waaramaa, T., & Laukkanen, A. (2021). Training the Vocal Expression of Emotions in Singing: Effects of Including Acoustic Research-Based Elements in the Regular Singing Training of Acting Students. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.12.032

1 INTRODUCTION

The expression of emotions in singing has traditionally been taught using different mind imagery exercises and techniques adapted from theatre work. A lot of emphasis is given to the song lyrics and their analyses. Singing instructors think about things like, *who is singing according to the text? What does this person want to say? Are there other people involved in this scenario? Where is it happening?* etc. We tend to think that working through these things with our students (talking about the story of the song, discussing the meaning it has to them and how they would like to present this idea that they have of the song) will magically make their performance better – and oddly enough it so happens in most cases. However, there is a small percentage of students that does not respond well to this kind of work. They are the students that say: *"I can't picture myself in this situation, I have no imagination," "So what exactly do you want me to do?"* or simply *"I don't know."* This study tries to provide a solution for these students. It looks at vocal emotion expression from inside the voice, analyzing it parameter by parameter and piecing it back together to form emotional voice qualities. To put it simply, it is an engineer's way to arrive at emotional expression in singing. To put it in a language of a singing teacher, it is a different route to the same destination. And to put it academically, it is an investigation into the effects of including acoustic research-based elements in training the vocal expression of emotions in singing.

In the teaching of singing, inference and abductive reasoning are the main forms of inquiry. All voice teachers operate by gathering auditory information about the singing voice into their own personal knowledge bank and then use that bank to make generalizations about a certain voice. Singing teachers can, for example, recognize a hyperfunctional voice by comparing the quality of the said voice to all the voices they have heard before. They then dive into their treasure chest of voice exercises and pull out an exercise that might help the student to lessen the strain. (This latter part of the process is hypothetico-deductive.) What I am trying to do in my work is to translate this intuitive process into the language of scientific inquiry and in that way bridge the gap between these disciplines. My research is about different sound qualities in the singing voice, which means that its practical implications in the voice studio will fall on the teaching of vocal technique(s). The

premise of my work lies in the notion that voice quality changes when expressing emotions (Juslin & Laukka, 2003; Scherer, Sundberg, Fantini, Trznadel, & Eyben, 2017; Sundberg, 1987). Therefore, I hypothesize that it should be possible to teach sound quality changes to aid vocal emotion expression and its recognition.

This study, in essence, is a practitioner enquiry, because my research question stands at the intersection of theory and practice (Robbins, 2014) and furthermore I am the teacher devising and teaching the parameter modulation technique to the students taking part in this study. The parameter modulation technique itself has been developed from practical grounds through observations and dilemmas pertaining to teaching emotion expression and voice quality changes in singing to novice students.

The term "voice quality" refers to the characteristic auditory coloring of an individual voice. It emerges as the joint function of the voice source and the vocal tract. Emotions change the habitual coloring of the voice to enhance communication and help the individual adapt to different situations, but it is also possible to deliberately change one's voice color to make the message one is trying to send clearer. The perceptual characteristics of voice quality may sound different to listeners from different cultural and aesthetic backgrounds, but certain tones of voice are recognized similarly across the world (Scherer, Trznadel, Fantini, & Sundberg, 2017). In this study, I view voice quality as an auditory-perceptual phenomenon with causal relations to the anatomical and physiological systems of voice production, which can be measured from the acoustic signal using parameter reduction.

My classroom functions as an inquiry site for intentional and systematic inquiry of my own teaching and students' learning, and in that way my research positions itself as teacher research (Robbins, 2014). I have been brought up by (and working in) the Finnish music education system for most of my life. From this point of view the basic assumptions that I have of what music and singing is and how they are taught, are largely informed by the policies of the national core curriculums for Finnish music education at the time of my own studies and through the curriculums of the schools that I've taught in[1]. It is fair to acknowledge that music transmission and learning are fundamentally social achievements. Musicians, teachers and students of music engage in cognitive, affective and kinetic operations that are informed by our participation in broader spheres of human culture (Szego, 2002). In

[1] The curriculum system refers to the overall curriculum, which is devised based on the Act on Basic Art Education (632/1998), Vocational education and training act (531/2017), Polytechnics Act (932/2014) and Universities Act (558/2009), (632/1998) at the time of relevance, the basics of the curriculum for music issued by the National Board of Education and the local curricula prepared on the basis thereof.

this research however I adopt a slightly narrower viewpoint and discuss the subtle quality differences of vocal techniques.

In this study I am operating in the realm of western music and under the umbrella terms of classical singing (technique) and contemporary commercial music singing (technique). By classical singing I refer broadly to Western lyric music, which is largely a written form, whose sub-genres are well established and clearly defined by composer, country, or era. By contemporary commercial music (CCM) I refer to genres based on an aural tradition, where the material is passed on by ear or recording/video. CCM musical scores rarely represent the notes performed by the original singer (Fisher, Kayes, & Popeil, 2019). More than the specific song genres however, I'm interested in the changes that occur in vocal technique(s) when expressing emotion. I define singing technique as a systematic way of using the singing voice acoustically and physiologically in a way that satisfies the aesthetic demands of a sung music genre while simultaneously being mindful of the individual anatomy and physiology (of the singer) to produce the vocal sounds economically and in a way that does not harm the body.

My specific research area provides both deductive-nomological and inductive-statistical explanations. I do statistical deduction from the obtained data asking the question "does the quality of voice change when using emotional expression?" in different experimental settings. This means that my results are inductive-statistical. To justify my measuring endeavors, I lean on deductive-nomological explanations about the singing voice, such as resonance, sound spectrum, and source-filter interaction. I argue that because there are certain laws of physics in action in the larynx and pharynx that result in a certain kind of sound, I can trace my measurements to these phenomena and ask questions about why the sound changes if the laws governing the sound production stay the same. As I cannot really pinpoint a causal interaction based on physical phenomena, emotional expression, and sound perception, I use statistics to offer an educated guess.

My theoretical thinking comes close to methodological instrumentalism. I view theory as a general conception that has resulted from rational or intellectual activity. It is a tool that transcends perceptions and helps to systematize them. In this study, I use Shannon and Weaver's (1949) theory of communication to align my experimental design, the source-filter theories to validate the acoustic measurements produced and analyzed, and the Brunswik/Scherer tripartite model to explain the acoustic communication of emotion (Bänziger, Hosoya, & Scherer, 2015; Brunswik, 1956; Fant, 1970; Scherer, 1986, 1995; Shannon & Weaver, 1949). I briefly touch upon different emotion theories to position my study in the field of emotion

research, and I turn to perceptual practical theory to explain anatomical and physiological phenomena related to acoustic emotion expression. Finally, I use theories of learning to explain how I would want my parameter modulation technique to be used in a voice studio. The philosophical position dictated by my research methodology is a realist one. Epistemological realism states that there is a world out there (outside our minds) and ontological realism says that we can get information about it (Niiniluoto, 2017, 2018). This is precisely what I am trying to do in my investigation. I do not relate my results to any social concept or construction *per se*, but merely state that this is what my measurements indicate in this particular setting.

This thesis does not build theory; it stops at the modeling level of knowledge construction. In my research area, the model is understood as a vehicle that links practice to theory. The parameter modulation technique tries to give guidelines towards establishing a common vocabulary to acoustic and physiological phenomena in the singing voice through a taxonomy of research-based findings. It aims to facilitate easy conceptualisation of voice qualities which would, in turn, yield readily to simple exercises that "anyone" can pick up on. It links different theories to a practice-based model that allows the exploratory modulation (or change) of voice qualities on the grounds of what is already known about the emotional voice.

The logic of the three articles comprising the empirical part of this study was to first establish whether it was possible to perceive emotion from the singing voice (Study I), then to take a deeper look into the acoustic compilation of the voices to find out what kinds of elements might account for the recognition (Study II), and finally to come up with a training pattern that would drill the acoustic elements found typical for expressing the aforementioned emotions to see if training in this way would help to make expressing these emotions easier (Study III). The general result of this study was that the communication of emotion (using the singing voice) became easier after incorporating vocological information into the regular singing training.

2 THEORETICAL BACKGROUND

Vocology is the science and practice of voice habilitation, which includes evaluation, analysis, and intervention (Titze & Verdolini Abbot, 2012). This study focuses on the artistic voice and teaching expressivity from a vocological standpoint. The singing voice is the focus.

2.1 Emotions and emotional expression

Emotions are an interesting subject of study, as they mirror our lives so pervasively. Emotions arise within an organism in response to an external or internal event (Bericat, 2016). They operate on the borderline of autonomous homeostatic systems (which control life processes) and rational thinking. The function of emotions is to get us to direct our focus to the present as opposed to purely rational thinking, whose aim is to concoct genius plans for our future well-being. For the most part, emotions operate on a subconscious level, which means that their command center is located in a different part of the brain from the conscious language and symbol-based logical thinking that we use in our day-to-day operations and actions. There is, however, some connectivity between reason and emotion: one can, for example, fairly often name one's emotions and one can control them at least to some extent. It is also possible to manifest an emotion just by thinking it (e.g., think about jealousy) (Nummenmaa, 2019). We can see and hear the biological effects of emotions as changes in appearance, sound, and behavior in our very daily lives, and we are, in fact, quite masterful at interpreting these changes to secure our own well-being (Bericat, 2016; Darwin, 1873; Ekman, 1992; Scherer, 2001). We are even able to scan brain activity and identify emotionally induced changes in energy flows in the brain and show the connections of these energy flows using realistic modeling, but we are still unable to unambiguously tell what an emotion is ultimately (D'Angelo et al., 2013; Purves, Cabeza, Huettel, LaBar, Platt, & Woldorff, 2013).

Investigations into the biology of emotions have revealed different brain functions behind the emotions. There are four different main functions that control or elicit emotions: 1) mechanisms that recognize emotion, 2) mechanisms that

regulate motivation, 3) mechanisms that produce and monitor body-functions, and 4) mechanisms that control and regulate emotional episodes. All of these systems are functioning when we experience emotions, but the amount of their engagement in the process varies depending on the situation. The deep parts of the brain are in charge of rigid automation, such as birthing emotions and maintaining the homeostasis of the body, while the outer part of the brain functions as an operator for perception, thinking, and memories. It is in the frontal lobe(s) that we become aware of our emotions. Even though emotions run mostly on autopilot, it behooves us to be cognizant of their existence. When they become available for conscious processing, we can start to work with them. We can express our emotions to others in a constructive manner to facilitate change in our interactions, talk about them, show them, or hide them, and possibly reduce their impact on our lives if they are not serving their purpose (e.g., seeking therapeutic help in a situation where one has become insanely jealous). However, we cannot get rid of them completely and that is why it is important to learn how to live with them (Nummenmaa, 2019).

2.1.1 Emotion theories and emotion models

One way of approaching emotion research is to look at its ontological and epistemological approach, which is usually divided into theories based on evolution, cognitive evaluation, or social construction.

A) Evolutionary theoretical thinking in emotion research views emotions as genetically encoded programs that are activated in evolutionarily important situations. Once activated, emotions direct bodily functions – such as perceptions, energy levels, and the body's physical reactions – to solve the problem presented by the situation at hand (Darwin, 1873; Levenson, Ekman, Heider, & Friesen, 1992; Niedenthal, Krauth-Gruber, & Ric, 2006). The reasoning behind this line of thinking is that the most competitive brains would have developed affective heuristics (which are completely biological, largely neuronal, but with strong bodily and cultural connections) to facilitate rapid decision making for individual benefits, but also empathy and the sharing of survival-related resources with a group (Panksepp, 2008). One of the first theories in this tradition is the James-Lange theory, which postulates that anatomic arousal differentiates emotions (one does not cry because one feels sad, but one feels sad because one cries) (Purves et al., 2013). Another

is the Cannon-Bard diencephalic theory, which proposes that the diencephalon directs emotional stimuli simultaneously to the neocortex for the generation of emotional feelings and to the rest of the body for the expression of emotional reactions (Purves et al., 2013). The Cannon-Bard theory represents one of the first parallel-processing models of brain function, and as such it has contributed greatly to the appraisal theories of emotion that define emotions as processes rather than states (Levenson, Ekman, Heider, & Friesen, 1992; Moors, Ellsworth, Scherer, & Frijda, 2013; Niedenthal, Krauth-Gruber, & Ric, 2006; Nolen-Hoeksema et al., 2009; Purves, Cabeza, Huettel, LaBar, Platt, & Woldorff, 2013).

B) Emotion theories based on cognitive evaluation rely on personality psychology, basing their central idea on the observation that individuals may experience completely opposite feelings, even if the situation in which the emotion is experienced remains the same. This type of thinking has led researchers to consider emotion theory based on biological stimulation alone to be inadequate. According to cognitive emotion research, emotions arise from the assessment of situations relevant to the individual and from the cause-and-effect relationships that have led to that situation (Niedenthal, 2010; Nolen-Hoeksema et al., 2009; Purves, Cabeza, Huettel, LaBar, Platt, & Woldorff, 2013; Scherer, 1986). One of the first theories in this line of thought is the Schachter-Singer theory of emotion (Schachter & Singer, 1962), which suggests that the physiological arousal occurs first, but to experience and label it as an emotion an individual must first identify the reason for said arousal. The critical factor in this theory (as compared to the James-Lange and Cannon-Bard theories) is that the situation and the cognitive interpretation both have an effect on what we feel. Another emotion theory in the cognitive evaluation category is the Cognitive Appraisal theory (usually credited to Arnold 1960 and/or Lazarus 1970), which states that the appraisal must occur first before experiencing emotion (Arnold, 1960a, 1960b; Lazarus, Averill, & Opton, 1970). The Component Process model of emotion (Scherer, 1984), which we follow in this study, is an offshoot of the Appraisal theories.

C) From the sociological point of view, emotions are seen as products of culture that are shaped by the influence of culture to fit the culture (Bericat, 2016). For this reason, it is considered that human emotions are social constructs that serve the general goals of society. Based on social construction theory, social scholars think that the expression of emotions is regulated by pre-defined roles in society and the status of the individual in the community in which the emotion occurs (Bericat, 2016). Emotions are thus learned: they are based on the attitudes reflected by the emotional growth environment (its norms, practices, and values). They should never be thought off as simple physiological responses to the situation at hand; rather, the complexity of the subject in the environment should be examined. The way the subject evaluates the situation (consciously and/or unconsciously), to whom or what the subject attributes the cause/responsibility of the situation, the subject's expectations and active social identity at each given moment, and the subject's identification with other persons or groups all have an effect on how emotion arises (Bericat, 2016).

The onto-epistemological premise draws up guidelines for how emotions might appear in everyday life and in this way defines or suggests what kind of research on emotions should be done. The theoretical starting point that we choose for our research dictates the methods of research and limits to some extent the way in which emotion can be defined.

Another fairly common classification for emotions in the field of vocological research is to divide emotions into categorical, dimensional, and component process models.

A) Categorical emotion models state that there are several distinct emotional systems in humans, each with its own adaptive behavioral function. This model suggests that humans have evolved a limited number of basic emotions that are always activated in certain types of situations and have certain characteristic manifestations that can be perceived as physiological changes and changes in expression and behavior. For example, facial expression research has provided support for viewing emotions as separate categories. It has been shown that certain emotions (such as joy, disgust, fear, sadness, and anger) are universal, and the facial expressions associated with these emotions are not only produced but also recognized in the same

way around the world and in different cultures (Darwin, 1873; Ekman, 1992, 1993; Izard, 1992, 2007; Levenson et al., 1992; Nolen-Hoeksema et al., 2009; Purves et al., 2013).

B) According to the dimensional model, emotions are seen as a point within a complex space that includes two or more continuous dimensions. This model is originally derived from Wundt's three-axis model, according to which emotions can be distinguished in a three-dimensional state on the axes of comfort – discomfort/rest – activity/relaxation – attentiveness (Laukka, Juslin, & Bresin, 2005; Scherer, Dan, & Flykt, 2006; Smith-Lovin, Lewis, & Haviland, 1995). Nowadays, valence (the affective quality referring to the intrinsic attractiveness or averseness of an event, object, or situation) and arousal (the physiological and/or subjective intensity of emotion) are considered to be critical dimensions by which emotions can be differentiated from each other (Juslin & Sloboda, 2010; Lewis, Haviland-Jones, & Feldman Barrett, 2010). Many studies (including this one) that ask participants to rate the emotional properties of music and sound use the circumplex model to differentiate emotions. The circumplex model organizes emotions around the circumference of a circle positioned at the intersection of the two orthogonal axes of arousal and valence (Purves et al., 2013; Russell, 1980).

C) The component process model sees emotion as interrelated changes in many of the organism's psychobiological functions. It considers the assessment of things and objects or people in relation to an individual's goals or needs and takes into account the changes in autonomous and physiological behavioral preparation and readiness, motor expression, and subjective feeling that result from this assessment process. The component model seeks to explain all emotion and emotional complexity in a single model and seeks to avoid truncating emotion to a few basic emotions or dimensions. According to the component model, motor expressiveness is a direct result of the evaluation process, which in turn is guided by, for example, the novelty value, comfort, goal control, survival potential, and normative significance of the thing to be evaluated. The evaluation process (or appraisal) is thus seen to play a key role in awakening and differentiating emotions. Assessment is seen as distinguishing an emotional stimulus from a reflex response and allowing for

flexible and adaptive interaction with the environment (Dael, Mortillaro, & Scherer, 2012; Nolen-Hoeksema et al., 2009; Purves et al., 2013; Scherer, 1984; Scherer, 2009; Scherer & Moors, 2019; Scherer, 2001; Scherer, 1984).

How we decide to define emotion determines how we can explore it. Basic emotions are studied from an evolutionary perspective; as they are seen as biological responses, they are often also studied by measuring biological responses. The dimensional model, on the other hand, easily bends to various surveys, for example. The component process model (CPM) is probably the most comprehensive of the models for defining emotion (although there are many more than these three models available to choose from). The problem of the CPM, however, is that the operationalization of this definition becomes quite challenging. It is very difficult to construct a research design in which all of the above-mentioned aspects of emotion are taken into account. Therefore, experimental setups/research questions are often limited to just some small slice of the broad spectrum of emotion manifestation as depicted by the CPM.

2.1.1.1 Appraisal theories of emotion

The way this study is modelled after Scherer's work on vocal emotion expression places it onto-epistemologically under the appraisal theories of emotion. The basic premise of appraisal theories is that "emotions are adaptive responses which reflect appraisals of features of the environment that are significant for the organism's well-being" (Moors et al., 2013, p. 119). They are componential theories because they view emotion as comprising changes in numerous organismic subsystems. These subsystems or components include an appraisal constituent that oversees the evaluations of the environment and person-environment interaction. The other subsystems are the motivational component responsible for different forms of action readiness, a somatic component that accounts for peripheral physiological responses, a motor component with expressive and instrumental behavior, and a feeling component that accounts for subjective experiences or feelings (Moors et al., 2013). Appraisal is a process that scans and evaluates the significance of the environment for well-being. It is an inherently transactional concept as it involves an interaction between the event and the appraiser (Lazarus, 1991).

Most appraisal theorists adhere to a dual- or triple-mode view of appraisal. The dual-mode view separates a *rule-based mechanism*, which consists of the on-line

computation of one or more appraisal value(s), from an *associative mechanism*, which consists of learned associations between representations of stimuli and appraisal outputs stored in the memory. The triple-mode view adds a *sensory-motor mechanism*, which consists of hedonic feelings, motor responses, and the activation of unlearned associations between sensory features (Moors et al., 2013). There are some questions about the relations between these mechanisms, automaticity, and formats of representation (e.g., verbal code), but appraisal theorists generally agree that various mechanisms can underlie appraisal and that they can operate on a wide range of representations. They believe that appraisal often proceeds automatically (i.e., uncontrolled in the promoting or counteracting sense, unconscious, efficient, and/or fast), but can also sometimes proceed nonautomatically (Moors et al., 2013). The representation of appraisal value(s) is unconscious by default, but part of it can become conscious and then that part becomes a part of the content of feelings (or subjective experience of emotions) (Scherer, 2009). If we say that the appraisal component of emotion shapes the motivational, somatic, and motor components, appraisal is then viewed as the core determinant of feelings. Changes in appraisal may lead to changes in physiological and behavioral responses. These changes may in turn affect the appraisal, for example, via a change in the stimulus situation. As a consequence of this iterative process, several emotional episodes may run in parallel (Moors et al., 2013).

Appraisal theorists allow variation in appraisal variables that are being processed simultaneously. If only a couple of appraisals are enough to bring forth an emotional episode, the emotional experience is seen as relatively undifferentiated and global, but if many appraisals are made, the emotional experience is highly differentiated and specific (Moors et al., 2013). This corresponds with the idea of primary and secondary emotions: primary emotions are seen as universal, physiological, of evolutionary relevance, and biologically and neurobiologically innate, whereas secondary emotions are considered to be socially and culturally conditioned (Bericat, 2016).

2.1.2 Locating vocal musical expression in emotion theories

We can intuit that music is the language of emotion, the singing voice is evolutionarily in a prime position to carry emotional messages in musical expression, and music conveying strong emotions is often regarded as somehow better than

plainer music. But what exactly is emotion expression in the singing voice, and how can we investigate it?

The first point of demarcation is to make a distinction between studies concerning emotions in music, emotions in singing/speaking, and emotions *per se*. This study investigates emotion expression in the singing voice, so it has an inherent musical component to it, but it is not a study of music. The study is close to the study of emotion expression in the speaking voice, but it focuses specifically on the singing voice, which has some implications for how the emotions can be expressed. Finally, the study is about emotion expression and emotion perception, which is not the same thing as emotion, but it is close to it. One can both express and perceive emotion without actually having an emotion (Juslin, 2013; Juslin & Laukka, 2004; Scherer, 2003, 2004, 2005).

A further task is to arrive at a working definition of emotion as it is experienced in music. Most emotion researchers agree that emotions can be seen as relatively brief and intense reactions to goal-relevant changes in the environment that consist of several components (Juslin & Laukka, 2004). But how do emotions manifest themselves in music? Juslin (2013) has identified eight mechanisms that might account for musical emotions. The key to his reasoning lies in tapping into the processes through which sounds are imbued with meaning: 1) The brainstem reflex may account for increased arousal and evoke feelings of surprise in the listener. 2) Rhythmic entrainment refers to a process where a strong rhythm in the music influences some internal bodily rhythm of the listener and can evoke feelings of communion and emotional bonding. 3) Evaluative conditioning is a Pavlov's dog-type reaction to music where emotion arises simply because the piece of music has been paired with other negative or positive stimuli in the past. 4) Emotional contagion refers to perceiving the emotional expression and then mimicking this impression internally. The contagion is especially potent with sung music, as the brain areas responsible for the contagion are localized in the pre-motor cortex through some kind of mirror-neuron system. 5) Visual imagery refers to coming up with vivid mind imagery while listening to music. This kind of process usually evokes positive emotions such as pleasure and deep relaxation. 6) Episodic memory works in bringing up memories through listening to music. The emotions evoked by remembering past events are typically nostalgia, melancholia, etc. 7) A certain type of emotion can be evoked by musical expectancy. These expectations are based on the listener's previous experiences of the same musical style and are therefore culturally tinged. The emotions awoken by musical expectancy can include emotions like surprise, thrills, but also irritation. 8) Finally, aesthetic judgment could evoke the

emotions if a piece of music is assessed as extraordinarily beautiful. According to Juslin, the emotion in question would then be some sort of awe or spiritual emotion (Juslin, 2013).

The idea behind this theoretical postulation is to fit the "real" emotions and aesthetic emotions under the same theoretical framework and underline the fact that what counts for one particular source of emotion in music and singing may not count for another source. Different theories may be required for explaining different sources. Most importantly, a failure to specify which source(s) of emotion one is studying could lead to unnecessary quarrels with people studying other sources and prevent the cumulativeness of research efforts (Juslin, 2013). As it is relatively impossible to come up with a comprehensive research design that would consider all the possible emotion components (see Table 1) and processes and to correlate them with sources of emotion in music, the next best thing is to clearly state the components, processes, and sources upon which one is focusing.

Besides situating itself in the realm of appraisal theories and component process models of emotion, this study focuses on emotion expression and emotion perception. The study design utilizes the idea of categorical emotions as well as dimensional ones, but it moves towards the component process model in going into great detail in investigating the expression component of emotion (see Table 1) while leaving other aspects of emotion relatively untouched. The voice samples used in this study have been most likely sung making a pre-set strategy of expression using episodic memory rather than actually feeling the target emotions. The voice samples have most likely been perceived using cognitive evaluation, but some sort of emotional contagion is also possible.

Table 1. Components of emotions (Juslin & Laukka, 2004) as seen from the viewpoint of the performer and listener. An exemplary explanation adapted from the original.

Components of emotion	From the viewpoint of the performer	From the viewpoint of the listener
Cognitive appraisal	You decide your approach to the expression of joy	You assess that the singing voice sounds joyous
Subjective feeling	You identify with the song; it reminds you of a happy event in your life	You feel joy listening to music
Physiological arousal	Your breathing intensifies when you interpret a fast and loud song	You get goose bumps from the singers' interpretation of a song
Emotional expression	You sing your heart out with a smiling face	You start to smile when listening to a happy song
Action tendency	You increase medial compression in your larynx when expressing joy	You want to start dancing to the music
Emotion regulation	You hold back tears of happiness so that they do not affect your vocal technique	You hold back tears so that you do not embarrass yourself in front of other listeners

2.1.3 The Shannon–Weaver model of communication

The way I view aesthetic emotional expression through the singing voice in this study is that it is a kind of strategy to get a message through to the listener. The Shannon–Weaver theory of communication (Shannon & Weaver, 1949) offers a great gateway

theory to the concepts behind vocological methodology and the way this study is constructed. This theory argues that communication can be broken down into six key concepts: sender, encoder, channel, noise, decoder, and receiver. The model's primary value is in explaining how messages are lost and distorted in the process of communication. The Shannon–Weaver model provides a basic framework for analyzing how auditory information is transmitted and thus it can help in organizing the basic layout of vocological study designs.

According to the model, the sender (information source) starts the process by choosing what kind of message to be sent, through what kind of channel, and to whom it will be sent. The message can be of any kind, but in this study, the focus is on vocal messages. The next step in the model is the encoder, which converts the idea into a signal that can be sent from the sender to the receiver. (The original model was intended to explain communication through machines, but the encoder can just as well be a person who converts an idea into a sound with the purpose of communication.) After the sender has encoded the message, they send it to the channel, which is the infrastructure that gets the information through to the decoder and the receiver. In the present study, this medium is essentially air because the human voice travels as changes in atmospheric air pressure, but there is the additional medium of audio to digital and back to audio conversion, which adds an additional layer of "noise" to the signal. Noise is defined by Shannon and Weaver as something that interrupts a message while it is on the way from the sender to the receiver. There are two types of noise according to the model: internal (which happens when a sender makes a mistake encoding a message or a receiver makes a mistake decoding the message) and external (when something external impedes the message, such as an additional competing sound source). The next step in the model is the decoder, which is the exact opposite of the encoder; it must interpret the meaning behind the voiced sound. The receiver is the destination of the message. The last step in the model is the feedback loop, which makes the model circular (introduced to the model by Weiner), but in this investigation, only the original linear construction of the model is used (Chandler & Munday, 2011; Shannon & Weaver, 1949; Wiener, 1971).

2.1.3.1 The Shannon-Weaver model in musical expression

Singing is a mixture of nonverbal and verbal communication, and it utilizes cues from both musical sounds and everyday speaking sounds. Music, on the other hand, is a form of nonverbal communication, which is abstract in its nature. Musical

communication refers firstly to a social signaling system (a code), secondly to an encoder who tries to pass on information using the code, and thirdly to a decoder who tries to interpret this code. The success of information transmission depends on the capacity of the channel used (in this case the effectiveness of the emotional expression of the singing voice) and the amount of noise infiltrating the channel. The capacity of the channel is equivalent to the amount of information this channel can transmit in a certain timeframe. If the capacity of the channel is small in relation to how complex the message it is supposed to transmit, the channel overloads. This means that the channel gets noisy and the message harder to interpret. One can try to reduce noise in the signal by using redundance (Juslin, 1997; Juslin & Sloboda, 2011). In the context of singing, this usually means multimodal emotion expression, using facial expression and body posture in conjunction with the voice when expressing emotions (Hawk, Fischer, & Van Kleef, 2012). However, there are situations when one has to impress using only one's acoustic signal (e.g., on the radio). In these cases, it is best to load the acoustic signal with overlapping emotion cues. The singing voice has a particularly good capacity for carrying an emotional signal because of its evolutionary background (Parada-Cabaleiro et al., 2017). Neuropsychological research has shown that there are certain musical variables, such as the timbre (or the quality) of sound, which are processed on the same neural pathways as speech. Musicians use these channels to express emotion via emotion-specific patterns of acoustic cues (Juslin & Laukka, 2003). The aim of this study is to investigate the conveyance of the emotional message using only the voice quality as the carrier.

2.1.4 Operationalizing emotion: The Brunswikian lens

Research on emotion expression in the human speaking and singing voice can be operationalized into various measurable quantities: sound parameters, pitch contour, amplitude envelope, breath patterns, body postures, emotional words, emotional perceptions, images of the phonatory system, and so on. In this study, we operationalize the emotional singing voice into sound parameter values and listening test answers. The theoretical background behind this operationalization is Brunswik's lens model (Brunswik, 1956). The model postulates that in order to adapt to the constantly changing environment, one must use a deduction method based on probability, utilizing small and uncertain bits of knowledge (proximal cues) to form a view of the world (the distal object) (Table 2). The lens model says that the

fragmented sensory information that we get from the outside world will condense, fanlike, into the lens and form one conclusion regarding that information. The way in which the conclusion corresponds to reality is called ecological validity (Brunswik, 1956). When applied to emotion recognition from the singing voice, the lens model would be as follows (Table 2):

Table 2. Brunswik's lens applied to emotion expression in singing.

A	B	C
Distal object	Observable and measurable cues	Perception
The emotion expressed by the singer	The vocal features affected by the emotion and used by the listener to infer emotion	A perception or perceptual judgment from a human observer

K.R. Scherer has elaborated on part B of the lens model and divided it into 1) distal objectively measurable cues and 2) subjective, proximal percepts of these cues. In this study, they are 1) acoustic parameter values and 2) voice quality impressions formed by the listener. Scherer argues that the signal gets easily distorted on its way from the sender to the receiver because of different noise factors, and therefore it is of pivotal importance to account for the distortion in the model by using both acoustic analyses and listening tests when investigating the transmission of emotional voice signals (Scherer, 2003, 2013). Later he further developed this model into a "tripartite emotion expression and perception (TEEP) model" (Table 3), in which he describes the communication process through four elements and three phases (Bänziger et al., 2015; Scherer, 2013). The major justification for the TEEP extension is that it provides a way to deal with the noise component that can occur in the transmission process. In order for his TEEP model to work flawlessly, Scherer suggests adding a group of expert judges to the receiver end before the actual receiver to verify the quality of the signal. When all of these elements are in place, one can perform a lens model equation, which is based on two regression equations and two correlations and gives descriptive information on effect sizes or the proportion of variance shared/explained between the respective variables (Bänziger et al., 2015).

Table 3. The tripartite emotion expression and perception model by K. Scherer (Bänziger et al., 2015; Scherer, 2003, 2013).

A	B		C
SENDER	DISTAL CUES	PROXIMAL PERCEPTS	OBSERVER
Externalization driven by external models and internal changes	Transmission (perceptual representation)		Cue utilization driven by inference rules and schematic recognition (inferential utilization)

2.2 Voice quality

Timbre is the perceptual quality in sound that allows listeners to detect differences between different voices or variation in the same voice when the basic loudness and pitch parameters are identical. Timbre/sound quality arises from sound characteristics that do not directly fall under the categories of frequency or intensity. These other elements include the way that energy is distributed across the sound spectrum, the amount of noise in the sound, the type of noise in the sound, and the rate of attack and decay in the stimulus (Purves et al., 2013).

The ANSI standard defines voice quality as a psychoacoustic attribute, a perceptual response to a voice signal and all its acoustic attributes (ANSI, 2020). The quality of how the voice sounds is the result of the combined output of the relative strength of the sound's different subcomponents, which can be obtained by spectral analysis. However, trying to match sound spectrum to voice quality is not just a question of amplitude mapping the harmonics. It is also defined by the fact that a human voice is never perfectly periodic. The voice contains small short-term irregularities in phonation in the period length and amplitude that occurs from cycle to cycle. Voice quality is always determined according to the norm that we call neutral. Individual neutral is determined by individual anatomical and physiological features (Laver, 1980). Habitual voice quality refers to individuals' characteristic way of using their own voice. Air pressure, vocal fold adduction, the symmetry of vocal

folds, muscle tension in the larynx, the shape and size of resonant cavities and articulators, and the amount of air escaping to the nasal cavities all affect the voice quality. By changing the habitual neutral settings of these factors, it is possible to change the habitual voice quality. Voice quality also changes during the psychophysical changes that occur during an emotional episode (Laver, 1980; Scherer, 1986; Sundberg, 1987).

Voice quality is based on both the laryngeal function of the vocal folds and the filter action of the vocal tract, but it is also dependent on the air pressing out from the lungs (Gobl & Ní Chasaide, 2003). This means that there are three separate mechanisms whose actions are very tricky to separate; they are designed to function simultaneously together when singing. There are different measurement techniques in use for gaining information about each of these functionalities separately, but this study uses acoustic analysis, as it is the measuring technique most analogous to the work of voice teachers that mostly happens by decoding information from the acoustic signal – analyzing voice quality.

2.2.1 Listening to the voice

In the context of voice quality, cognitive factors cannot be disregarded, since voice quality relates to the suitability of its function in a specific situation, and such suitability cannot be assessed without cognitive functions. Perceived changes in the face, voice, and body posture of the interlocutor can affect the emotional state of the perceiver, but there are many possible mechanisms that can make this happen, including the perceiver's sense of self, different physiological processes, and learning processes. Even if the listener cannot consciously recognize the emotion expression, the motor activity affects the perceiver. Listening to the human voice activates action codes in the listener. This means that the listener models the singing actions in their own vocal organs. Motor mimicry is thought to be an integral part of mood contagion, which happens when the person unconsciously imitates the emotional state of the interlocutor (Neumann & Strack, 2000). The individual differences in how we perceive music and especially vocal sounds in music (our educational and cultural background, our personal tastes) make us exhibit different attitudes even to the same piece of vocal music. Music psychologists have proposed a distinction between perceived expressions of emotion and felt expressions of emotion, the difference being that perceived emotions can be recognized from music without necessarily feeling them, while felt emotions are actual emotions being induced in

the listener by the music (Gabrielsson, 2001; Hodges, 2010). In recognition studies where emotion is appraised from short samples (such as in this present study), the emotions are most likely perceived, as the induction of emotion generally takes a little more time (Parada-Cabaleiro, Costantini, Batliner, Schmitt, & Schuller, 2019).

The singing voice is probably the oldest sound associated with quality ratings, but finding a simple and general model of voice quality that could be generalized to all emotional singing is very hard, if not impossible (Cunningham, Weinel, & Picking, 2018; Eyben, Salomão, Sundberg, Scherer, & Schuller, 2015; Jansens, Bloothooft, & de Krom, 1997; Livingstone, Choi, & Russo, 2014; Scherer, Sundberg, et al., 2017; Sundberg, Salomão, & Scherer, 2021). Perhaps the only common factor is the listener. The most reliable way to study the aesthetic value or informative feature of the sound is to conduct psychophysical experiments with a group of human subjects or assessors. Listeners are used to interpreting many things from the voice, such as the speaker's age, gender, social class, and emotional state (Juslin & Laukka, 2001; Kempton, Scherer, & Giles, 1981). The development of sound-quality models and theories involve subjective listening experiments from which data are collected and analyzed statistically. Then, based on this information, different techniques that allow for detailed investigation of the factors and features of sound are used to explain the pattern of assessment (Pulkki & Karjalainen, 2015).

Decoding emotional information from the voice seems to function primarily via the subcortical route: even small babies can recognize the emotional content of speech before they can understand the semantics (Murray & Arnott, 1993). As we get older, the interpretation is broadened by our cultural understanding and it becomes a blend of preconscious and conscious appraisal (Davis et al., 1996; Purves et al., 2013). Recognizing the emotional state of our interlocutors helps us to make fast interpretations about their intentions and behavior. In music, we make similar fast categorizations about the function of music – whether it is for a party, grieving, battle, or a lullaby for sleeping. In party songs, the tempo is fast, the volume is up, and the sound color bright, whereas in songs for grieving, the tempo is slow, the volume is quiet, and the timbre is soft (Juslin & Laukka, 2003). The acoustic parameters inside these song categories are similar to the ones we use in vocal communication to convey emotion. There has been reported evidence of music and speech using the same neural pathways in processing emotional information, but the extent of overlap in these systems is still debated (Nair & Large, 2002; Simpson, Oliver, & Fragaszy, 2008).

Whereas the ability to code emotional information to the voice varies greatly from individual to individual, the ability to decode emotional information from the voice

seems to be more stable among humans, and the decoding errors seem to be more systematic (Juslin & Laukka, 2003).

2.2.2 Vocal emotional expression in singing

(Pre-composed) Music itself exerts a lot of restrictions on the way singers may express emotions vocally. The pitch is predetermined by the composer, the harmonic content suggests consonance or dissonance, and the rhythms are fixed. The tempo is often static and predetermined, and dynamic contours might also be written into the music. Therefore, the singers' expression is almost always some sort of a compromise between what they think the song is about and what the composer had in mind.

Because emotions bring on physiological changes in the respiratory , laryngeal, and vocal tract musculature, they can be heard in the singing voice as changes in the quality of the voice (Eyben et al., 2015; Hakanpää, Waaramaa, & Laukkanen, 2019; Jansens, Bloothooft, & de Krom, 1997; Kotlyard & Morozov, 1976; Livingstone, Choi, & Russo, 2014; Scherer, Sundberg, Fantini, Trznadel, & Eyben, 2017; Scherer, Trznadel, et al., 2017; Siegwart & Scherer, 1995; Sundberg, 2000; Sundberg, Iwarsson, & Hagegård, 1995; Sundberg, Salomão, & Scherer, 2021). The singers for their part can encode many things that suggest emotional information to the singing voice by making slight pitch shifts, temporal shifts, and volume and timbral changes (Juslin & Laukka, 2003, 2004; Sundberg, Lã, & Himonides, 2013). These encodings can be very strategic and placed in the acoustic signal with great skill and mastery of the instrument, but they can also happen automatically. This difference in encoding styles can be explained by the dual process models, which distinguish between the associative and reflective information processing systems. Implicit expressions of emotions can therefore be tracked to the associative system, while explicit expressions of emotions can be tracked to the reflective system. Depending on the individual's level of introspection, the reflective system can access the information being processed in the associative system. On the other hand, affective processing can escape our conscious focus, but it can still affect our appraisals, behavior, and physiology (Quirin, Kazén, & Kuhl, 2009; Smith & DeCoster, 2000). We can hear traces of affective acoustic communication in everyday situations where there is an immediate vocalized reaction to an event, such as "Oooooouch!" when you miss a nail and hit your finger with a hammer, or "Aaaaawwww!!!" when you love dogs and see a cute puppy (Scherer, 1995). These are short, fast, and automatic exclamations

of meaning encoded to the voice mainly by changes in the voice quality. When we want to exploit this system of voicing and make artistic interpretations with it, we need to first become aware of its function and then determine how to double process it for the use of the reflective system and the practice of singing.

One quite simple way of organizing vocal emotion expression for singing is through the activity level of emotions. Compared to the expression of low activity emotions, vocal expressions of high activity emotions are typically characterized by a faster tempo, greater loudness, quick changes in loudness, a lower level difference between the lowest partials H1 and H2, a flatter spectrum, large vibrato extent, local departures from pitch contour at tone onsets, and a higher degree of perturbation and noise (Eyben et al., 2015; Livingstone et al., 2014; Scherer et al., 2015; Sundberg, 1998, 2000). The extent to which one can use these quality differences in singing depends on the aesthetic demands of the music genre and the kind of voice use which is considered appropriate within that genre. Previous research has found that singers use different singing techniques for different styles of singing, and that they vary their techniques even within a musical style (Bourne & Garnier, 2012; Bourne, Garnier, & Samson, 2016; Yanagisawa, Estill, Kmucha, & Leder, 1989). This lends support to the idea that some technical variation is tolerated in most singing styles, and thus expression via voice parameter modulation, or the deliberate changing of one's own voice quality, could be considered as an acceptable technique for training emotion expression.

2.3 Acoustics

In order to fully grasp the meaning of voice quality in the singing voice, I find it helpful to start piecing the concept together from the smallest common denominator – the acoustic parameter. In this chapter I discuss the composition of voiced sound through the sound parameters measured in the empirical part of this study. I also give an overview of the underlining theories behind the reason for choosing to measure these parameters. This next chapter deals with the acoustics part of vocology.

2.3.1 The source-filter theory and nonlinear interaction

The source filter theory is credited to Gunnar Fant (1970), and it takes a step deeper into the core of vocological research. A source-filter decomposition of the voice

wave data implies a detailed mapping of the singer's vocal tract during the singing process so that its transmission properties can be specified. The lungs are seen as a pressure compressor that gives the driving force to the glottal oscillator (vocal folds) which creates the sound source). The filter system is comprised of the vocal and nasal tracts. Finally, the signal radiates from the lips or the nostrils to a listener (or a microphone) (Fant, 1970; Pulkki & Karjalainen, 2015; Sundberg, 1987). The subsystems are coupled in a series: the excitation (generation), a two-port transmission line, and a radiation impedance (acoustic radiation load). Functioning of the system is described by acoustic variables: sound pressure and/or volume velocity at each point of the system. If we consider only a single transfer function in the frequency domain as a Fourier transform formulation, the pressure of the radiating waveform comprises of the glottal excitation (for example volume velocity), the corresponding transfer function of the vocal/nasal tract system and the transfer function of mouth/nose radiation. The input signal is volume velocity and the output signal is the radiated pressure signal (Pulkki & Karjalainen, 2015; Sundberg, 1987). The sound radiates from the mouth to the atmosphere resulting in the voice. Because of this radiation, the sound source is in proportion to the time derivative of the glottal flow. This is why in voice literature, it is the time derivative of the glottal flow and not the glottal flow, which is considered as the voice source (Zhang, 2016). Fant devised one of the first mathematical models with which the acoustic production of speech signals can be approximated. It is a simple linear function where source (S) and transfer function(T) are seen to create the final product of corresponding speech sound (P) (Fant, 1970).

$$P = S \cdot T$$

Different vocal tract filters have a different characteristic frequency response, which accounts for a large part of our individual voice quality. Perceptually, vocal tract resonances are responsible for distinguishing different vowel sounds and coloring the singing voice, giving it its special timbre(s). If the short-term amplitude spectrum is constant with time, then the timbre is also constant. The vocal tract length is the distance from the glottis to the lips, and the cross-sectional area (or the shape) of the tract varies with the distance to the glottis. By moving the articulators and changing the oral cavity shape, the singing voice timbre also changes (Pulkki & Karjalainen, 2015; Titze, 1994; Titze et al., 2015). The effect these configurations have on the spectral partials is called the area function. It determines how the component frequencies of the voice source spectra are modified (Sundberg, 1987;

Welch et al., 2000). Transfer function refers to the frequencies that are transferred most efficiently by the vocal tract (aka. formants). The partials in the spectrum of the voice source, which are closest to a formant frequency, are radiated from the lip opening with a greater amplitude then other partials (Sundberg, 1987). The transfer function changes every time the shape of the cavity changes, and quite small movements of the articulators can cause formants to change in frequency, increase in amplitude or to attenuate.

Strictly speaking the linear source-filter function is a simplistic approximation of what really happens in the transmission process. Nonlinear source-filter coupling is a more accurate conceptualization of what happens in voicing. This theory assumes that the acoustic airway pressures contribute to the production of frequencies at the source. The diameter of the epilaryngeal tube plays a leading role in the nonlinear coupling either matching or mismatching the output impedance of the glottis to the input impedance of the vocal tract. When the glottal impedance is high and the epilarynx tube impedance is low the coupling would be weak and when the impedances are comparable a strong coupling (or nonlinear interaction) is obtained. This theory assumes that the transglottal pressure includes a strong acoustic component (Rothenberg, 1981; Titze, 2008a; Titze, Riede, & Popolo, 2008). When the fundamental source frequency lies well below the formant frequencies of the vocal tract, the source is influenced mainly in terms of glottal flow pulse skewing and pulse ripple. The skewing of the flow pulse produces new harmonic frequencies in the glottal airflow that are not part of the glottal area waveform. Titze (2008) conceptualized this as Level 1 interaction:

> "Level 1 interaction contributes to the spectral slope and the spectral ripple in the sound source, even when the spectrum is purely harmonic and no bifurcations in vocal fold vibration occur. The supraglottal and subglottal impedances are additive for this interaction. If both impedances are inertive (positive), a maximum skewing of the flow pulse is achieved, which increases the maximum flow declination rate and thereby vocal intensity. Individual harmonics can be enhanced or suppressed by frequency-dependent reactances that change from positive to negative" (Titze, 2008b p.2747).

When the fundamental frequency crosses the formants bifurcations in the dynamics of vocal fold vibration can occur that may create a sudden jump of fundamental frequency, or changes at the overall energy level at the source. Titze (2008) conceptualizes this phenomenon as Level 2 interaction:

> "Level 2 interaction is realized more in high *fo* productions for which the dominant harmonics are near the formants. Frequency jumps and a variety of new source frequencies or instabilities can be produced, including subharmonics and non-random noise. The instabilities occur mostly when one of the dominant harmonics encounters

sudden changes in reactance, destabilizing the modes of vibration of the tissue that are affected differently by reactance" (Titze, 2008b p. 2748).

There are plenty of variation in nonlinear interaction between singers and the occurrences of nonlinear coupling may vary within subjects for repeated vocalizations (Titze, 2008b; Titze et al., 2008).

To sum up: The linear source-filter theory is a simplified model which assumes that the source frequencies are produced independently of the acoustic pressures in the airways. The source-filter interaction is a nonlinear phenomenon because new harmonic distortion frequencies are created by the vocal tract. For the purposes of using the parameter modulation technique in teaching of singing, comprehending the linear source-filter theory will for the most part suffice. In this study we use only the radiated spectrum to measure the acoustic parameters from the singing voice and thus have no way of knowing what kind of nonlinear phenomena has occurred during voicing. The source-filter system can be viewed as an *encoder* of singing voice sounds which attaches it nicely to the Shannon-Weaver theory of communication (see chapter 2.1.3.).

2.3.2 Acoustic parameters and their perceptual correlates

Different measuring techniques have been established to enable the dissemination of different aspects of the human voice. These aspects are called sound parameters. There are two ways of looking at these parameters: a) the physical measurement of mechanical vibrations and b) the hearing-related concepts that correlate (but are not necessarily matched one to one) with these physical phenomena. Below is a list of the parameters investigated in this research.

2.3.2.1 Fundamental frequency – pitch

The voice from a physical point of view is a pressure wave propagating in a medium. It consists of many sinusoidal pure tones, called partials, each having their own frequency, amplitude, and phase. This complex tone is periodic in nature, which means that it repeats itself and the partials have harmonic relationships to each other. The lowest frequency and the longest wave of the complex waveform is called the fundamental frequency, or *fo* for short (Pulkki & Karjalainen, 2015). Measuring the *fo* from a singing voice is always an approximation, because the singing voice rarely stays stable for more than a few moments, even on the same note. For example,

vibrato, which is considered an integral part of a healthy singing voice, is the rapid fluctuation around the target pitch, and it can vary even by a few semitones without sounding too out of tune. The unit of measurement for *fo* is Hertz (Hz) (Laukkanen & Leino, 2001; Plack, 2013; Pulkki & Karjalainen, 2015; J Sundberg, 1987; Suomi, 1990; Vurma & Ross, 2006).

Pitch is the perceptual correlate of the periodicity of an acoustic waveform. A harmonic complex tone is the most commonly considered form of pitch-evoking sound. The harmonic complex waveform repeats at a rate that corresponds to the fundamental frequency (*fo*) and can be taken apart to sinusoidal harmonics or overtones, which have frequencies at multiple integers of the *fo*. The relative amplitudes of the harmonics within a complex tone are of great importance in determining the sound quality, or timbre, of the singing voice (Oxenham, 2012).

Pitch cannot be measured or expressed in physical units because it is a subjective quality (though the Mel scale has been developed using psychoacoustic tests to quantify pitch (Pulkki & Karjalainen, 2015)). However, it is reflected in the repetition rate of a sound, and quite often it is used synonymously with fundamental frequency in colloquial language.

The clarity (or the strength) of pitch depends significantly on the nature of the sound. Periodic signals have a higher pitch strength than aperiodic ones, while sinusoids produce the clearest pitch perception (Pulkki & Karjalainen, 2015). Pitch detection is best at a frequency range between 30 Hz and 4 kHz, but also sound duration has an effect on pitch perception. At the lower frequencies of the entire human hearing capacity (400 Hz to 6 kHz) just a little over two period lengths or less than 20 ms are required for pitch recognition. The best accuracy in pitch perception is achieved after 100-200 ms of sound onset (Pulkki & Karjalainen, 2015). The mechanism leading to pitch perception has been explained by the place theory, which assumes that the frequency-place mapping happening in the cochlea accounts for pitch perception, and the timing theory, which states that a time-domain analysis of periodicity is also performed in pitch perception. On the (vocal)pitch production side of the theoretical discussion, there are also a few competing theories: 1) perception necessarily precedes vocal production, 2) vocal stimuli are first processed for motor-relevant features and only after that are sent forward into our conscious perception for symbolic representations, and 3) vocal stimuli are processed for motor-relevant features and conscious, symbolic representations along two different and independent pathways (Hutchins & Moreno, 2013; Pulkki & Karjalainen, 2015).

In this study, I am interested in pitch perception and fundamental frequency in the context of emotion recognition and expression. I ask whether the perception of

emotion changes at different pitches and whether the expression of emotion changes when expressed at different fundamental frequencies (Studies I & II).

2.3.2.2 Sound pressure level (SPL) – loudness

The most important physical measure in this study is sound pressure. Sound pressure is the deviation of pressure from the static pressure in a medium due to a soundwave at a specific point in space. In the context of this study, the medium in question is air. Sound pressure values are typically much smaller than the static pressure, but they can nevertheless be captured quite easily with a good condenser microphone, which converts them into an electrical signal (voltage) with high accuracy (Pulkki & Karjalainen, 2015). All the acoustic measurements presented in this study are based on pressure differences captured by a microphone.

The unit for pressure measurement is the Pascal (Pa), and it describes the force applied per unit area in a perpendicular direction to the surface of an object [Pa] = [N/m^2]. As the sound pressure varies over a large range in Pascal units, it is more customary to use decibels [dB] to describe the ratio between two amplitudes. A *level* expressed in terms of decibels describes a ratio relationship between two values and it is not an absolute measurement. If we say that Amplitude 1 is our reference, then the relative amplitude of another sound in decibels can be calculated as:

Level in decibels = $20\log_{10}$(Amplitude2/Amplitude1).

It is customary to use the concept of the decibel in a specific way in acoustics. In relation to the fixed role of Amplitude 1 as a denominator in the equation above, decibels are seen as *level* units. The reference sound pressure $p_0 = 20 \cdot 10^{-6}$ Pa is used so that the sound pressure level (SPL) Lp [dB] is:

$$L_p = 20 \log_{10}(p/p_0).$$

This value p_0 is selected so that it roughly corresponds to the threshold of hearing, the weakest sound that is just audible, at 1 kHz (Pulkki & Karjalainen, 2015; Svec, 2018). The quietest sound that the human ear senses is thus denoted by 0 dB SPL.

The perception of sound pressure level is referred to as loudness. Loudness can be described as "that attribute of auditory sensation in terms of which sounds can

be ordered on a scale extending from quiet to loud" (ANSI-S1.1[2], 2020). Much like in *fo* and pitch, the SPL does not estimate perceived loudness directly. The change in sound pressure relative to the sound pressure level works so that when the pressure is doubled or halved, the sound pressure level changes by ±6dB SPL. However, this is not directly comparable to the sense of doubling or halving the volume obtained through hearing. Hearing perception only detects a change of 10 dB SPL as a doubling of the volume. The unit of loudness is the sone, and it is defined by stating that a loudness of 1 sone is equivalent to the loudness of a 1 kHz tone at 40 dB SPL (Pulkki & Karjalainen, 2015). The unit of loudness level is the phon, and that is defined so that at 1 kHz the sound pressure level in dB and the loudness level in phons have the same magnitude (Pulkki & Karjalainen, 2015). In addition, the perception of loudness also depends on the pitch, spectral content, and duration. The perceived loudness increases when the band-pass spectrum is made wider while keeping the sound pressure constant. Our hearing analyzes the loudness of different auditory bands separately and adds the obtained results to form a total loudness for the system, which then can add up to more than it should be in principle. In essence the singing voice consists of sounds whose spectrum changes all the time, so it would be logical that the perceived loudness would constantly change with it. However, it has been found that loudness seems to be the result of a short-term integration of sound power followed by the processing of the peak values by a different processing system. The perceived loudness can therefore be defined by loudness peak values (Barrichelo, Heuer, Dean, & Sataloff, 2001; Lee et al., 2008; Pulkki & Karjalainen, 2015; Sundberg, 2001; Suomi, 1990).

2.3.2.3 Alpha ratio – sound balance

The alpha ratio is a ratio developed by Frøkjaer-Jensen and Prytz (1976), which describes the slope of the spectrum. Originally it was calculated from the long-term average spectra as

$$\alpha = \frac{intensity\ above\ 1000Hz}{intensity\ below\ 1000Hz}$$

and it reflects the mean strength of the higher spectrum partials as compared to the lower ones. We use the slightly adapted version of calculation using the formula SPL

[2] The American National Standards Institute – Acoustical Terminology. This American National Standard provides definitions for the terms used in acoustics and electroacoustics.

1500-5000 Hz – SPL 50-1500 Hz, which has been found more suitable for investigating the singing voice (Frøkjaer-Jensen & Prytz, 1976; Hakanpää, Waaramaa, & Laukkanen, 2021a; Lã & Sundberg, 2012).

The value of the alpha ratio is usually negative due to the fact that the higher sound partials are weaker than the lower ones. Typically, the alpha ratio in speech falls between about -10 dB and -30 dB, with a higher number representing a loud voice or pressed sound quality and a lower number corresponding to a breathier sound. Thus, the alpha ratio numerically describes the distribution of sound energy across frequency bands and is perceptually related to sound qualities related to vocal fold adduction (Guzman et al., 2015; Laukkanen & Leino, 2001). The alpha ratio increases if the high-frequency content of the voice increases; therefore, it is strongly influenced by variations of vocal loudness (Sundberg & Nordenberg, 2006).

2.3.2.4 Harmonics-to-noise ratio (HNR) – clarity of sound

The HNR is an assessment of the ratio between periodic components and the non-periodic component comprising a segment of voiced speech (Murphy & Akande, 2007). The periodic component is derived from the vocal fold vibration and the non-periodic from the glottal noise, expressed in dB. If the signal is 99% periodic with 1% noise, the HNR is calculated as $10\log_{10}(99/1) = 20\,dB$ (Boersma & Weenink, 2014). An HNR measure of 0 dB indicates an equal amount of harmonic content and noise. Typical harmonicity for sustained [a:] is around 20 dB. Lower values can be interpreted as an indication of hoarseness. In this way, the HNR can indicate the quality of sound (Boersma & Weenink, 2014).

The evaluation between the two types of components reflects the efficiency of singing. If there is a lot of noise present in the signal, the sound will be more on the breathy side, and the signal-to-noise ratio will be low. The higher the decibel value, the more the periodic signal is present in the sound, and the sound will be clearer. Perceptually, a high HNR value refers to a clear voice from which, for example, the pitch is easier to perceive (Laukkanen & Leino, 2001; Van Puyvelde et al., 2018).

2.3.2.5 Formant frequencies – sound timbre

Formants are distinctive frequency components of the acoustic signal produced by singing (or speech). They are the broad peaks or the local maxima in the spectrum that result from acoustic resonances of the human vocal tract. When the nasal and

oral cavities remain in the same position, they will amplify the same frequency ranges regardless of the fundamental frequency of the source. The frequency response of the vocal tract filter represents the effect that the vocal tract shape would have on any sound that travels through it (Story, 2016). The term "broad peak" is usually associated with the radiated spectrum (the produced sound itself), and "local maximum" is used when the term *formant* is used to describe the production mechanisms of sound including both the vocal tract transfer function and the sound source (Titze, 1994; Titze et al., 2015; Vurma, 2020). The current ANSI (2020) definition for formant is "a range of frequencies in which there is an absolute or relative maximum in the sound spectrum. The frequency at the maximum is the formant frequency." In this study, formant is defined as the peak of enhanced spectral energy in the output spectrum and resonance as the natural frequency of the vocal tract (Story, 2016; Titze et al., 2015).

The interruption of airflow by the vibrating vocal folds creates a sound pressure wave that consists of the fundamental frequency and multiple overtones, which are called harmonic partials or harmonics (*fo* is the first harmonic, and it is denoted as 1*fo*). Partials radiating from the glottis are evenly spaced (they are multiple integers of *fo*), and they decrease in amplitude the faster they oscillate (the higher their frequency).

As this sound pulse (all the partials of the emitted sound) moves through the air molecules in the vocal tract, it changes as it hits obstacles (narrow and wide spaces). The obstacles are put in place by moving articulators, and they constantly shape the voice source spectra during singing. The four most influential vocal tract articulatory chambers are the epilarynx, the pharynx, the oral cavity, and a small area located just behind the lower front teeth (in the oral cavity). When sound travels through these chambers, it can be reinforced or dampened. The larger the chamber, the lower the partials that are amplified, and conversely, the smaller the chamber, the higher are the amplified partials. By shortening, lengthening, narrowing, or expanding any of these chambers or by shortening or elongating the whole vocal tract, the resonance frequencies of the vocal tract and thus the frequency regions that are amplified within the spectrum are moved lower or higher within the spectrum. When one or all of the frequency regions is changed, the perceived voice quality also changes (Welch, Thurman, Theimer, Grefsheim, & Feit, 2000).

The product leaving the lips is called the radiated spectrum, and it is from this transformed spectrum that the formant frequencies can be detected. The output pressure spectrum is the combination of the glottal flow spectrum (source spectrum)

and frequency response of the vocal tract. Formant frequencies can be estimated from the frequency spectrum of the sound using a spectrum analyzer. Estimating acoustic resonances of the vocal tract can be done with linear predictive coding, and an intermediate approach can be taken by eliminating the fundamental frequency from the spectral envelope and then looking for the local maxima in the spectral envelope (Laukkanen & Leino, 2001; Pulkki & Karjalainen, 2015). The amplitude at each frequency in the output spectrum is the sum of the amplitudes of the source spectrum and filter resonances. The fundamental frequency and all of the overtones are present in the output spectrum, but their amplitudes have been modified by the vocal tract resonances; harmonics near a formant frequency are enhanced in amplitude, while those distant from the formants are suppressed (Story, 2016).

The ability of the vocal tract to transfer sound is increased near and between two formants if the frequency distance between these formants is decreased either by voluntary or involuntary articulatory movements. This is what happens in the so-called singer's formant, where formants F3-F5 move closer to each other to form a formant cluster. When two or more sounds contribute to the same sound field, they can affect the total field in a different way. As the broadbands for filter resonance are quite broad, the formants can coexist in the same resonating frequency area, making their pressure values add up. This phenomenon is called the linear superposition of waves. If the sound sources are coherent, meaning that they or their partials have the same frequencies, then depending on their phase they can either add constructively (same phase) or destructively (opposite phase). This corresponds to a 6 dB increase/decrease in sound level. If the sound sources are incoherent, meaning that their frequencies do not coincide, the powers of the signals are summed, adding up to a 3 dB increase/decrease in sound level (Pulkki & Karjalainen, 2015; Sundberg, 1987; Titze, 1994; Titze et al., 2015; Welch et al., 2000). The sound at the formant frequencies is generated by one sound source, which means that even though the sound itself has many partials, they all exist at different frequencies and thus do not coincide with each other. This means that the extra power gained with clustering the formants would give the singer a 3 dB advantage in sound projection when using the "singer's formant." The effect the singer's formant has on loudness is greater than its effect on SPL due to the sensitivity of hearing in the frequency area of the cluster (Pulkki & Karjalainen, 2015; Sundberg, 1987; Titze, 1994; Titze et al., 2015; Welch et al., 2000).

2.3.2.6 Jitter/shimmer – hoarseness, noise

Jitter and shimmer measures indirectly assess laryngeal function by quantifying acoustic correlates of irregular vocal fold vibration. Jitter measures *fo* perturbation and shimmer measures amplitude perturbation caused by vibratory variations from one vocal fold cycle to the next (Boersma & Weenink, 2014). Jitter is caused mainly by the lack of control of vocal fold vibration and shimmer by the instability of SPL-related tension in the vocal folds (Farrús & Hernando, 2009). When one glottal pulse differs somewhat in shape, peak amplitude, and duration from neighboring pulses, even when *fo* and other phonatory conditions are held constant, the spectral consequence is a perturbation of formant amplitudes related to the changes in the zero-pattern of the source spectrum. This is seen as a factor contributing to the voice quality (Fant, 1986).

Although sound perturbations refer to instability of the period length, amplitude, and waveform of the vocal fold cycle, the sound is somewhat tolerant of the wave's asymmetry. Small irregularities in the acoustic wave are considered as normal variation associated with physiological body function and voice production (Brockmann, Drinnan, Storck, & Carding, 2011; Orlikoff & Baken, 1989; Titze, 1991). In fact, some perturbation occurs in every natural sound, and it is even important for a sound to be perceived as a normal human voice. The excessive occurrence of jitter or shimmer in voiced sound can be perceptually described as a hoarse, husky, or rough voice (Boersma & Weenink, 2014; Guzman et al., 2012; Laukkanen & Leino, 2001; Laver, 1980; Seikel, King, & Drumright, 2009).

Brockman et al. (2008) investigated jitter and shimmer values in healthy speakers under different loudness conditions and found that both jitter and shimmer increase when SPL decreases. Below 80 dB, even small changes in SPL affected jitter and shimmer values significantly, which was thought to be caused by small intrinsic vocal fold tension and vibration amplitudes potentially causing more variability in mucosal movement and therefore resulting in higher perturbation (Brockmann, Storck, Carding, & Drinnan, 2008).

Under clinical conditions, jitter and shimmer are usually measured from a sustained vowel at a comfortable loudness and pitch level. Sustained vowels are used because for the algorithms to work, they need a long enough sample with harmonic contents. The vowel /a/ has been found to enhance measurement reliability (Brockmann et al., 2011).

2.3.2.7 Frequency and amplitude modulation – vibrato

Vibrato in everyday language refers to a modulation of fundamental frequency, and it has two primary acoustic attributes: frequency and amplitude. In voice studies, we usually differentiate two kinds of vibratos; the *fo* vibrato, which is characterized by an undulation of the fundamental frequency, and amplitude vibrato, which is characterized by a pulsation of subglottal pressure, the frequency variation of formants, or the frequency variation itself (Sundberg, 1994).

The *fo* vibrato is thought to be produced by the pulsating contractions of the cricothyroid muscle. It is characterized by four parameters: the rate and the extent of undulations, the regularity of the undulations, and the waveform of the undulations. The rate specifies the number of undulations per second and the extent describes how far phonation frequency rises and falls during a vibrato cycle. The regularity of vibrato is considered a sign of a singers' vocal skill – the more regular the undulations the better. The waveform corresponds more or less to a sinewave. The other, perceptually different type of vibrato that is characterized by a pulsation of subglottal pressure is sometimes referred to as the "hammer vibrato" (Sundberg, 1994). This type of vibrato is produced with a tremolo mechanism by rapidly alternating abduction and adduction of the vocal folds (Bunch, 1997; Sundberg, 1987; Sundberg, 1994).

There are three different sources of amplitude modulation that can be identified in vibrato: 1) The frequency variation itself – as the frequency undulates, so do the frequencies of its partials. This can have an effect on their amplitude as they get closer and further from the resonance frequencies of the vocal tract. 2) The characteristics of the voice source due to a pulsation of the subglottal pressure or the glottal adjustment may cause amplitude variation. 3) Rhythmical variations of the vocal tract shape can cause formant frequencies to move slightly up and down in frequency (Bunch, 1997; Sundberg, 1987, 1994). The amplitudes of all spectrum partials are determined by their distance from the formant frequencies. If the partials approach a formant or *vice versa*, the amplitude of the partials will increase. Bearing this in mind, we can say that the frequency variation of the vibrato is sufficient to induce variations of the overall amplitude, as the phase relationship between amplitude and frequency vibrato depends on the frequency relationship between the first formant and the strongest spectrum partial (Sundberg, 1994).

The perceptual significance of amplitude vibrato is not great, however. The concept is perhaps empirically useful in differentiating the aforementioned hammer vibrato, tremolo, and trillo (performed by cutting off the airflow all together) from

other types of vibratos. However, the main perceptual effect of the vibrato is dependent on the frequency modulation.

The vibrato is characterized by a fundamental frequency undulation ranging from 5 to 8 Hz in rate and less than ±1 semitone in extent. The extent of the vibrato undulations varies with the loudness of phonation. The vibrato rate is often constant within the individual singer, and it is dependent on factors such as the age, gender, physical activation level, and the emotional involvement of the singer. Female singers generally have a slightly faster vibrato than males (Sundberg, 1994). The limits for the rate and extent of an acceptable vibrato are quite narrow. A rate that is slower than 5 undulations per second tends to sound unacceptably slow, and vibrato rates exceeding 8 undulations per second tend to sound nervous. Similarly, vibrato rates over ±2 semitones are generally regarded as awkward (Sundberg, 1987).

There have been many attempts to link vibrato with emotional expression in singing. Previous studies have suggested that in high energy emotion expression, the vibrato rate and extent are higher, and that in low energy expressions, they tend to be lower. The extent of the vibrato undulations varies however with the loudness of phonation, and that makes it difficult to distinguish if it is the emotion causing the changes in the vibrato pattern or simply the SPL (Dromey, Holmes, Hopkin, & Tanner, 2015; Eyben, Salomão, Sundberg, Scherer, & Schuller, 2015; Guzman et al., 2012; Park, Yun, & Yoo, 2010; Seashore, 1923, 1931).

2.3.2.8 Attack, sustain, & release – amplitude envelope of sound

Human hearing is very sensitive to temporal changes in sound, as the attributes of natural sounds typically change with time. A simple way to depict the temporal variations in voiced sound is to plot the moment-by-moment amplitude. This variation of amplitude is called the envelope of sound. This research utilizes three instances of volume against time: 1) attack, which represents the phase when the amplitude of the voice sample rises (and reaches its peak); 2) sustain, which represents a constant amplitude phase; and 3) release, a decrease of the amplitude until silence.

Sundberg (2000) found that tone onsets could be important in encoding emotional information. He suggested that a delayed syllable starting with an elongated consonant before the following vowel could be one way of adding emotional information to the voice (Sundberg, 2000). In this study, voice onset is defined as the time interval between the release of a plosive/fricative and the

beginning of vocal fold vibration associated with the subsequent vowel. It gives information about the coordination of the articulatory system.

2.4 Anatomy and physiology

When we work our way up from the sound parameter level, we stumble into anatomy and physiology. How exactly do we "make" the voice quality? What are the systems that make these sounds and how do they work?

2.4.1 Systems of singing

A system is defined as a functionally defined group of organs. The singing voice capitalizes on the organic functions of other systems, which often serve a more important role in the survival of the species (such as the respiratory system). We can, however, categorize the systems of speech into 4-5 different parts: the respiratory, phonatory, articulatory, resonatory, and nervous systems.

The *nervous system* (NS) is used to sense and control bodily events, and it controls the muscle and the brain work needed for singing. Since singing is bodily in a secondary function – i.e., singing capitalizes on anatomical structures originally or primarily intended for other functions – there are sometimes some discrepancies between how we would want the NS to work in singing and how it actually works.
The anatomical components of respiration for voice are one with *the respiratory system*, i.e., they include the oral, nasal, and pharyngeal cavities; the trachea, bronchi, and lungs; the muscles that move them; the surrounding bones, cartilage, tendons, and other tissues. *The phonation system* includes the larynx and its function. The *resonatory system* can be viewed as consisting of the nasal cavity, soft palate, and parts of the respiratory and digestive tract (such as the oral cavity and tongue) (Seikel, King, & Drumright, 2009). *The articulatory system* uses the anatomical structures of the respiratory and digestive systems (such as the tongue, lips, teeth, soft palate, oral cavity, etc.).

The acoustic features of the human voice are always defined by individual biology. However, the human anatomy and physiology of individuals is adequately similar to enable generic instruction on how to use the singing voice. The basic principles of voice production are the same for all human beings, regardless of their voice range or the style of music they are singing. This being said, it must be stressed that different music genres require different techniques and the way that the voice

instrument is delicately tuned to match the aesthetic demands of a music style is often quite limiting towards the way the instrument can be used. Things are further complicated by the fact that the same anatomical structures can participate in the operation of different systems and have a different function depending on the objectives set for them. In the work of a singing teacher, this is reflected, for example, in the fact that the same exercise can act as an optimal vocal breathing practice in one context and as articulation training in another. The good thing about this is that these systems can be separated and combined as needed depending on the student's individual needs.

2.4.2 Nervous system

Voice production is a psychophysiological process influenced by environmental and internal challenges. Both processing and producing singing voice sounds rely on the cooperation of approximately 100 muscles, which are innervated by a network of cranial and spinal nerves as well as cortical and subcortical parts of the brain and cardiorespiratory processes (Van Puyvelde et al., 2018). During singing, we control breathing by shortening inspiration and lengthening expiration. Both respiratory and laryngeal muscles are controlled by the efferent fibers of the Nervus vagus, which divide into pharyngeal, superior laryngeal and recurrent laryngeal nerves. The recurrent branch of the Nervus vagus carries the motor signal to the adductors and abductor of the intrinsic laryngeal muscles, which is responsible for closing and opening the glottal gap. The superior laryngeal nerve innervates the cricothyroid muscle, which is involved in vocal fold stretching (pitch control) (Câmara & Griessenauer, 2015; Seikel et al., 2009).

Recent studies have indicated that vocal expression and musically expressed emotions can elicit activity in the amygdala and hippocampus. Both vocal emotions and musical emotions involve similar brain mechanisms for decoding and responding to emotional cues. The amygdala is thought to provide the affective evaluation of the stimulus and contribute to the emotional reaction (fast and pre conscious evaluation), whereas the hippocampus is thought to be responsible for memory encoding (the cognitive route) (Frühholz, Trost, & Grandjean, 2014).

2.4.2.1 The neural system in making art

Artistic emotion is not a spontaneous emotion in the same way as a stimulus-induced reaction in everyday life (Williams & Stevens, 1972). The artistic representation of spontaneous emotion is not necessarily a one hundred percent conscious process but may be instinctively approached by activating subcortical areas of the brain (Davis et al., 1996). Even though emotion and emotion expression are two different things, when we are asking someone to express emotions, we cannot fully compartmentalize actual emotion from the artistic expression. Sometimes it happens that students get stuck with their actual emotions when trying to portray an artistic one.

If we view the voice output as a psychophysiological response to the human integrative psychophysiological stress system, as we do when we argue that emotions change the way we use our voice, we can see that the activation of these responses is dependent on individual features, such as individual anxiety traits and stress-coping mechanisms as well as the environmental challenges the singers are confronted with. This can make the pedagogical usage of neural networks difficult in a teaching situation.

2.4.2.2 The neural system in learning

The neural system has the capacity to change in response to experience. The concept behind this plasticity is that an impulse is transmitted from one neuron to another via the axon of the sending neuron. Because there is a synaptic gap between the axon and the dendrite, the sender must secrete a neurotransmitter, which diffuses across the synaptic gap and stimulates the receiving neuron. The two key ideas regarding learning are that the change in the synapse is the neural basis of learning and that the effect of this change is to make the synapse more (or less) efficient (Nolen-Hoeksema et al., 2009). Therefore, it is beneficial for learning to repeat and build information. Repetition is the key component of skill acquisition.

One of the most important neural structures for social cognition is the mirror neuron system found in the 1990s. Mirror neurons were found to respond both when a monkey performs an action and when it observes, hears, or knows that someone else is performing a similar action (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Keysers et al., 2003; Kohler et al., 2002; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996; Umiltà et al., 2001). This is a system that allows us to understand other individuals. The brain can feel

other people's actions through a process of motor embodiment – in other words, by feeling what one would feel during the execution of a similar action. The dorsal and ventral premotor, supplementary motor, posterior parietal, temporal, and somatosensory cortices are activated during both action execution and action perception. Broca's area is also constantly active in the observation of action sounds (Gazzola, Aziz-Zadeh, & Keysers, 2006). It is widely agreed upon that mirror neurons transform the sight or sound of action into a corresponding motor representation (Keysers & Fadiga, 2008). This means that the motor representation has the same goal as the action observed. Its function is to help us understand what another individual did and how they did it. By influencing the observer's actions, the mirror neuron system creates a reciprocal bond between the behaviors of social partners. Both partners are both agent and observer, both the target and the source of social contagion. The mirror neuron system is proof that social cognition is at least partially based on our own bodily representations (Keysers & Fadiga, 2008). The perception of other people's emotions activates regions involved in experiencing similar emotions and producing similar facial expressions. In addition, the sight of emotional body movements leads to similar actions in the observer (Jabbi, Swart, & Keysers, 2007; Pichon, de Gelder, & Grèzes, 2008; Singer et al., 2004; van der Gaag, Minderaa, & Keysers, 2007; Wicker et al., 2003). The existence of such a neural network has implications for both performance and learning situations. Pichon et al. (2008) investigated neural reactions to portrayals of anger performed by actors and found that brain areas coupled with autonomic reactions and motor responses related to defensive behavior were activated in the test participants. This means that the mirror neuron system is active also in acted emotion expressions and most likely striving towards an authentic performance as possible would increase its function. In a learning/teaching situation, it has a different role as a facilitator of communication.

2.4.3 Respiratory system

The respiratory system consists of a gas-exchange mechanism supported and protected by a bony cage. The lungs are situated within the thorax, which is a bony structure suspended from the vertebral column. The thorax consists of the ribs and sternum. The ribs slope downward when the rib cage is inactive but elevate as breath is drawn in to facilitate the increase in lung capacity. The lungs expand as a result of enlargement of the structure surrounding them, because they are attached to the

structure via the pleural lining. This lining provides a mechanism for translating the force of thorax enlargement into inspiration (Seikel et al., 2014; Watson, 2019).

The primary muscle of inspiration is the diaphragm, which attaches to the lower margins of the rib cage, sternum, and vertebral column, and it separates the thoracic and abdominal chambers from each other. When the muscle contracts, the muscle-fibers shorten and pull the central tendon (the interior part of the diaphragm formed of the leafy aponeurosis) downward and forward. Contraction of the diaphragm enlarges the vertical dimension of the thoracic cavity, while elevating the rib cage enlarges the transverse dimension. The most important elevators for the rib cage are the external intercostal muscles. They provide a significant proportion of the total respiratory capacity, and they have functions that are uniquely speech-related (Seikel et al., 2014; Watson, 2019).

The lungs are composed of a porous tissue that is highly elastic, like a sponge. Their natural tendency is to expand after being compressed and return to their original shape after being stretched. For instance, when inhaling, the rib cage rises through muscular activity and the abdomen might protrude because the downward force of the diaphragm is stretching the abdominal muscles. When exhaling, the inspiratory muscles of the rib cage and abdomen relax and return to their resting position. When the abdominal muscles relax, they tend to push the abdominal viscera back in and force the diaphragm up. In addition to elasticity, gravity is the second force acting in support of passive expiration. When in an erect position, gravity helps the ribs to return to their normal position after the muscles of inspiration have expanded the rib cage for inspiration. One final force which could also assist in expiration is torque, but this force is only active after significant inhalation reaching over 60% of vital capacity (not an unusual number when singing long phrases and using a loud singing voice) (Lã & Gill, 2019; Seikel et al., 2014; Watson, 2019).

2.4.3.1 Respiratory airflow in singing

Humans are capable of two types of breathing: passive and active. Quiet inspiration followed by passive expiration is the most common style of breathing, which we normally use most of the time when nothing special is happening. The other type is forced inspiration coupled with active expiration. This type of breathing is in use when we need to vocalize or when we do something physically straining. It is also possible to do quiet inspiration and active expiration (e.g., in the case of repelling

something from the airways using a cough) or forced inspiration with passive expiration (e.g., breathing exercises in meditation).

In order to understand respiratory airflow in singing we need to have some knowledge of lung volumes and capacities. Volumes refer to the amount of air the lungs can hold, and capacities refer to combinations of volumes that express physiological limits (Seikel et al., 2014).

Tidal volume is the amount of air humans breathe in during a respiratory cycle (inhalation + exhalation). Tidal volume (TV) varies as a function of age, body size, and physical exertion (meaning that if we are doing something strenuous, TV increases markedly). Inspiratory reserve volume (IRV) is the volume that can be inhaled after a tidal inspiration, and expiratory reserve volume (ERV) is the amount of air that can still be expelled after a passive tidal expiration. Residual volume (RV) refers to the air that remains in the lungs even after exhalation. Humans also have a little dead space air residing in the upper respiratory passageway and the conducting passageways to the lungs. These volumes can be combined in multiple ways to characterize the way we use our breathing system (Seikel et al., 2014; Watson, 2019).

The vital capacity (VC) is most often sited in our literature, because it represents the capacity available for speech and singing (VC=TV+IRV+ERV).[3] The functional residual capacity (FRC) is the volume of air remaining in the lungs after passive exhalation (FRC=ERV+RV), total lung capacity (TLC) is the sum of all the volumes, and inspiratory capacity (IC) is the maximum inspiratory volume possible after tidal expiration (Seikel et al., 2014; Watson, 2019). During singing, both inspiration and expiration (and thus the lung volumes and capacities) are facilitated through muscle activity. An experienced singer can adjust the volume of each breath to correspond with the length and demands of the sung phrase. In this way, the singer ensures that there is enough air to sing through the phrase on the one hand, and that the excess air does not have to be expelled at the end of the phrase, before the next inhalation, on the other hand (Watson, 2019).

Five different types of pressure are classified for speech and non-speech functions: alveolar pressure, intrapleural pressure, subglottal pressure, intraoral pressure, and atmospheric pressure. The atmospheric pressure is our constant zero against which we can compare respiratory pressures – it is our reference point. The subglottal pressure (P_{sub}) is the pressure below the vocal folds and the intraoral pressure (P_m) is the pressure within the mouth. As stated before, the pressure

[3] Resting lung volume (ERV) is 38% of vital capacity in the average adult, while tidal volume takes up only about 15% of VC. The missing 47% is covered by inspiratory reserve volume (IRV) (Seikel et al., 2014).

measurements are made relative to the atmospheric pressure, which usually in our concept of the singing voice correlates with increased values due to muscular effort, which can generate pressures above and beyond atmospheric pressure. To initiate the singing voice in atmospheric pressure, it needs to generate enough power to make a pressure change. When we close the glottis for voicing, a significant blockage is generated between the lungs and upper respiratory pathway. This will cause an immediate increase in the pressure below the vocal folds (P_{sub}) as the lungs continue expiration. Simultaneously, the intraoral pressure above the vocal folds drops to near atmospheric pressure. This results in a large difference between the subglottal and supraglottal spaces. If this pressure difference below the vocal folds exceeds 3-5 cm H_2O, the vocal folds will be blown open and the voicing will begin. Respiration for speech and singing requires maintenance of a relatively steady flow of air at a relatively steady pressure. Research has indicated that for conversational speech, we keep our VC near the ERV regions (35-60% of VC), but for loud voice use, we need to inspire more (>80% of VC) (Herbst, 2017; Lã & Gill, 2019; Seikel et al., 2014; Watson, 2019).

We know that in inspiration, we exert a force to overcome gravity and the elastic forces of tissue. It is usually an active movement, requiring muscular action to complete it. Expiration on the other hand capitalizes on elasticity and gravity to take back some of the energy "wasted" during inspiration. When the muscles of inspiration contract, they stretch tissue and distend the abdomen (Iwarsson, Thomasson, & Sundberg, 1998; Seikel et al., 2014). When breathing in, the muscles and bones move to make more room for air, and when breathing out, these muscles and bones return to their normal state. These restoring forces generate recoil pressures themselves. During lung inflation, the alveoli[4] increase in volume, and energy must be used to overcome the surface tension of the liquid film that coats their walls. When the muscles of inhalation relax, surface tension causes the alveoli to collapse. This phenomenon is known as elastic recoil, and it contributes markedly to subglottic pressure at high lung volumes (Watson, 2019). "Recoil of the chest obeys the laws applying to any elastic material: The greater you distend or distort the material, the greater is the force required to hold it in that position and the greater is the force with which it returns to rest" (Seikel et al. 2014 p. 165). Titze et al. (2012) claim that it is this recoil pressure that trained vocalists use as the main component of lung pressure in speaking and singing. The relatively high rib cage position

[4] Small air sacs at the end of the terminal bronchioles of the lungs, which are responsible for gaseous exchange. Oxygen diffuses through the liquid film that lines the alveoli into the bloodstream and binds to the red blood cells and hemoglobin (Watson, 2019).

employed by most professional singers allows for the recoil pressure to be utilized over a significant part of the expiratory phase. It lessens the effort one has to use in expiration for voicing but requires more work in the inspiration phase to inflate the lungs against the elastic recoil (Titze & Verdolini Abbot, 2012).

Salomoni et al. (2016) investigated the respiratory kinematics of the singing voice on Classical singers and non-singers using respiratory inductance plethysmography bands and pneumotachograph. They found that while both groups adapted to rhythmical constraints with a decreased time of inspiration and increased peak airflow, the Classical singers altered their coordination of the rib cage and abdomen more during singing. Classical singers used a greater percentage of abdominal contribution to lung volume (i.e., stronger activation and deeper descent of the diaphragm) during singing and greater asynchrony between the movements of the rib cage and abdomen when phonating (moving the abdomen inward prior to phonation). They used inhalation maneuvers that involved complex movements of the respiratory compartments. The abdominal wall began to rise before the rib cage and abdominal peak volume occurred earlier than rib cage peak volume at the transition from inhalation to phonation. As a consequence, the inward movement of the abdominal wall muscles started before the maximum lung volume was reached (Salomoni et al., 2016). In another (real-time MRI) study of professional singers, Traser et al. (2016) found that during exhalation in normal breathing, the diaphragm and the rib cage moved synchronously to reduce lung volume, but during phonation different functional units could be identified. In phonation, the reduction of lung volume was mainly generated by the elevation of the posterior and middle part of the diaphragm, while the anterior diaphragm and the rib cage remained in a more inspiratory position at the beginning of the phrase. When reaching 50% of the maximum phonation time, the movement of posterior and the middle part of the diaphragm slowed down and simultaneously the elevation of the anterior part of the diaphragm and the lowering of the rib cage increased their movement velocity (Traser et al., 2016). They suggest that concerning phonation, the anterior diaphragm and the rib cage could be regarded as one functional unit and the middle and posterior parts of the diaphragm another.

The tricky part of breath regulation for singing is that there is no one correct way to do it. Studies have continuously identified different breath management patterns in professional singers (Lam Tang, Boliek, & Rieger, 2008; Salomoni et al., 2016; Traser et al., 2016). Breathing patterns in individuals repeating the same phrase have been found to be very consistent, but the breathing patterns differ greatly between individuals (Lã & Gill, 2019; Thomasson & Sundberg, 1999; Thorpe et al., 2001;

Watson, 2019). Furthermore, the singing can be judged to be poorer if the singer is instructed to use a breathing pattern that is not habitual to them (Collyer et al., 2011). This makes the teaching of breath management pedagogically challenging. We can, however, move forward with a notion that the main function of the respiratory system during singing is to regulate subglottic pressure.

2.4.3.2 "Support"

The supported voice is associated with greater subglottal pressure, greater sound pressure, and higher peak airflow (Salomoni, Van Den Hoorn, & Hodges, 2016). It is common practice among voice teachers to explain the physiology of "support" by describing it as an activation of the inhalation musculature in the exhalation phase. Perceptually this makes sense, as the body expands during inhalation and then tries to stay expanded for as long as it takes to sing a phrase.

If we view the breath cycle as consisting of four different phases – inhalation, preparation for the onset of sound, exhalation (with voicing), and release before the next breath – we can deduce that support is working in half the cycle. Abdominal movement is usually greater during phonation than during quiet breathing (Estenne, Zocchi, Ward, & Macklem, 1990; Macdonald, Rubin, Blake, Hirani, & Epstein, 2012; Watson, Hoit, Lansing, & Hixon, 1989). Projected singing requires even more abdominal muscle involvement (Macdonald et al., 2012; Thorpe et al., 2001; Watson, Williams, & James, 2011). Contracted abdominal muscles prevent the rising of the diaphragm during singing and provide the opposing force that allows the rib cage to create the strong subglottal pressure needed for louder voice use and pitch generation (Pettersen & Westgaard, 2004; Watson & Hixon, 1985). The abdominal wall and the rib cage have been found to move independently and asynchronously during phonation. This can lead to a paradoxical movement where the rib cage volume increases during the exhalation/phonation phase of the breath cycle (Watson & Hixon, 1985). Sundberg et al. (1987) have shown that singers use their diaphragms in coactivation with the intercostals to secure a greater speed and precision of pressure changes, maintaining their diaphragmatic contraction over a sung phrase, therefore lending support to the idea of using the inhalation muscles in the exhalation phase to give support to the voice (Leanderson, Sundberg, & Von Euler, 1987). Control of the amount of P_{sub} at the start of a sung phrase is necessarily brought about by the muscles that control the fall of the chest or the ribs, such as the external intercostals, scalenes, or sternocleidomastoids, i.e., the muscles of

inhalation (Watson, 2019; Watson et al., 2011). Seikel et al. (2014) summarize the control of air pressure in voicing as follows:

> "The pressure we generate for voice use is a direct function of the forces of expiration, which are generally passive above 38 percent of VC. Because we operate almost exclusively within the region above 38 percent of VC, generation of subglottal pressure is mostly a function of the forces of expiration, specifically elasticity and gravity. We must use the muscles of inspiration to check that outflow of air, so the muscular effort in speech beyond the initial inspiration is dominated by using the muscles of inspiration again to impede the outflow of air. By delicate manipulation of these muscles, we are able to maintain a constant subglottic pressure that can be used for the fine control of phonation by the larynx." (Seikel, Drumright, & King, 2014, p. 164)

Some pre-phonatory posturing is also needed prior to the onset of phonation. The subglottic pressure must be raised to the level necessary to support the sung phrase. According to Watson (2019) this is done by engaging the abdominal muscles (especially the lateral abdominal muscles) to push the diaphragm upwards (Watson, 2019). The chest and the abdomen need to be brought to an optimal configuration for the onset of phonation(Iwarsson & Sundberg, 1998; Iwarsson et al., 1998).

In order for the support system to stay supple, many singers and singing teachers advocate for a proper release of muscle tension (built up in the body during supported singing) before every inhalation (see e.g., Chapman, 2006).

Commonly in the later part of the 20th century, singing teachers adhered to either the "belly-in" or the "belly-out" schools of how to teach support. In the "belly-in" movement, the emphasis is on keeping the chest high and stable, leaving the diaphragm and abdominal muscles to draw air in or push it out. In the "belly-out" strategy, less emphasis is given to chest stability and more to maintaining abdominal pressure. The abdominal wall is allowed to expand on inspiration and remain relatively expanded at the onset of phonation (Watson, 2019). The differences between these two support techniques have been found to be subtle at best. Thomasson (2003) found no significant differences between the belly-in and the belly-out techniques in subglottic pressure or airflow through the glottis (Thomasson, 2003). Furthermore, subjective comparisons by a panel of expert vocal judges in perceptual analyses regarding the output of belly-in and belly-out voicings failed to demonstrate a consistent influence on vocal quality between these two strategies (Collyer et al., 2011). This does not mean that the voice teachers of the late 20th century were all wrong in teaching support techniques; instead, it seems to suggest that in the future it might be good to view these strategies as two ends of a continuum rather than an either/or type of selection.

As already shown e.g. by Draper, Ladefoged & Whitteridge 1959, modern research views the concept of support as an interaction between different subsystems (Draper, Ladefoged, & Whitteridge, 1959; Herbst, 2017; Lã & Gill, 2019; Laukkanen & Leino, 2001; Watson, 2019). To sustain a note or a phrase at a constant dynamic, we need to maintain a constant subglottic pressure for its entire duration. This demands a continuous adjustment in the level of inspiratory and expiratory muscle activity. When we start to sing, the elastic recoil forces generate a pressure that is generally too high. This pressure needs to be counteracted by gradually decreasing the activity level of the inspiratory muscles (e.g., external intercostal muscles) until the recoil forces equal the required pressure. Subsequently, activity level in the expiratory muscles (e.g., internal intercostals) must rise to maintain the pressure at the required level (Sears & Davis, 1968; Watson, 2019). Iwarsson et al. (1998) found that the vertical larynx position can be influenced indirectly by the pulmonary system so that when the lung volume is high (and the diaphragm low), the larynx position tends to be lowered as well (Iwarsson et al., 1998). This can promote the so-called 'flow phonation' with relatively large airflow peak amplitude during vocal fold vibration (Sundberg, Leanderson, & von Euler, 1986). This was also confirmed by Iwarsson et al. 1998.

Singers make constant adjustments in the glottal and laryngeal configuration when producing the supported voice (Griffin, Woo, Colton, Casper, & Brewer, 1995). Adduction plays an important part in generating P_{sub}. The main determinant of the glottal flow resistance is the degree of glottal flow adduction (Alipour, Scherer, & Finnegan, 1997; Grillo & Verdolini, 2008). Furthermore, the vocal tract adjustments can influence the behavior of the voice source by means of nonlinear source-tract interaction(Rothenberg, 1981; Titze, 2008b). The key question in defining what support in singing means is whether it should be defined only through the function of the pulmonary system, or whether other voice subsystems (such as laryngeal function and the resonant characteristics of the supraglottal vocal tract) should be included in the definition as they likely contribute to the complex. Herbst (2017) promotes a model where support is seen through these three interactive subsystems. This question remains partially unanswered as a consensus has not yet been reached.

2.4.4 Phonatory system

Phonation is the process by which the vocal folds produce sounds through quasi-periodic vibration. In singing, we normally strive towards optimal sound balance where the phonation happens perfectly in line with our voicing target and in compliance with our individual anatomy and physiology. Controlled phonation requires well established regulation over several variables, like the balance between subglottal pressure and vocal fold approximation, tension, stiffness, and resistance. All this must be done during respiratory regulation throughout the expiratory phase of breathing for singing. (Alipour et al., 1997; Herbst, Hess, Müller, Švec, & Sundberg, 2015; Herbst, Howard, & Svec, 2019; Seikel et al., 2014; Titze, 2015; Zhang, 2015) The affective state and stress also affects phonation changing the habitual voice quality (Helou, Jennings, Rosen, Wang, & Verdolini Abbot, 2020; Helou, Rosen, Wang, & Verdolini Abbot, 2018; Johnstone, 2001; Zhou, Hansen, & Kaiser, 2001).

2.4.4.1 The role of vocal folds in phonation

The voice source is created by the vibrating vocal folds and their interaction with glottal airflow. The steady tracheal airflow is converted into a time-varying glottal flow via flow-induced self-sustained oscillations of laryngeal tissue (Chan & Titze, 2006; Herbst et al., 2019; Van den Berg, 1958). The vocal fold vibration is facilitated by a time-varying glottal shape and a delayed vocal tract response caused by an inert acoustic load, which also aids in closing the vocal folds (Herbst et al., 2019; Titze & Alipour, 2006; Zhang, 2016). To put it simply, the vocal folds create the sound source automatically, governed by the laws of physics (for further information, see Van den Berg, 1958), when the conditions (pulmonic airflow and laryngeal muscle activity) are favorable. Hence, one does not move one's vocal folds, they move themselves. The vocal folds attach to the cartilaginous structure formed by the thyroid, cricoid, and the arytenoid cartilages. The vocal folds are attached to the front, internal surface of the thyroid, at the point where the lateral plates fuse together. The posterior ends of the vocal folds attach to the vocal processes of the arytenoids (Seikel et al., 2014). The vocal folds have a layered structure with different biomechanical properties. The layers include the body of the vocal folds, consisting of the thyroarytenoid muscle and the deep layer of the lamina propria, and the cover part, consisting of the superficial and intermediate layers of the lamina propria and the epithelium (Hirano, Kakita, Ohmaru, & Kurita, 1982; Titze, 2000). The body and

the cover can move as separate units in vocal fold vibration (Hirano, 1974). In every glottal cycle, the airflow from the lungs sends up a traveling wave propagating on the surface of the vocal fold tissue, moving first in a down-up direction and then laterally on the surface of the vocal folds (Herbst et al., 2019; Kumar et al., 2020; Timcke, Von Leden, & Moore, 1958, 1959). The passive vocal fold oscillations are superfast and should not be confused with the active muscular positioning adjustments of the vocal folds that are done in order to regulate pitch, register, and quality of phonation (pressed-breathy) (Herbst et al., 2019).

There are three types of muscular tension relevant for voice production: longitudinal tension, adductive tension, and medial compression. Longitudinal tension relates to pitch changes, adductive tension is needed for moving the vocal folds closer to each other for phonation, and medial compression refers to the degree of force that can be applied by the vocal folds at their point of contact. Increased medial compression is a function of the increased force of adduction, and this is a vital element in vocal intensity change (Laver, 1980; Seikel et al., 2014).

Adductive tension is used when vocal folds move closer to each other to phonate. Biologically the adductive process is very important, as it protects our lungs from any foreign object entering them (Herbst et al., 2019; Seikel et al., 2014). There are two paired and one un-paired adductor muscles (lateral cricoarytenoids, transverse arytenoids, and the oblique arytenoid) and only one (paired) abductor (the posterior cricoarytenoid). The vocal folds themselves can aid in closing the glottis by contracting their lateral part, which helps the lateral cricoarytenoids to bring the vocal processes of the arytenoids together (Laver, 1980; Seikel et al., 2014).

The abduction part is fairly simple: the contraction of the posterior cricoarytenoid muscles pulls the muscular processes of the arytenoids back, which makes the arytenoids rotate, pivoting the other ends of the cartilage (vocal processes to which the vocal folds are attached to) outwards. This movement repeats every time we inhale (Laver, 1980). The adduction part has more variation. The (paired) lateral cricoarytenoids arising from the outer and upper surface of the cricoid cartilage and attaching to the muscular processes of the arytenoids can drag the arytenoids forward and turn them inward so that the vocal processes come together and adduct the focal folds along their entire length (Laver, 1980). The (un-paired) transverse arytenoid originates from the muscular process and lateral border of one arytenoid and attaches to the lateral edge of another arytenoid. It closes the glottis by drawing the arytenoids medially towards each other with a gliding action. This movement approximates the vocal folds, but it is considerably less adductive than the effect of the cricoarytenoids. This muscle action is a component force in generating medial

compression. The oblique arytenoid arises from the posterior base of the muscular processes and course obliquely up to the apex of the opposite arytenoid. When they contract, they pull the apexes medially, tilting the tops of the arytenoids towards each other, promoting adduction, enforcing medial compression and rocking the arytenoid (and the vocal folds) down and in (Laver, 1980; Seikel et al., 2014). Sometimes the transverse arytenoid and oblique arytenoids are viewed as one singular interarytenoid muscle.

Different parts of the glottis are important for different types of phonations. The glottis is the whole length of the opening between the true vocal folds, from the front angle of the thyroid cartilage to the back of the arytenoids. The part where the arytenoids are situated is referred to as the cartilaginous glottis, and the part where the thyroarytenoid muscles are situated is called the membranous glottis. The glottis has length and width, but it also has a vertical dimension (depth). The changing vertical thickness of the vocal folds from the outer wall inwards to the vocal ligaments at the edge of the glottal space mirrors the interaction of the different tensions that are exerted on and in the folds by the laryngeal musculature. The way the glottis is shaped by these tensions is one of the factors in differentiating the modes of phonation (Laver, 1980).

The glottis is divided into two parts: the cartilaginous glottis and the membranous glottis, separating cartilaginous adduction and membranous medialization. The cartilaginous glottis is the most posterior part of the glottis, consisting of the arytenoid cartilages and their vocal processes. It is maintained by the lateral cricoarytenoid muscles, the interarytenoid muscle, and the posterior cricoarytenoids (for abducting) (Seikel et al., 2014). The type of movement these muscles exert is called cartilaginous adduction, and it is said to affect the type of phonation along the dimension of breathy-optimal-pressed (Herbst et al., 2019). The membranous glottis is the portion made up of the vocal folds from the anterior commissure to the tip of the vocal processes, and it can be adducted mainly by the thyroarytenoid muscles, which cause the vocal folds to bulge medially and reduce the width of the glottis anteriorly. This maneuver is called membranous medialization, and it is said to affect the phonation register (Herbst et al., 2019). Lindestad and Södersten (1988) also recognized the phenomenon in their videofiberscopic study of male singers, calling it the centrally located medial compression of the vocal folds.

Herbst et al. (2011) shows that singers (and non-singers) can create four distinct voice qualities by using different combinations of cartilaginous adduction and membranous medialization (Herbst, Schutte, & Švec, 2011). Both trained and untrained singers can separately influence the degree of cartilaginous adduction and

membranous medialization and create abducted and adducted falsetto qualities and abducted and adducted modal (or chest) qualities (Herbst et al., 2019, 2011; Herbst, Ternström, & Švec, 2009). In singers' terms, they would roughly translate to the head voice with a pressed quality, breathy head voice, chest voice with a pressed quality, and breathy chest voice.

2.4.4.2 Vocal registers in singing

Registers are defined as "a series of consecutive and homogenous tones going from low to high, produced by the same mechanical principle, and whose nature differs essentially from another series of tones equally consecutive and homogenous produced by another mechanical principle" (Henrich, 2006 p.3; Hollien, 1974). The two main registers in singing are the modal/chest and the falsetto/head. The main difference between them lies in the action of the thyroarytenoid muscle (also known as the vocalis muscle) (Herbst et al., 2019; Vennard & Minoru, 1970).

The modal/chest register is the most commonly used mode of phonation, as we use it in our daily speech communication. This register is called modal because it includes the range of fundamental frequencies most frequently used in everyday speaking. In singing, it resides at the low to middle portion of the whole voice range. In the modal register, the glottis is fully involved in phonation. The thyroarytenoid muscle is active and the body of the vocal folds thicken, shorten, and bulge medially. Simultaneously, the cover part of the folds slackens. (Hirano, 1974; Titze, 2000). Because the two parts of the folds are in different forms of being (slack and thick), a phase difference in the vocal fold vibration called the mucosal wave is created (Hirano, 1974; Hirano, Kakita, Kawasaki, Gould, & Lambiase, 1981). The mucosal wave created in the modal register is strong and relatively slow. The mucosal wave aids the energy transfer from the pulmonary airstream to tissue during the open phase of vocal fold vibration and prolongs the closed phase of vocal fold vibration (Herbst et al., 2019; Titze, 1988b). The minimum driving pressure of the vocal folds in modal phonation (trans-glottic pressure, i.e., pressure difference between subglottic and supraglottic air pressure) is approximately 3-5 cm H_2O subglottal pressure; if the pressure difference is lower than this, the folds will not be blown apart (Seikel et al., 2014). The modal register is often characterized by a brassy or bright voice timbre. This is thought to correlate with the longer closed phase of the vibratory cycle as it will enhance the output of high frequency energy (Herbst et al., 2019). The vocal folds tolerate some longitudinal tension in the modal register (e.g.,

voicing in the belt style), but if they are stretched too much, the mode of phonation will switch to falsetto.

Falsetto/head is the high register of phonation, rivaled only by the even higher whistle register. It is located in the middle to upper portion of the whole vocal range and it is characterized by a flutey sound that is indicative of a steeper spectral slope as compared to the modal/chest register. The vertical phase difference typical of the modal/chest register can be greatly reduced in the falsetto register. Because of the higher longitudinal tension in the vocal folds, their vibration is smaller in amplitude, but higher in frequency (Herbst et al., 2019). In falsetto, both the vocal fold body and cover are stretched, and it is less likely that the mucosal wave can be seen. The superficial vocal fold layer is under great tension, which reduces the amplitude of the oscillation at the same time as the mucosal wave speed increases (Hirano, 1974). In falsetto singing, the vocal tract settings affect vocal fold oscillation more easily (Titze, 2008a). P_{sub} for falsetto is often lower than in the modal register because the glottis can remain slightly open. Acoustically, falsetto is characterized by a high fundamental frequency, a flute-like quality due to the partials being a long way from each other, and a steep spectral slope due to the relatively longer open phase and less abrupt changes in the airflow of the oscillatory cycle. (Herbst et al., 2019; Laver, 1980; Seikel et al., 2014; Titze, 2008).

Whistle is typically characterized by a flute-like high pitched squeal. It is a phonation type that occurs mainly on the 6^{th} octave (Echternach, Döllinger, Sundberg, Traser, & Richter, 2013; Garnier, Henrich, & Crevier-Buchman, 2012; Titze, 2000). Various mechanisms have been suggested for producing whistle or flageolet phonation, i.e. turbulent sound source like in whistling or vibrating only part of the membranous vocal fold length (Echternach et al., 2013; Garnier et al., 2012). However, at least in a trained operatic singers whistle phonation is reported to be created by tightly stretched and slightly abducted vibrating vocal folds by means of airflow modulation (Echternach et al., 2017, 2013; Švec, Sundberg, & Hertegard, 2008).

Vocal fry is the low frequency quasi-periodic vibration of a small section of the vocal folds. The vibrating margin of the vocal folds is flaccid and thick, while the lateral portion of the thyroarytenoid muscle is tensed. There is a strong medial compression coupled with low P_{sub} of 2cm H_2O, which helps to sustain this mode of phonation. The glottal excitation pulses are dampened by the vocal tract and trachea, and because of this, the acoustic output goes to zero in between the pulses (Laver, 1980; Seikel et al., 2014; Sundberg, 1987; Titze, 1988, 2000).

The chest and falsetto (and even whistle) register can be produced with or without glottal closure, depending on the amount of vocal fold adduction (Herbst et al., 2019).

2.4.4.3 Vocal attack and variations of the modal voice

An increase in P_{sub} leads to an increase in the airflow if the glottal resistance is kept constant. Glottal resistance is defined as the ratio between P_{sub} and transglottal airflow, and it varies considerably between the extremities of breathy and pressed. The glottal resistance is determined mainly by the degree of adduction of the vocal folds (Sundberg, 1987). The process of bringing the folds together to start phonation is called vocal attack. This action requires muscular action. In simultaneous vocal attack, we coordinate adduction and the onset of respiration so that they occur simultaneously. The vocal folds reach the critical degree of adduction exactly at the same time the airflow from the lungs is strong enough to support phonation. In breathy vocal attack (soft attack), the airflow starts before the vocal folds adduct, and in the glottal attack (hard attack), adduction of the vocal folds happens before the airflow (like in a cough) (Seikel et al., 2014). We use all of these three ways of attack all the time in speech and in singing, and this is quite normal. The problem arises if the habitual attack style veers towards either breathy or pressed. A breathy phonation is an instance of low glottal resistance, which is normally preceded by soft attack. A pressed phonation style is an example of high glottal resistance and is usually preceded by hard attack (Sundberg, 1987). Varying vocal fold adduction along the axis from breathy to pressed induces a change from a steep to gentle spectral slope (flattening of the spectrum). Loose adduction causes a large level difference between the lowest partials (H1-H2, i.e., H1 dominates the spectrum), while the opposite is seen when the adduction is tight (pressed, strained voice). Likewise, strong adduction results in stronger relative spectral energy above 500 Hz (or 1 kHz) (Fant, 1970; Pulkki & Karjalainen, 2015; Sundberg, 1987; Titze, 2000; Titze et al., 2015) This relation between adduction and spectrum concerns particularly source spectrum, since vocal tract resonances affect the radiated spectrum, changing the amplitude relations between partials. Steepness of the waveform at the closing phase of the glottal cycle is acoustically the most important determinant of perceptual quality along the breathy-pressed continuum. A steeper closing phase will result in an acoustic output with a larger degree of high frequency energy and a flat spectral slope. When the glottal flow waveform has a more gradual

slope, with less high frequency energy and a steeper spectral slope, the voice is perceived as less brilliant or even breathy (Herbst et al., 2019).

Breathy and pressed phonation styles are variations of cartilaginous adduction: "While in breathy voice the arytenoid cartilages are set apart, in pressed voice they are usually squeezed together" (Herbst et al., 2019, p. 12). In pressed phonation, the medial compression is increased excessively. A pressed voice may sound strident or harsh, and sustaining this phonation type for long periods of time may be harmful to the voice. The greater medial compression creates a stronger, louder phonation. The pressed voice is characterized by too much muscular effort, which leads to strong impact stress in the vocal folds that can damage the tissues of the folds. A breathy voice in turn is created if the vocal folds are inadequately approximated and excessive airflow escapes between the vibrating margins of the vocal folds. Breathiness is inefficient and causes air to be wasted, but it will not damage the vocal mechanism. The breathy voice is characterized by weak muscular effort in the adductory muscles (Seikel et al., 2014).

A harsh voice is acoustically characterized by spectral noise resulting from irregularity of the glottal waveform. A prominent physiological correlate of harshness in normophonic people is excessive laryngeal tension. It can be the excessive approximation of the vocal folds, as in hard glottal attack, but strong tensions in the throat and neck area and the hypertension of the whole body can also contribute to the harshness of the voice. If harshness becomes extremely severe, the ventricular folds start to partake in phonation (Laver, 1980). Ventricular phonation is a rare type of phonation that demands large amounts of muscle activity. In singing, ventricular phonation might be used as an embellishment or vocal effect (Neubauer, Edgerton, & Herzel, 2004).

2.4.5 Resonatory system

Both the intensity and efficiency of sound production can be affected by vocal tract inertance. This is what Titze and Verdolini (2012) call the resonant voice. The resonant voice refers to the interaction effects between the vocal tract and vocal fold vibration (Titze & Verdolini Abbot, 2012).

2.4.5.1 Source-filter interaction in phonation

Interaction is a concept that refers to the backward flow of energy from the vocal tract to the source (the source refers to the glottal excitation, or, in other words, the pulsating airflow that emits through the glottis). The interaction of vocal fold movement with the acceleration and deceleration of the supraglottal air column is considered one of the mechanisms for self-sustained oscillation (Titze & Verdolini Abbot, 2012). The vocal tract air column is accelerated during glottal opening due to increased glottal flow. This produces positive supraglottal pressure. When both supra- and subglottal pressure are positive, the resulting positive intraglottal pressure pushes the vocal folds apart. This reinforces the movement of the vocal folds. When the glottis closes, the inertia (the resistance of any physical object to any changes in its velocity) creates a negative supraglottal pressure as less air is filling in behind the moving air column. The pressure in the glottis is also reduced. The reduced pressure can even create a pull on the vocal folds instead of a smaller push. The strength of this push-pull interaction depends on the vocal tract inertance. The push-pull action helps to sustain vocal fold vibration (Titze & Verdolini Abbot, 2012). If the intraglottal pressure is negative when the glottal flow increases and the folds are trying to move a part but positive when glottal flow decreases and the folds are trying to adduct, the result is contrary and hinders the vibration.

According to Titze and Verdolini-Abbott (2012) inertance can be increased by lengthening the epilarynx (the tube consisting of the laryngeal ventricle, the space between the ventricular folds, and the laryngeal vestibule) and by reducing its cross-sectional area.

Phonation threshold pressure (PTP) is the minimum lung pressure required to initiate vocal fold oscillation. The PTP is lowered by the push-pull action from the vocal tract. The greater the inertance of the supraglottal air column, the greater the push-pull effect is on the vocal folds. A low threshold pressure will allow a greater amplitude of vocal fold vibration and that in turn will increase the source strength.

2.4.5.2 Systems of loudness control

In order to increase vocal intensity, in general we increase the amplitude with which the vocal folds open and close the glottis. There are three components to SPL control in the voice: the subglottic pressure, phonation type (type of adduction), and the resonances of the vocal tract. Increased subglottal pressure has an impact on vocal loudness, and it could be considered the most important of the factors

mentioned here (Echternach, Burk, Burdumy, Traser, & Richter, 2016; Herbst, Hess, Müller, Švec, & Sundberg, 2015; Sundberg, 2017; Zhang, 2015). Medial compression of the vocal folds comes as a close second. Vocal intensity can be increased and decreased by varying the grade of adduction of the vocal folds. The so-called flow phonation[5] (the perfect balance between P_{sub} and adduction) has been shown to have the greatest maximum flow declination rate, which in turn is associated with the greatest SPL (Sundberg, 1987). If the vocal tract shape remains constant, for intensity to increase, the energy of the source must also increase. The vocal folds are in tight compression in loud phonation. Therefore, it takes more air pressure to blow them apart, the folds make a wider movement and close the glottis more rapidly, and they tend to stay closed longer because of the elastic and aerodynamic forces and tissue compression. Because so much energy is required to hold the folds in compression, the release of the folds is also markedly stronger. The powerful movement of the vocal folds creates an explosive compression of the air travelling through the glottis. The more energetic the opening of the vocal folds, the greater is the amplitude of the cycle of vibration. For every doubling of the subglottal pressure, there is an increase of 8 to 12 dB in vocal intensity (Seikel et al., 2014; Vilkman, Alku, & Vintturi, 2002). Intensity and frequency are controlled independently, but they depend upon the same basic mechanisms (tension/compression and subglottal pressure). It is possible – albeit hard – to keep in the same pitch while varying the loudness to a large extent. In the Classical singing technique, this is called "messa di voce."

Lastly, modifying the sound channel resonance characteristics can have an effect on SPL and/or perceived loudness. The epilarynx tube plays an important role in reshaping the vocal tract for volume. Matching the vocal tract input impedance with the source impedance provides the best transfer of power from the glottis to the lips, and the epilarynx tube area is the key to this matching. A wide epilarynx tube requires a low glottal impedance for maximum power transfer. A narrow epilarynx tube requires more adduction (high glottal resistance) for maximum power transfer. The singing style qualities of crooning and the sob are produced with a wide epilaryngeal and pharyngeal vocal tracts, while the belt and the twang opt for a more narrow tube and high glottal resistance (Titze & Verdolini Abbot, 2012). The raising of the larynx results in a shortening of the pharynx and the narrowing of the lower part of it as the wall tissues pile up and fill a part of the lower pharynx. When the vocal tract shortens, the formants rise. A lower larynx position stretches the pharyngeal sidewall

[5] Flow phonation is a term used in voice pedagogy and voice therapy to describe a production that feels effortless and efficient because ample airflow is passed through the glottis when the vocal folds vibrate (Titze, 2015).

tissues, resulting in a widening of the lower larynx. Lengthening the vocal tract always results in the lowering of all formant frequencies (Sundberg, 1987).

Modification of the lower vocal tract might produce a clustering of vocal tract resonances 3 to 5, resulting in boosted spectrum partials in this region (2000-3000 Hz, singer's formant). The singer's formant cluster is a quality in the singing voice that is described as resonant, projective, and sonorous. Acoustically it corresponds to the clustering of the three higher resonances of the vocal tract, f_{R3}-f_{R5}, so that instead of F3-F5 being apart in the spectrum, they form a single peak within the region between the frequency band of 2.5 to 3 kHz, preceded and followed by a minimum in the spectrum envelope. From a physiological point of view, this can be achieved by lowering and widening the pharynx, straightening the aryepiglottic region, and enlarging the sinus pyriformis (Sundberg, 1974) Sundberg has noted that the cross-sectional area in the pharynx at the level of the larynx tube (epilarynx) should be more than six times the area of that of the epilaryngeal cross-sectional area for the singer's formant to form (Sundberg, 1974).

As a rule of thumb, when formants pack closer together, the sound is perceived to be louder; when they scatter, the sound is perceived to be softer (Sundberg, 1987). Raising the first resonance frequency of the vocal tract increases the SPL as F1 comes closer to F2 (Fant, 1970). Increase in F1 also strengthens the singer's formant by several dB and rises the center of spectral gravity (Vurma, 2020). The first formant makes a big contribution to the overall sound spectrum envelope at lower frequencies (Sundberg, 1987). The first vocal tract resonance (f_{R1}) is raised by a greater lip and jaw opening. This is a strategy that singers capitalize on in very high pitches (>500Hz) where the natural placement of the first formant falls below the fundamental frequency. By adjusting the jaw and lips, the singers lift the first formant to the level of the fundamental frequency and in this way can gain more power for their sound (Sundberg, 1987). Raising the formant frequencies (e.g., in smiling or using a more frontal articulation) makes the voice timbre brighter, which increases perceived loudness as it increases the sound energy in the higher frequency range (between 2 and 4 kHz) where the human ear is more sensitive. Lowering the formant frequencies (like in yawning or protruding the lips or vocalizing with a retracted tongue or small mouth opening) darkens the timbre. Adduction and vocal tract acoustics interact, e.g., so that when the adduction is loose, it increases damping, lowers the amplitude of the formants, and broadens their bandwidths.

Practicing jaw opening, lip opening, and increasing pharyngeal volume while keeping the laryngeal position low could have an effect on loudness. Echternach et al. (2016) found articulatory differences with respect to changes of both pitch and

loudness. They investigated Classical singers with different voice classifications using magnetic resonance imaging (MRI). The singers were asked to sing at different loudness levels and the success of their efforts were corroborated with the acoustic measurement of SPL. Lip opening and pharynx width were increased and correlated more with the sound pressure level than pitch. Vertical larynx position was found to rise with pitch and lower with greater loudness. Changes in jaw protrusion, uvula elevation, and tongue position did not yield significant results in the SPL (Echternach et al., 2016).

2.4.6 Articulatory system

The articulatory system is the system of immobile and mobile organs brought into contact for the purpose of shaping the sound of the singing voice. There are three immobile articulation points: the alveolar ridge of the upper jaw, the hard palate, and the teeth. Mobile articulators are the tongue, the lower jaw, the soft palate, the lips, the cheeks, and the region behind the oral cavity (fauces and the pharynx), which can be moved through muscular action. The muscles involved in the shaping of the resonant cavities of the vocal tract are responsible for vowel and consonant pronunciation, and they play an important part in molding the perceived voice quality (Seikel et al., 2014; Sundberg, 1987). According to Laver (1980), the neutral configuration of the vocal tract can be modified by three different groups of settings: 1) by the modification of the longitudinal axis of the tract, 2) by the modification of the latitudinal, cross-sectional axis of the tract, and 3) by velopharyngeal modifications. The lips, tongue, and velum are the three most significant structures for articulation. Movement of the lips is produced by the muscles of the face. The tongue capitalizes on its own musculature and the muscles of mandible and hyoid for its movement. The muscles of the velum elevate that structure to seal off the oral region from the nasal cavities.

2.4.6.1 The lips

The longitudinal axis is modified by the laryngeal position (high/neutral/low) and the lip opening. Lip rounding, and especially lip protrusion, has a similar acoustic (and also perceptual) effect on the voice as lowering the larynx. It lowers the frequencies of the lower formants and brings about a darker timbre in the voice. The effect is not identical though, as in the lowered larynx voice, the lower formants (F1-

F2) are most affected, while in labial protrusion and lip-rounding, it is the higher formants which are most changed (Laver, 1980).

In latitudinal settings, the lips' function is to create a particular constrictive or expansive effect on the cross-sectional area relative to the cross-sectional area habitual to the neutral vocal tract. This effect is created in the interlabial space, which is defined as "the maximum horizontal and vertical dimensions of the aperture through which the airstream can pass" (Laver, 1980, p. 35). Laver (1980) identifies a total of 18 categories of labial settings available for creating different voice qualities. We already established that there is a protruded setting in use in the longitudinal modification of the vocal tract, so naturally there needs to be a non-protruded version to counterbalance that. Then we have eight latitudinal settings pertaining to the interlabial space: 1) horizontal expansion, 2) vertical expansion, 3) horizontal constriction, 4) vertical constriction, 5) horizontal expansion with vertical expansion, 6) horizontal constriction with vertical constriction, 7) horizontal expansion with vertical constriction, and 8) horizontal constriction with vertical expansion. Finally, we have the neutral setting (with or without protrusion).

The lips form the focus of the facial muscles. Numerous muscles insert to the orbicularis oris inferior and superior muscles, providing a system for lip closure, protrusion, retraction, elevation, and depression. These muscles are responsible for the bulk of facial expression, and they are important for articulation. The buccinator and the risorius insert into the corners of the mouth and retract the lips. The zygomatic major elevates and retracts the angle of the mouth, as in smiling. The orbicularis oris are continuous with the buccinator muscle, and ultimately with the superior pharyngeal constrictor. The depressor labii inferioris depresses the lower lip and the depressor anguli oris depresses the corners of the mouth. Levator labii superioris, zygomatic minor, and levator labii superioris alaeque nasi muscles elevate the upper lip. The levator anguli oris pulls the corner of the mouth up and medially (Seikel et al., 2014).

While protrusion makes the perceived voice quality darker, the acoustic effect of horizontally expanding the interlabial space is chiefly to raise the formant frequencies and make the timbre brighter. The latitudinal rounding of the lips has a similar effect as longitudinal protrusion – it tends to lower the formant frequencies. The different shapes of the lip-openings are created by the actions of different muscles or by different degrees of tension in the same muscle. The acoustic properties that the muscles of the lips exert on the voice can therefore vary with different tension settings and give slightly different auditory impressions to the voice. The different settings affect the radiation and absorption of sound through the cavity walls, which

is then reflected in the acoustic signal as changes in the bandwidths of the formants most affected (Laver, 1980).

2.4.6.2 The tongue

The muscles of the mouth are dominated by the intrinsic and extrinsic muscles of the tongue and muscles elevating the soft palate. The superior surface of the tongue is called the dorsum, the anterior portion is the tip, and the portion of the tongue that resides in the oropharynx is called the base. The finer movements of the tongue are produced by the contraction of the intrinsic musculature. The superior longitudinal muscles move the tongue tip up, while the inferior longitudinal muscle pulls the tongue tip down. If these muscles contract together, they help with the retraction of the tongue. The transverse muscles of the tongue can narrow the tongue, and the vertical muscles flatten the tongue. The extrinsic muscles tend to move the tongue as a unit. The genioglossus is the prime mover of the tongue, and it moves the tongue in an anterior-posterior direction: contraction of the anterior fibers retracts the tongue, and constriction of the posterior fibers draws the tongue forward. The hyoglossus pulls the sides of the tongue down in direct antagonism to the palatoglossus, which elevates the back of the tongue. The palatoglossus has a double role as the soft palate depressor, and the muscle makes up the anterior faucial pilar. The styloglossus draws the tongue back and up, and the chondroglossus depresses the tongue (Seikel et al., 2014).

The various constrictive settings of the tongue result from the movement of the location of the tongue's center of mass. Phonetics distinguishes nine different tongue settings, but for our purpose two suffice: the tongue-fronted voice and the tongue-retracted voice (Laver, 1980). Raising the tongue and giving a slight fronting to the mass of the tongue (e.g., in the vowel /i/), gives a brighter timbre, raising the formant frequencies. In this setting, the second formant is maximally high and close to F3. The distance between the first and the second formant is usually large. Retracting the tongue, moving the center of the tongue's mass backwards and slightly downwards (e.g., in the vowel /u/), gives a darker timbre, lowering the formant frequencies. In this type of setting, F1 is usually a little higher than in a neutral setting, and the second formant typically lower than in neutral (Laver, 1980). Settings where the mass of the tongue moves downward and forward as a continual tendency are rare, but in voice pedagogics, this is a setting that is often promoted. For example, exercises where the tongue is placed on top of the lower lip while uttering an /a/ or an /ä/ sound can be used for training a larger pharynx and oral cavity space in

singing. Lindblom and Sundberg (1972) found in their study investigating the tongue contour lengths in the production of sung and spoken vowels that spoken vowels would have shorter tongue lengths than their sung counterparts. The root of the tongue can be adjusted to expand or constrict the pharynx. Advancement of the root of the tongue results in a more frontal sound (Edmondson & Esling, 2006). The tongue body settings underline basically every segment of continuous singing. The tongue tip/blade has less effect on the sound quality even though it is important in articulation.

2.4.6.3 The jaw

The jaw has four types of movements: vertical, horizontal, lateral, and rotational. In other words, the jaw can open and close the mouth, protrude and retract, move a little to the left and to the right, and rotate a little. The principal voice-related dimension is the vertical dimension of jaw open – jaw closed. The neutral mandibular setting interferes at least with the tongue and the lips. The neutral jaw position is achieved by lifting the jaw, which means that there is some muscular tension present in the neutral position. The neutral position lies somewhere between the two extreme possible settings of maximally open or clenched shut (Laver, 1980).

There are three paired muscles that raise the jaw and three that close it. The masseter is the most powerful jaw muscle; it lifts the jaw, and it is the muscle responsible for the action of grinding the teeth together. The internal pterygoid muscle lifts, protrudes, and pulls the jaw to one side. The temporalis can be used for lifting and retracting the jaw. Lowering the jaw happens with the co-operation of these aforementioned muscles. They need to give way to the muscles that actively open the jaw: the external pterygoid, geniohyoid, the anterior belly of digastricus, and the mylohyoid (Laver, 1980; Seikel et al., 2014).

The acoustic effect of jaw movement can be seen in the first formant. F1 rises as the jaw opening becomes larger. In the closed jaw setting, the frequency of the first formant tends to drop and its range decreases. Higher formants are proportionally less affected by the jaw opening, but they too tend to rise with the degree of jaw opening (Lindblom & Sundberg, 1971). The relationship between jaw and lip positions is such that each can magnify or diminish the other's effect. That is why it is important to check the lip effect before specifying acoustic phenomena to any particular mandibular setting (Laver, 1980).

2.4.6.4 The velum

The velum is an important part of the vocal tract for singing. Much emphasis is given to the correct form of the velopharyngeal area in voice pedagogy and there are heated discussions around the matter of how much is too much velopharyngeal activity. The perceptual marker of inadequate velopharyngeal activity is a nasal voice. A lot of the misconceptions about nasality come from the oversimplistic view of the velopharyngeal action as involving only the positioning of the velum as a hinge-like trapdoor that just lifts to seal of the passage to the nasal cavities (Birch, Gümoes, Prytz, et al., 2002; Birch, Gümoes, Stavad, et al., 2002; Gill, Lee, Lã, & Sundberg, 2020; Laver, 1980; Sundberg et al., 2007). In reality, there is some "anticipatory" nasality present in nearly every spoken or sung segment, as nasal and non-nasal speech sounds alternate rapidly in normal communication.

"Action of the velopharyngeal system is a mixture of the valvular movement on the part of the soft palate and sphincter movement by the superior constrictor and its related fibers" (Laver, 1980, p. 77). The paired muscles of the soft palate (or velum) include elevators, a tube dilator, and depressors. The levator veli palatini is the palatal elevator making up the bulk of the soft palate. The musculus uvulae shortens the soft palate, bunching it upwards. The tensor veli palatini functions as a dilator of the auditory tube. The palatopharyngeus assists in narrowing the pharyngeal cavity as well as lowering the soft palate. It can also help to elevate the larynx (Seikel et al., 2014). Contracting the tensors spreads and tenses the soft palate laterally. Contraction of the levator lifts the body of the velum, which has been tensed by the tensor (Laver, 1980; Seikel et al., 2014).

In singing, the elevated soft palate is usually preferred, although recent studies have indicated that a slight velopharyngeal opening might contribute to the relative enhancement of the higher spectrum partials (Birch, Gümoes, Prytz, et al., 2002; Gill et al., 2020; Sundberg et al., 2007; Vampola, Horácek, & Laukkanen, 2021) and possibly help facilitate a seamless timbral transition in a tenor passaggio area (Birch, Gümoes, Stavad, et al., 2002).

Nasality sets in when the opening to the nasal cavity is relatively larger than the opening to the oral cavity. Nasal resonance is produced by a side chamber (usually the nasal cavity) that has an equal size or larger entry area and a smaller exit than the other cavity (Laver, 1980). The acoustic effects of a nasal vowel quality have been found to result from both widening the bandwidth of the first vowel formant and from a decrease in its amplitude. Also a spectrum peak near 250 Hz is often observed in nasalized vowels (Gill et al., 2020). Hixon (1949) found that nasal speakers retract

and raise their tongues more compared to normal speakers. It was thought to be because the palatoglossus in pulling the velum downward would simultaneously pull the tongue body upwards and backwards (Hixon, 1949, as sited in Laver, 1980). The marked auditory and acoustic effect of nasality is the loss of power in the lowest formant frequencies. Nasalization has been found to emphasize the frequency range around the singer's formant cluster (F3-F5) (Birch, Gümoes, Prytz, et al., 2002; Gill et al., 2020; Sundberg et al., 2007; Vampola, Horacek, Radolf, Švec, & Laukkanen, 2020).

2.4.6.5 Twang

Twang is not a nasal quality in essence, although it too can be nasalized. It was classified as a (singing) voice-quality category that is separate from, for example, opera, belting, and sob, by Yanagisawa and Estill in 1989, and it has since consolidated its position in the terminology used by singing voice pedagogues (Yanagisawa et al., 1989). The narrowing of the aryepiglottic sphincter, pharyngeal constriction, shortening of the overall vocal tract length, and retraction of the corners of the lips are all associated with twang quality (Sundberg & Thalén, 2010; Titze, Bergan, Hunter, & Story, 2003; Yanagisawa et al., 1989). The closed quotient is found to be greater in twang, whereas the pulse amplitude and the fundamental have been found to be weaker and the normalized amplitude lower as compared to a neutral (singing) voice use (Sundberg & Thalén, 2010). Formants F1 and F2 are higher in twang (as compared to normal) and formants F3 and F5 are lower than in the normal singing voice. The lowering of the F3 can be seen as indicative of twang being produced with a front cavity between the tongue tip and the lower incisors. This setting can be coupled with a retraction of the tongue, which is needed for narrowing the pharynx (Sundberg & Thalén, 2010). Twang quality can be heard as a type of brightening of the vowel sounds in the voice (Titze & Verdolini Abbot, 2012).

2.5 Teaching emotion expression through voice qualities

Singing voice pedagogy usually aims at a healthy voice production, stamina, and projection within the vocal demands of a desired musical style. Voice quality is often coupled with a resonant voice. A resonant voice is a voice quality that makes good use of the auditory mechanism and uses subjective bodily functions to perform

consistently (Alderson, 1979; Bunch, 1997; Chapman, 2006; Harrison, 2006; Miller, 1996; Sell, 2005). One of the key characteristics of a resonant voice in the voice pedagogic literature is that perceptually it sounds good. It is often described as free, flowing, projecting, and pleasing to the ear. A scientific description of the resonant voice is any voice production in which the vocal tract assists the sound source in producing acoustic energy (Titze & Verdolini Abbot, 2012). Clinically, the resonant voice can be defined as a type of voice production that is both easy to produce and vibrant in the facial tissues. It has been described as having low-impact vocal fold collision and a large-amplitude vocal fold oscillation, providing both preventive and healing effects. Perceptually, a resonant voice is described as neither pressed nor breathy (Li et al., 2008; Peterson, Verdolini-Marston, Barkmeier, & Hoffman, 1994; Titze & Verdolini Abbot, 2012; Verdolini, Druker, Palmer, & Samawi, 1998). Teaching acoustic emotion expression through different voice qualities means taking a step back from the traditional way of teaching a solid resonant quality that can persevere through the entire vocal range in all dynamic conditions. The effect an emotion has on voice production is that it changes it, and if we want to mimic this biological tendency, we must allow for a little staggering in the voice quality.

The physical and acoustic elements of teaching emotion expression through changing voice qualities are thankfully exactly the same as for teaching the regular singing technique. The only difficulty in teaching emotional voice qualities is to know when to stop pushing the boundaries. There are two things to consider: 1) Does the emotional voice quality comply with the aesthetic demands of the music style? 2) Is the emotional voice quality safe to perform?

From the teacher's point of view, it is usually as simple as asking for a little more or less of something in a specific area of voice production. The gist of the parameter modulation model proposed in this study is that it gives one a direction from which to start looking for exercises for teaching expressive emotive vocalizations. It cannot give direct, ready-made exercises as the proper amount of voice quality change has to be evaluated for each occasion separately (according to musical demands and the singers' individual anatomy and physiology). Fortunately, all of the regular vocal technique exercises we use in our normal teaching can be harnessed for teaching emotion expression through changing voice qualities following the model presented in Section 4.4.

2.5.1 Perceptual motor learning

Music has been taught for ages through the master-apprentice teaching-learning method, where young people wanting to learn how to play an instrument or tap into a musical tradition have sought the instruction of someone who has already mastered the said instrument and/or tradition. Singing is no exception. Learning to sing is a form of perceptual motor learning. Motor learning is defined as a set of processes associated with practice or experience leading to relatively permanent changes in the capability for movement (Guadagnoll & Lee, 2004; Maas et al., 2008; Titze & Verdolini Abbot, 2012). Understanding how the motor system reorganizes itself based on behavioral intervention can provide important insights into the way we teach singing. Motor skill learning is facilitated by a number of factors concerning the structure of practice, stimulus selection, and the nature of feedback (Maas et al., 2008). Mastering different singing strategies involves relatively stable changes in average performance over time. Learning indicates that the same thought processes and resulting configuration of the physiological voice (settings in the voice organ) occur in new situations based on success in previous situations that were somewhat different, but related (Titze & Verdolini Abbot, 2012). A permanent change in the capability to perform different singing techniques (skilled movement) has to be measured by retention/transfer tests. Retention refers to the skill level that the student is able to maintain after the practice session has ended. Transfer refers to the generalization of knowledge/skills (e.g., the student can use the same vocal technique in different songs). The transfer and retention measures are important in voice research concerning learning, as performance during practice could be affected by factors – such as warm-up, fatigue, or lack of attention – that do not necessarily reflect learning (Maas et al., 2008).

2.5.1.1 Schema theory

A possible framework for a theoretical approach to the teaching of singing is the Schema theory (Schmidt, 1975, 2003). The theory states that the production of rapid discrete movements involves units of action (motor programs) that are retrieved from memory and then adapted to a particular situation. It further postulates that the motor programs are generalized so that they catch the essence of a certain movement, but fine tuning needs to be done to determine the absolute timing and force of the muscle contractions. A general class of movements may be governed by a single generalized motor program, which can be scaled to meet the current task

demands (Schmidt, 1975). An example could be that a pitch is created with a sufficient P_{sub} and adduction in conjunction with the right vocal fold length (essence of movement), but the way one executes it in a song is dependent on the tempo and overall loudness of the accompaniment (fine tuning).

In order to work correctly, the motor system must be familiar with the relations among the initial conditions (current position of vocal tract and breath support system), the generated motor commands (timing and amplitude of the trunk and vocal tract muscle contractions), the sensory consequences of these motor commands (proprioception of movement, auditory feedback loop), and the outcome of the movement (did the singer hit the correct pitch). This knowledge is referred to schemas in schema theory, and they mean the memory representations that encode the relations among these types of information, based on past experience with producing a similar action. When performing a specific action, these memory representations become temporarily available in the short-term memory, and they can be used for two different types of schema creation: 1) recall schema and 2) recognition schema. To produce a certain movement, the system supplies the recall schema with the movement goal (intended outcome) and information about the current conditions, from which the recall schema computes the appropriate parameters. The recognition schema allows the system to evaluate movement by comparing the actual sensory consequences with the expected sensory consequences of a correct movement. If there is a mismatch between the actual and the expected consequences, the system will generate an error signal that can be used to update the recall schema (Maas et al., 2008). Learning a specific singing technique usually requires some feedback from an instructor, especially at the beginning of the learning. The reference of correctness is not directly available or interpretable to the learner, or, to put it differently, the singer cannot yet hear the optimal vocal organ settings through acoustic information. In these cases, the singer calibrates the expected sensory consequences with an externally provided reference (teacher instruction) of correctness and receives the error signal in this way.

2.5.1.2 Internal/external focus

Instructions are a central part in teaching motor skills. When we are trying to teach a particular sound, such as twang, we usually give quite detailed instructions as to how to achieve it. We might ask our student to lift the larynx, tilt the epiglottis, retract the corners of the mouth, lift the soft palate, show teeth, etc. Previous research has suggested that conscious attention to the mechanics of a motor task can affect

learning and performance negatively. Attention to task outcomes, on the other hand, can benefit performance and learning (Titze & Verdolini Abbot, 2012; Wulf, Lauterbach, & Toole, 1999). An internal focus refers to concentrating on the kinetic, kinematic, and somatosensory aspects of movement (e.g., the force of adduction or the tongue position). An external focus means concentrating on external but task-related aspects of movement (e.g., what we hear). The goal of movement is the same in both the external and internal focus conditions – to produce a certain kind of voiced sound (e.g., twang) and the feedback in both conditions is also the same (the produced sound itself) (Maas et al., 2008). Wulf et al. (2001) postulated that the internal focus in the sense of trying to consciously control otherwise automatic motor processes might cause the system to "freeze." They think that aiming the focus away from the motor system's actual function would allow for more automatically executed motor routines. They also showed that participants' reaction times to a secondary task were faster when performing a primary task with an external focus (Wulf & Prinz, 2001). This might be of significance when thinking about making music together. Focusing on listening rather than physical sensations might allow for a prompter reaction to other musicians' ideas during playing.

Attention is nevertheless required before automaticity is achieved. Instead of detailed anatomical and physiological instructions that turn the focus of attention inward, it might be sometimes useful to direct the attention to an external focus by, for example, asking the student to make the sound of a wicked witch when practicing the twang sound. Auditory feedback affects the performance of singers, and the integration of auditory feedback is an important part of finding the ideal sound. Furthermore, auditory feedback contributes significantly to singers' pitch control (Mürbe, Pabst, Hoffmann, & Sundberg, 2002). The auditory feedback system is a form of automonitoring through inputs from cochlear mechanoreceptors in response to their vibratory simulation by vocal tract sound waves (Laukkanen, 1995). This feedback system offers a fast corrective tool that often works better than verbal instructions about the biomechanics of a voice in a specific task.

As the human voice is generated internally, vocalists rely also on proprioceptive feedback, such as vibratory (tactile) and kinesthetic (muscular) sensations. These sensations help singers to control their singing technique. As proprioceptive control develops through practice, it replaces external auditory feedback as a more reliable source of technique-related input. The perceived voice is always modified by the acoustics of the environment, and non-experienced singers compensate more when it comes to different acoustic environments (Bottalico, Graetzer, & Hunter, 2016). The route to body awareness for most singers goes through listening to the sound

(and feedback from the teacher) and connecting it with a certain "setting" of the voice apparatus.

2.5.1.3 Task explanation and motivation

Both the emotional and cognitive sides are important in teaching, and a lot of what makes a good teacher eventually comes down to personal relationships – the way a teacher can empathize with the student and to what extent the instructor can communicate effectively (Orón Semper & Blasco, 2018). Using pedagogical tools such as constructive alignment (Biggs, 1996; Biggs & Tang, 2011) can ameliorate the quality of teaching in a situation where the communicative process is compromised. The Feather Newton value expectation theory is a good guiding principle in pupil-oriented teaching. It states that in order for students to engage with their studies and in order for deep learning to occur, they must gain something useful (tangible) from the education, and the students must be able to perform the tasks they are given (Feather & Newton, 1982). Motivation can be enhanced by understanding the relevance of the rehearsed task (Maas et al., 2008). This speaks for the importance of explaining the acoustic and physiological linkage in the human voice. So, we should not just dispose of theory, even though it might slow down the learning process due to the increased processing capacity required for understanding the terminology and concepts. We should, however, include the student in the selection of functionally relevant learning targets, as this has been found to increase motivation in previous studies (concerning the integral simulation method) (Strand & Debertine, 2000). When we teach what they want to learn, it should motivate students to better tolerate theoretical jargon. Grasping the concepts and understanding the task are important for learning, but we as instructors can ration the semantics. Instead of overly lengthy and complex instructions, we can sometimes just model the target behavior by singing. Students can then see and hear what the expected result should be like.

Bloom (1979) has categorized the way we learn in his famous taxonomy. He states that we need to first gather up, memorize, and understand a bank of information before we can start applying and analyzing it, coming up with critiques and assessments, and creating something of our own (Bloom, 1979). In order to learn, we need information from which to learn. Typical forms of information in the voice studio include auditory, visual, and conceptual information. The student learns through observation and mimicry, operant conditioning (facilitated by the continuous feedback provided by the teacher), constructive memory, and conceptual

thinking (Guadagnoli & Lee, 2004; Nolen-Hoeksema et al., 2009; Woody, Sloboda, & Lehmann, 2007). It is obvious that thought processes play a significant part in all learning, but not all learning is attentional. (An example of unattentional learning in the voice studio would be picking up the teachers' vocal mannerisms by default.) When we look at the concepts of declarative (facts and events) vs. procedural memory (knowing how to do things) and learning through these different pathways, we come to see the potential problems in the voice studio. Declarative memory is the type of knowledge we use in exam situations: we declare that we know the answer to the question either because we read it somewhere beforehand, heard it in a lecture, or have an empirical recollection of it. Procedural information, on the other hand, is something we learned while doing it, such as tying one's shoelaces. Procedural knowledge is a lot easier to show than to verbalize. This is also the case for singing. We learn to sing by singing, which in itself is a procedural technique, but in the singing studio we verbalize the auditory and physiological events that happen and thus bring the learning experience to the realm of declarative learning. Studies on music teaching and learning have found that especially new teachers tend to spend a lot of time explaining tasks to students, while the more experienced teachers tend to give short and concise instructions with the majority of the lesson time devoted to music making (Woody et al., 2007). If there is too much information available during the learning task, it will slow down the learning process. Thus, the learning achievement depends on an optimal amount of information, which differs as a function of the skill level of the individual (Guadagnoll & Lee, 2004). This means that different types of information and different types of exercises are required for singers at different skill levels. To ensure that learners understand the task, they should also be provided with information about which productions are acceptable and which ones are not, and also why this is so (Maas et al., 2008). One way to facilitate this is to show exaggerated examples of aesthetically unacceptable voice use.

2.5.2 Joy, Tenderness, Sadness and Anger as voice qualities

In this thesis, I study four emotions (anger, sadness, tenderness, and joy). These emotions have been selected because they can be placed on a fourfold table of valence and activation. Anger, sadness, and joy are regarded as basic emotions and should by definition be easy to recognize (Ekman, 1992; Izard, 1992). Tenderness is included because an emotion with a positive valence and low activity level was

needed to complete the fourfold table. All of these emotions occur frequently in song interpretations in both the Classical and contemporary commercial worlds, and they are thus familiar performance tasks for most singers.

2.5.2.1 Neutral

The neutral voice in this study refers to the habitual way singers use their voice. Its purpose is to provide a standard reference sound to which the emotional sounds can be compared.

2.5.2.2 Joy

Typical musical expressions of joy are characterized by a fast tempo, low tempo variability, major mode, simple and consonant harmony, medium to high sound level, small sound level variability, high pitch, great pitch variability, wide pitch range, ascending pitch, perfect 4th and 5th intervals, rising micro intonation, strengthened singer's formant, staccato articulation, large articulation variability, smooth and fluent rhythm, bright timbre, fast tone attacks, small timing variability, sharp contrasts between "long" and "short" notes, medium-fast vibrato rate, medium vibrato extent, and micro-structural regularity (Juslin & Laukka, 2004).

For the singing voice, the composer has pre-determined the pitch, pitch variability, and at least a part of the articulation and tone attacks due to the rhythm of the melody. The overall instrumentation dictates the loudness level to some extent, and the music style usually has standards for the emergence of the singer's formant and the acceptable vibrato rate and extent.

The list of voice quality features found in the expressions of sung joy in the field of voice science include high loudness, low loudness variation, moderate loudness rise and fall slopes (Scherer, Sundberg, et al., 2017), high equivalent sound level, low Hammarberg index,[6] low-level difference between partials 1 and 2 (H1/H2) (Sundberg et al., 2021), higher mean sound level, more short-term variability of sound level (Sundberg et al., 1995), high SPL (Hakanpää et al., 2021a), low formant bandwidth, low formant amplitude, high formant frequency positioning, moderate

[6] The Hammarberg index is defined as the intensity difference between the maximum intensity in a lower frequency band [0-2000 Hz] versus a higher frequency band [2000-5000 Hz] (Hammarberg, Fritzell, Gaufin, Sundberg, & Wedin, 1980).

low energy frequency variation (Scherer, Sundberg, et al., 2017), low proportion energy <.5 kHz, Low proportion energy <1 kHz, high alpha ratio, high spectral flatness, high spectral centroid (Hakanpää et al., 2021a; Sundberg et al., 2021), shallow spectral slope (Jansens et al., 1997), low perturbation variation, high perturbation level (Scherer, Sundberg, et al., 2017), less jitter, and more shimmer (Hakanpää et al., 2021a).

From the viewpoint of the singing instructor, this information would point at least towards some sort of manipulation of loudness/SPL, formant positioning, and possibly vibrato. We established before that the loudness control includes an increase in P_{sub} and medial compression. In pedagogical terms, (breath) support and phonation balance. The high formant positioning, low formant bandwidths, and the way that the energy is distributed on the spectrum suggest a louder, brighter sound that could be established using twang or a singing style with rather more than less adduction in phonation and a tract setting that allows for the singer's formant or some other kind of sound energy concentration (whose perceptual correlate would be, e.g., ring) to emerge. The brightness element could be attained by raising formant frequencies, by fronting the root of the tongue, or retracting the corners of the lips (Hakanpää, Waaramaa, & Laukkanen, 2021b).

2.5.2.3 Tenderness

Typical musical expressions of tenderness include a slow tempo, major mode, consonance, medium to low sound level, small sound level variability, low pitch, fairly narrow pitch range, lowered singer's formant, legato articulation, small articulation variability, slow tone attacks, moderate timing variability, accents on tonally stable notes, medium fast vibrato, small vibrato extent, and micro-structural regularity (Juslin & Laukka, 2004). Again, some of these parameters in pre-composed music affect the singer's freedom of expression.

The list of voice quality features found in the sung expressions of tenderness in the field of voice science include low loudness, high loudness variation and rise and fall slopes, low dynamics (Scherer, Sundberg, et al., 2017), low equivalent sound level, high Hammarberg index, high level difference between partials 1 and 2 (H1/H2) (Sundberg et al., 2021), low mean sound level, less short-term variability of sound level (Sundberg et al., 1995), low SPL (Hakanpää et al., 2021a), low vocal energy, broad formant bandwidths, moderate formant amplitude, lower formant frequencies, a tendency for high low-frequency energy, small low-energy frequency variation (Scherer, Sundberg, et al., 2017), high proportion energy <.5 kHz, high

proportion energy <1 kHz, low alpha ratio, low spectral flatness, low spectral centroid (Sundberg et al., 2021), high perturbation variation, low perturbation level, little waveform irregularity (Scherer, Sundberg, et al., 2017; Klaus R. Scherer et al., 2015), more jitter, and more shimmer (Hakanpää et al., 2021a).

From a pedagogical point of view, this kind of compilation of acoustic parameters could be obtained by loosening the medial compression of the vocal folds and lessening the subglottal pressure to reduce the loudness. The broad formant bandwidths suggest softer phonation and softer muscular tissues of the oral cavity, which could be done by lengthening rather than shortening the vocal tract and relaxing some of the lip muscles so that the cheeks become a little softer (Hakanpää et al., 2021b).

2.5.2.4 Sadness

Typical musical expressions of sadness include a slow tempo, minor mode, dissonance, low sound level, moderate sound level variability, low pitch, narrow pitch range, descending pitch, "flat" or falling intonation, small intervals, legato articulation, small articulation variability, dull timbre, slow tone attacks, large timing variability, soft contrasts between "long" and "short" notes, pauses, slow vibrato, small vibrato extent, ritardando, and micro-structural irregularity (Juslin & Laukka, 2004).

The list of voice quality features found in the expressions of sung sadness in the field of voice science include low loudness, high loudness variation, moderate rise and fall slopes (Scherer, Sundberg, et al., 2017), low equivalent sound level, high Hammarberg index, high-level difference between partials 1 and 2 (H1/H2) (Sundberg et al., 2021), low mean sound level, less short-term variability of sound level (Sundberg et al., 1995), low SPL, low level of dynamics (Jansens et al., 1997), low intensity, high formant bandwidth, low formant amplitude, low formant frequencies, small low-energy frequency variation (Scherer, Sundberg, et al., 2017), high proportion energy <.5 kHz, high proportion energy <1 kHz, low alpha ratio, low spectral flatness, low spectral slope, low spectral centroid (Sundberg et al., 2021), broad formant bandwidth (Hakanpää et al., 2021a), high perturbation variation, low perturbation level (Eyben et al., 2015; Scherer, Sundberg, et al., 2017), more jitter, and more shimmer (Hakanpää et al., 2021a).

Sadness has the same activity level as tenderness, and they share almost all of the acoustic parameter directions with each other. Tenderness is described as having a tendency for high low-frequency energy and sadness is described as having a low

level of dynamics. From a teachers' point of view, there seems to be nothing that especially separates the acoustic expression of sadness from tenderness. In this type of situation, one must look at the portions of parameters to achieve the desired effect. One can try experimenting with formant positioning, making the timbre darker, one can reduce the mandibular movement or lip activity to make the sound muffled, or one can add extra vibrato to create an unstable sound (Hakanpää et al., 2021b).

2.5.2.5 Anger

Typical musical expressions of anger include a fast tempo, large tempo variability, minor mode, dissonance, atonality, high sound level, small loudness variability, high pitch, small pitch variability, ascending pitch, major 7th and augmented 4th intervals, strengthened singer's formant, staccato articulation, moderate articulation variability, complex rhythm, sudden rhythmic changes, sharp timbre, spectral noise, fast tone attacks, decays, small timing variability, accents on tonally unstable notes, sharp contrasts between "short" and "long" notes, accelerando, medium-fast vibrato rate, large vibrato extent, and micro-structural irregularity (Juslin & Laukka, 2004).

The list of voice quality features found in the expressions of sung anger in the field of voice science include high loudness, low loudness variation, rise and fall slopes (Scherer, Sundberg, et al., 2017), high equivalent sound level, low Hammarberg index, low-level difference between partials 1 and 2 (H1/H2) (Sundberg et al., 2021), high mean sound level, more short-term variability of sound level (Sundberg et al., 1995), high SPL, high vocal energy, high dynamics (rate, F0 contour, loudness variation), low formant bandwidth, moderate formant frequency, high low-energy frequency variation (Jansens et al., 1997; Scherer, Sundberg, et al., 2017; Scherer et al., 2015), low proportion energy <.5 kHz, low proportion energy <1 kHz, high alpha ratio, high spectral flatness, high spectral slope, high spectral centroid, narrow bandwidth, weak low frequency energy (Sundberg et al., 2021), flat highly balanced spectrum indicating strong energy in the higher partials, low perturbation variation, high perturbation level, less jitter, and more shimmer (Eyben et al., 2015; Hakanpää et al., 2021a; Scherer, Sundberg, et al., 2017).

From a teachers' perspective, anger is produced with a high P_{sub} and high medial compression, even a pressed voice production. Articulation is sharp, which calls for the active and fast movement of the articulators. The flat spectrum and narrow bandwidths indicate increased adduction and short open quotient. The overall shortening of the vocal tract by lifting the larynx could help, as could increasing the

position of the F1 through opening the jaw and increasing the interlabial space (Hakanpää et al., 2021b).

3 STUDY QUESTIONS

The main aim of this study was to investigate whether it is possible to teach emotion expression in singing by using vocological information about voice quality and implementing this information to form a training pattern (the parameter modulation technique).

We investigated this by proposing several smaller study questions:

- Are listeners able to recognize emotions in a singing voice from short vowel samples?

- Is there a difference in the recognition of emotions when they are sung using a Classical singing technique compared to when they are sung using a CCM style of singing?

- Does pitch affect the recognition of emotion in the Classical-/CCM-style singing voice samples?

- Are valence and the activation of the emotions perceptible in the sung samples?

- Are there acoustic voice quality differences between emotional expressions for short singing excerpts of the vowel [:a] from Classical and CCM singers?

- Does the specific training (parameter modulation technique) improve the recognition of emotions from the singing voice?

- Do the acoustic differences between emotional expressions increase after the particular training (parameter modulation technique)?

4 METHODS

This study is an experimental comparative study using the hypothetico-deductive method as its basic scientific approach (we are evaluating general explanations of observed regularities by generating and testing hypotheses).

Experimental research operates using dependent and independent variables. The independent variable is manipulated and the effect that this change has on a dependent variable is examined. This enables a researcher to identify a cause and effect between variables.

The independent variable is the predictor variable being manipulated by the experimenter in order to observe the effect on a dependent variable (Coolican, 2009). In this study, the independent variable is the voice sample performed with five different emotion expressions. To be specific, vocal emotion expression is the actual independent variable we are manipulating (by asking the performer to sing in different expressions), and the voice sample is the way we have operationalized it in this study (the way we capture the expression). In other words, we are looking to hear/see the differences in emotion expression from the voice samples. As there are four different emotions that we have asked the singers to portray in their voice in addition to the neutral state, we have a categorical independent variable consisting of five levels (joy, tenderness, neutral, sadness, anger).

The dependent variable is the event expected to change when the independent variable is manipulated (Coolican, 2009). In this study, it is either the answers given in the listening tests or the measurement result in acoustic analyses. In the case of the listening tests, the dependent variable is always categorical, nominal, and dichotomous. That means that there are two levels to this variable: correct and incorrect answers. The other possible dependent variable in this study, the acoustic parameter value, is a continuous variable. This means that the acoustic parameters are measured along a continuum, and they have a numerical value. To be very specific, the continuous variables in this study are so-called ratio variables that have the condition that the zero (0) of the measurement means that there is none of this variable. The name ratio variable reflects the fact that one can use the ratio of measurements. So, for example, the SPL doubles every 6 dB, or the ratio of frequencies of two notes an octave apart is 2:1.

The difficulty concerning the variables in this study is that we are trying to obtain measures that reflect a relatively unobservable construct, emotion. The extent to which our measurements actually coincide with the construct of emotion is referred to as construct validity. The way we try to address this problem and increase the validity of this study is by triangulation. We use acoustic parameters to look for a correlation between them and the emotion enacted by the singer. We can also use acoustic parameters to look for a correlation between them and listener evaluations of enacted emotions. We then of course look at the correlation between enacted emotion and listener evaluations. This model of triangulation originally comes from the Brunswikian lens model, but has been further modified to fit vocological study by K. Scherer (Bänziger et al., 2015; Brunswik, 1956).

Another key feature of this study is that it is a comparative study. In this thesis, we use two types of comparison studies: the cross-sectional and the longitudinal. Cross-sectional study compares samples drawn from separate distinguishable sub-groups within a population, in this case singers using the Classical singing technique and singers using CCM singing techniques from a population of singers. The longitudinal study carries repeated measurements on the same group of people over a period of time. In this study, we use the repeated measures design in two ways: 1) to investigate the effects of the teaching intervention, and 2) to investigate the effects of different emotion expressions in the singing voice. In Study III, we use a control group for a comparison with the "Test" group that is undergoing the teaching intervention.

The final defining feature of this study is that it uses statistical analyses to validate hypotheses.

4.1 Participants/ sample

The participants of the individual studies that comprise this thesis were either listeners, singers, or singing acting students. They provided datasets of voice samples and listener evaluations. One of the main aims of scientific study is to be able to generalize from samples. The rationale is that if we take a fair enough sample – one that is representative of all cases in a group that we are studying – then we may generalize our results from that sample to the overall population *with a certain degree of caution* (Coolican, 2009). When we are investigating differences between categories of groups of people, such as popular music singers vs. Classical singers, we need samples representative of the populations from which they are drawn (e.g., all

Classical singers and all popular music singers). We should also make sure that the sample represents all kinds of singers in the category. So, for example, only investigating the highest paid opera stars of the world would most likely yield different results from investigating the struggling artist who just graduated in Classical music. Obtaining a good representative random sample is fairly hard to do in any research. The next best thing is to try to obtain as unbiased a sample as possible, and this is the tactic we have used in the studies of this thesis.

For all of the studies, we have recruited participants using the grapevine and different social media platforms, such as Facebook and LinkedIn. We also had to recruit students, as the interest in participating among the population was low. Our sample was self-selecting, as all of the participants were volunteers. All results should be interpreted against this background: this is a study of a fairly regional self-selecting random sample.

4.1.1 Study I

The main data in this study comprise evaluations of expressed emotion in the listening test. Therefore, the main participants in this study were those who completed the listening test. The number of people who completed the test was 29 (22 females, 7 males, no reported hearing defects). Eight of the listeners were professionally involved in assessing the human voice (singing teachers and vocologists) and 21 were laypeople. Seventeen of the listeners were singers (14 amateur and 3 professionals). The voice samples they listened to were of course sung by singers, but as they are mostly the same participants as in Study II, I will only mention two of them in connection to this: in addition to the female singers, this study also contained two male singers (one Classical and one CCM) who were both professionally educated singers. The voice samples themselves are also explained below under the rubric of Study II.

The data I used in this study consist of 9,048 listening test answers given by the 29 listeners who each went through a 300-sample test battery. The listeners completed a multiple-choice questionnaire on which emotion they perceived (anger, sadness, joy, tenderness, neutral) for each sample. The way I looked at the data was that they were either correct or incorrect in their evaluations.

4.1.2 Study II

Participants: The data in this study consist of voice samples gathered from 11 Finnish female singers, six with a Classical training and five with a CCM background (mean age 32 years, mean singing experience 10 years, minimum of three years of singing lessons at a professional level). All of the singers were portfolio-type singers (singing multiple styles of CCM and/or Classical) who performed at regional or local venues.

Task: The singers were instructed to perform an eight-bar excerpt from a song expressing the emotions of joy, tenderness, sadness, and anger using either the CCM or Classical technique. The song was Gershwin's "Summertime," with Finnish lyrics by Sauvo Puhtila. This song was chosen because it has been composed as an aria, but is widely popular amongst CCM singers as well, so it fits both the Classical and CCM repertoire. The Finnish lyrics depict a nature scene that contain no particular emotion information as such. The singers used a backing track with a neutral accompaniment suitable for both Classical- and CCM-style singing. All subjects were instructed to use the same pitches (males one octave lower) regardless of genre.

The key of the song was D minor, and the pulse was 80 bpm (beats per minute) for all test subjects and every emotion portrayal. The emotion samples were performed in a randomized order and repeated three times. The singers also gave a neutral voice sample without expressing any emotion. This sample was also repeated three times.

Sample data: Vowel [ɑ:] was extracted from three different pitches in each sample for further analysis. The pitches were for the females a, e1, a1 (A3, E4, A4 according to the American system), and A, e, and a for the males (A2, E3, A3). The [ɑ:] samples were extracted from the Finnish words *aikaa* [ˈɑikɑ:](time), *hiljaa* [ˈçiljɑ:] (softly), and *saa* [sɑ:](to enable). The nominal durations of the extracted vowels (including the preceding consonant) were 2.25 s for a, 4.5 s for e1, and 2.25 s for a1 according to the notation and tempo of the song.

4.1.3 Study III

This study had both singers and listeners. The listeners provided evaluations of emotion expressions heard from the voice samples. The singers provided the samples and half of them underwent a teaching intervention.

Listeners: Thirty-two people completed the test (27 females, 5 males, no reported hearing defects), and their answer sets of perceived emotions were all selected for further analysis.

Task: The task was to complete a multiple-choice questionnaire on which emotion they perceived (anger, sadness, joy, tenderness, neutral) for each sample.

Sample data: The data consisted of 8,160 listening evaluations that were either correct or incorrect in relation to the expressed emotion of the singer.

Singers: The singers giving voice samples in this study were Finnish-speaking actor students. They were divided into two groups: a) the test group, which underwent a teaching intervention, and b) the control group, which did not undergo the intervention but had all other regular voice training involved in actor training during the "waiting period."

Test group: The participants of this study were six Finnish actor students, both male (3) and female (3), with a minimum of two years of singing lessons at a professional level. The mean age was 24 (25±4). The mean years of singing experience was six (median 2.25).

Control Group: The control group was a group of six Finnish actor students, both male (3) and female (3), with a minimum of two years of singing lessons at professional level. The mean age was 24 (25±3). The mean years of singing experience was two (median 1).

Task: The actor students were instructed to perform an eight-bar musical excerpt composed especially for the test situation expressing the emotions of joy, tenderness, sadness, and anger plus a neutral state. They all sang using the syllables pa[pa:], da[da:], and fa[fa:], which in themselves carry no emotional content as such. The excerpt was composed using the pentatonic scale in order to avoid sounding too major or too minor. The pulse was 115 bpm for all test subjects and every emotion portrayal. The participants were asked to identify the take they liked the best and that take was selected for further analysis. The same test was issued before and after the teaching intervention.

Sample data: Vowel [ɑ:] was extracted from the last bar in each sample for further analysis. The pitch was f1 (F4, 349.23 Hz) for females, and f (F3, 174.61 Hz) for males. The nominal duration of the extracted vowel (including the preceding consonant) was 3.13 s according to the notation and tempo of the song. A phrase was extracted consisting of two bars from the beginning of the melody. Twelve vowels and 12 phrases were selected for further analysis and the listening test from each participant. The number of vowel samples was 120 and the number of extracted phrases was 120, thus totaling 240 extracted voice samples.

4.2 Techniques of measurement

In an ideal experimental study, the experimenter manipulates the independent variable, *holding all other variables constant*, and measures any change in the dependent variable (Coolican, 2009). This is done so that it would be possible to establish a causal effect. When investigating emotional expression in the singing voice, the process of holding all variables constant is impossible: we have no clear consensus of what an emotion is and entails. We have a somewhat clear consensus on that it manifests differently in different individuals, we know that individuals have their own unique way of expressing themselves, and it has been said that a person's singing and voice use is as individual as a thumbprint. Furthermore, we know that the same sound can be produced using different configurations of the vocal tract (Brown, 2007; Lewis et al., 2010; Miller, 1996; Niedenthal & Ric, 2017; Purves et al., 2013; Seikel et al., 2014).

We cannot know exactly how different individuals express emotion vocally or why listeners perceive certain emotions from samples. There are always situation-specific conditions that affect human action and that is completely fine from the viewpoint of science; we can still very much investigate even if our test subjects are shifting and inconsistent. We just need to adjust for that. In this study, we respond to this problem by having multiple individuals performing the same task and by bringing in "impartial observation" in the form of the voice analysis program Praat. In addition to this, we are very meticulous with our recording procedures and task instructions.

4.2.1 Studies I & II

All recordings were made at the recording studio of Tampere University Speech and Voice Research Laboratory using a Brüel & Kjær 4188 microphone, which was connected to a Brüel & Kjær Mediator 2238 sound level meter with a built-in preamplifier. Audio signals were calibrated for SPL measurements using a Brüel & Kjær 4230 calibrator. The distance between the microphone and test subjects' lips was 40 cm according to the standard of Tampere University Speech and Voice Research Laboratory Studio recordings. The distance of 40 cm is long enough to avoid the proximity effect and close enough so that the signal-to-noise ratio does not deteriorate and the room acoustics does not distort the sound (i.e., the voice sound remains reasonably strong compared to the ambient sound in the room) (Leino & Laukkanen, 1993; Svec, 2018). Samples were recorded with Sound Forge

7 digital audio editing software using a 44.1 kHz sampling rate and a 16-Bit external soundcard (Quad-Capture Roland). All samples were saved as wav. files for further analysis with Praat.

An effort was made to make the experiment as lifelike as possible. Therefore, the singers used a backing track with a neutral accompaniment suitable for both Classical- and CCM-style singing that was played to them via an S-LOGIC ULTRASONE Signature PRO headset. The studio setup also featured a Shure SM58 vocal microphone, which allowed their singing voice to be mixed in with the backing track as they were singing.

The [ɑ:] vowels were extracted from the sung excerpts using Sound Forge 7 audio editing software. The samples were cut right after the preceding consonant. The duration of the sample varied between 0.6267 s and 4.8063 s depending on how the test subject had interpreted the time value of the notation. The tail end of the vowel was left as the singer interpreted it (the nominal note durations were 2.25 s or 4.5 s).

The listening test was a web-based test that the participants could access from their own devices. The test was coded to an altervista.org homepage using html (language) and php scripts (the way that the page functions).

4.2.2 Study III

All test condition recordings were made at the recording studio of Tampere University Speech and Voice Research Laboratory similarly to studies I & II. Samples were recorded with Sound Forge Pro 11.0 digital audio editing software using a 44.1 kHz sampling rate and a 16-Bit external soundcard (Focusrite Scarlett 2-i-4).

All control samples were recorded at the recording studio 365 of the University of Arts Helsinki using a Brüel & Kjær 4188 microphone, which was connected to a Brüel & Kjær Mediator 2238 sound level meter with a built-in preamplifier. Audio signals were calibrated using a Brüel & Kjær 4230 calibrator. The distance between the microphone and test subjects' lips was 40 cm according to the standard of Tampere University Speech and Voice Research Laboratory. Samples were recorded with Cubase 10 digital audio editing software using a 44.1 kHz sampling rate and a 16-Bit external soundcard RME Babyface Pro.

The singers used a backing track with a neutral accompaniment that was played to them via a SONY MDR V-700 headset through Zoom H-4 in the test condition and a Sennheiser HD 25-SP II 60 ohm headset in the control condition.

The [ɑ:] vowels and the phrases were extracted from the sung excerpts using Reaper audio editing software. The vowel samples were cut right after the preceding consonant. The duration of the sample vowels varied between 1.2 s and 4.04 s depending on how the test subject had interpreted the time value of the notation. The tail end of the vowel was left as the singer interpreted it (the nominal note durations were 3.13 s).

All samples were saved as .wav files for further analysis with Praat, and the listening test was coded in the same way as in Studies I & II.

4.2.3 Praat & measurement of acoustic parameters

Praat (Version 6.0.19) was used to run the acoustic analysis. It was originally designed to do speech analysis in phonetics, but it can be used for analysis of the singing voice as well. The program can run on a wide range of operating systems and continues to be developed by its creators, Paul Boersma and David Weenink of the University of Amsterdam (Boersma & Weenink, 2014). It allows spectral, fundamental frequency, formant, and SPL analyses and the examination of jitter, shimmer, and voice breaks, among other things. Praat has a high content validity – i.e., many experts agree that it measures what it claims to measure. It also has concurrent validity in comparison to other sound analysis software and predictive value in the sense that if the extracted sound parameter values were to reconstruct a sound, we could imagine how that sound would sound by merely looking at the numbers.

Twenty different measures were used in this thesis.

Fundamental frequency -
fo using autocorrelation with a bandwidth set to 75-600 Hz.

SPL – SPL was measured with reference to the calibration signal recorded. First, the difference between the known SPL of the calibration signal and the SPL value given by Praat for the calibration signal was calculated. This difference was then added to the SPL value given by Praat for the speech sample of interest, thus achieving the real SPL of the speech sample.

F1-F5 - Formant frequencies F1-F5 were measured from the spectral structure as a function of time up to 5500 Hz. Praat performs a short-term spectral analysis, approximating the spectrum of each analysis frame by the number of formants.

HNR - with the Harmonicity object.

ALPHA RATIO - The cut-off frequency was set to 1500 Hz instead of the more traditional 1000 Hz in order to better suit the analysis of the singing voice.

SHIMMER - For Shimmer, the three-point amplitude perturbation quotient (Shimmer apq3 – the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude) and the five-point amplitude perturbation quotient (apq5 – the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbors, divided by the average amplitude) were measured.

JITTER - Two measures of jitter were used: Relative average perturbation, RAP (defined as the average absolute difference between an interval [glottal period] and the average of it and its two neighbors, divided by the average time between two consecutive points), and five-point period perturbation quotient ppq5 (the average absolute difference between an interval and the average of it and its four closest neighbors, divided by the average time between two consecutive points) (Boersma & Weenink, 2014; Teixeira, Oliveira, & Lopes, 2013).

VIBRATO – The vibrato rate (number of *fo* undulations per second) and extent (how far *fo* departs from its average during a vibrato cycle) were measured as well as the rate and extent of the amplitude vibrato. We sampled the vibrato on a 0.01sec timeframe looking for the local positive and negative peaks in the sound (Hz and dB) and calculated the vibratos as the mean of time differences between the local peaks and negative peaks in the sound. First the 0.01s timeframe was extracted {time t [s], f_0 [Hz], SPL [dB]}. Then, we looked at the f_0 [Hz] column for the time moments where $f_0(i-2)<f_0(i-1)$ and $f_0(i-1)>f(i)$, i.e., we looked for the local peaks in the sound. We marked the events in time $tmax(i) = t(i-1)$. Then we looked at the f_0 [Hz] column for the time moments where $f_0(i-2)>f_0(i-1)$ and $f_0(i-1)<f(i)$, i.e., we looked for the local negative peaks in the sound. We marked these times $tmin(i) = t(i-1)$. Vibrato is calculated as the mean of time differences between the local peaks and negative peaks in the following way:

mean Hz & std Hz = rate (divide the results by two to obtain the correct value)

Mean amp & std amp = extent (measured from maximum peak to minimum peak)

mean(Hz) = mean(1/(tmax(i)-tmin(i)))

std(Hz) = std(1/(tmax(i)-tmin(i)))

mean(amp) = mean(f_0(imax)-f_0(imin))

std(amp) = std(f_0(imax)-f_0(imin))

The amplitude vibratos were calculated in the same way referring to the SPL [dB] column.

ATTACK, SUSTAIN, RELEASE – We also looked into how long it takes from the moment the sound initiates until it reaches its 90% peak level (attack time) and the time that the note is sustained at 25% to 75% dB level in comparison to the maximum level (sustain time). We also calculated the time when the SPL drops to zero from the ¼ level of the SPL (release time). A timeframe of 0.01 sec was used for sampling {time t [s], f_0 [Hz], SPL [dB]}. For every sample, we searched for the timeframe where t_imax = 9/10*max(SPL). Attack time and level were set by choosing the smallest timeframe, t_min, where 90% of the maximum SPL level is reached. Then we looked for the timeframe where t_imax = 3/4*max(SPL); we chose the largest timeframe, t_max1, where 75% of the maximum SPL is reached. Sustain time is the subtraction result of the last possible moment where SPL 75%*max(SPL), and attack time difference with a level of 75%*max(SPL), sustain = {t_max1-t_min, 0.9*max(intens)}. Lastly, we extrapolated the timeframes when t_max2 and when the SPL is zero, and chose the release time with the largest timeframe t_max2 release = {max(t_max2), 0}.

4.3 Statistical tests

When we are assessing a group of people for an experimental variable, we are typically not interested in individual samples but rather in the underlying population from which they are drawn. The measures we draw from the sample are often taken as an estimate of the same measures of a population. We assume that the sample statistics reflect the population as a whole. This is why in order to run these statistical tests and to have confidence in their results, the data population from which the samples have been drawn should be a normal distribution.

The parametric tests that we use in this study work as they should when the data is normally distributed. In this study, we have both larger and smaller data sets. In the smaller data sets, it is usually enough to look at the samples and distribution of data in terms of their numerical central tendency and dispersion. This is the case in the acoustic analyses in Study III: the data units are so small that it does not allow for a strong statistical deduction. This does not mean that the data we gathered is of no importance; it just means that it cannot be extrapolated to automatically cover the whole population of singers. Instead, what we did in Study III was to use non-parametric statistical tests (which fit better with small, non-normally distributed data sets) to give some support to our arguments concerning acoustic analyses.

All statistical analyses were done with SPSS (v.17; SPSS Inc., Chicago, IL).

4.3.1 Binomial one proportion z-test

Binomial test (one proportion z-test) was used to evaluate the probability that the observed percentage of the correctly recognized emotions could have resulted from random guessing.

A z score is the number of standard deviations a score is from the mean, and it is a specific type of standard score. A one proportion z-test is a hypothesis test that tries to make a statement about the population proportion (p) for a certain population attribute. In this case, we used it to distinguish knowing from guessing in the listening test. The data used in this test were binomial. This means that there were only two possible measurements for the data: correct and incorrect. The participants either perceived the acoustic emotion expression correctly, or they did not. With a large number of samples, we could approximate the binomial distribution by normal distribution. The one proportion z-test has two non-overlapping hypotheses, the null and the alternative hypothesis. The null hypothesis is a statement about the population proportion, which corresponds to the assumption of no effect. The listening test contained five different emotions, which meant that the expected percentage of correctly recognized emotions in the case of random guessing would be 20%. Therefore, the null hypothesis is H0: recognition % = 1/5. The alternative hypothesis is the complementary hypothesis to the null hypothesis: H1: recognition % <>1/5. The null hypothesis is rejected when the z-statistic lies in the rejection region, which is determined by the significance level (α). This means that the observed percentage of correctly recognized emotions differs statistically significantly from random guessing if the p-value of the test is $p<0.05$. The alpha

level in this study was set to 0.05. If the p-value is greater than alpha, the null hypothesis that the recognition % is the same as it would be in guessing (0.2) stands (Mathcracker.com, 2020, Cooligan, 2009).

4.3.2 Pearson's chi-squared test of homogeneity

Pearson's chi-squared test of homogeneity compares the distribution of counts (number of cases observed) for two (or more) groups using the same categorical variable. It allows one to determine whether the proportions are statistically significantly different in the different groups.

In our study, the test was used to evaluate the probability that two groups of listening test results would have the same percentage of correctly recognized emotions (Study I) and to compare the correct recognition within the same population under different conditions (Study III).

The statistical significance of the difference between the correctly recognized proportions of emotions in different subgroups was tested using null hypothesis H0: $p_1 = p_2$ against the alternative hypothesis H1: $p_1 \neq p_2$, where p_1 is the proportion of correctly recognized emotions in subgroup 1 and p_2 is the proportion of correctly recognized emotions in subgroup 2. The null hypothesis is rejected if the significance level of the test shows a low value $p<0.05$.

The statistical test was performed by presenting observed numbers of emotions recognized in a two-by-two contingency table and applying Pearson's chi-squared test of homogeneity. The contingency table was formed such that the results of the first group were in the first row of the table and the results of the second group were in the second row of the table. The numbers of correctly recognized emotions were in the first column of the table and the numbers of incorrectly recognized emotions were in the second column of the table. Application of this chi-squared test requires that the number of observations in all elements of the table should be larger than five (Coolican, 2009). This requirement was fulfilled by the data.

4.3.3 Cronbach's alpha

Cronbach's alpha was used to evaluate the reliability of the internal consistency of listener evaluations. Values >0.7 indicate acceptable internal consistency of the data.

An important aspect of reliability in the study is that the inter-rater reliability is high. This means that we want the listeners to agree independently in their assessment of the emotion they perceive from the voice samples. Cronbach's alpha is one of the most commonly used statistics for estimating a test's reliability. Its value depends on how people vary on individual items, in this case the evaluation of a certain voice sample. If they vary a lot (on the individual item) relative to how much they vary overall in the test, then the test is assessed as unreliable and the value for alpha will be low. Alpha is equivalent to the average of all possible split-half reliability values that could be calculated for the data set. Good reliability alpha values range from around .75 to 1 (Coolican, 2009).

4.3.4 RM-ANOVA

To evaluate whether the parameter values extracted with Praat differed across emotions for each parameter, we computed a repeated measures analysis of variance (RM-ANOVA) of the general linear model (GLM) with SPSS (v.17; SPSS Inc., Chicago, IL).

We used the one-way repeated measures analysis of variance to determine whether there were statistically significant differences between the means of five levels of a within-subjects factor. (Factor is just another name for an independent variable, and within-subjects refers to one participant singing the same melody using five different emotion expressions. In other words, the participants are measured on the same dependent variable while undergoing the same conditions, i.e., we are tracking e.g., the SPL changes (the dependent variable) of 12 singers at five different emotion expressions (the independent variable) to determine whether the SPL is a contributing parameter in emotion expression in singing). The null hypothesis for our RM-ANOVA is: H_O: $\mu_{joy} = \mu_{tenderness} = \mu_{neutral} = \mu_{sadness} = \mu_{anger}$, and the alternative hypothesis is Ha: $\mu joy \neq \mu tenderness \neq \mu neutral \neq \mu sadness \neq \mu anger$ (μ is the population mean of a measured sound parameter (e.g., SPL).)

The RM-ANOVA does not tell in which particular way the means of the different levels of the within-subjects factor differ in population, only that they do differ in some way. To further investigate how the emotion expressions differed from each other, we ran pairwise comparisons with Bonferroni corrections.

The Bonferroni correction is considered to be one of the most suitable adjustments for making a multiple post hoc comparison for a one-way repeated ANOVA (Maxwell & Delaney, 2004). It tests all possible pairwise combinations of

levels (emotion categories) of the within-subject factor (sound parameter value extracted from the melody sung with emotional expression). The test gives a statistical significance level (p) for each pairwise comparison and gives confidence intervals for the mean difference for each comparison. It basically runs multiple paired-samples t-tests, but adjusts the confidence intervals and p-values to take account of making multiple comparisons and thus avoids a type-I error (a false positive).

4.3.5 Univariate analysis (GLM)

Univariate analysis gives a regression analysis and the analyses of variance for one dependent variable (acoustic parameter change due to emotional expression) by one or more variable, in this case pitch and SPL. With the GLM procedure in SPSS, it is possible to test the null hypotheses about the effects of other variables on the means of various groupings of a single dependent variable. The effects of covariates and covariate interactions with factors can be included (IBM SPSS Manual, 2020). The null hypothesis for this test is the same as in RM-ANOVA,
H_O: $\mu1 = \mu2 = \mu3 = \mu4 = \mu5$, and the alternative hypothesis Ha: $\mu1 \neq \mu2 \neq \mu3 \neq \mu4 \neq \mu5$.

It is known that SPL and *fo* have an effect on different sound parameters such as perturbations, HNR, and alpha ratio. To further validate the results obtained from the RM-ANOVA, we used the univariate analysis of the GLM to determine if these parameters varied individually according to emotion and not just related to F_0 and SPL. In univariate analysis, each parameter was set as a dependent variable at a time, emotion was set as the fixed factor, and F_0 and SPL were set as co-variates.

4.3.6 T-test (unrelated samples)

The unrelated samples t-test was used to compare the number of correct answers given from test group samples and the number of correct answers given from the control group samples. The null hypothesis here is that the two populations from which the two samples have been drawn have equal means (H0: $p_1 = p_2$ against the alternative hypothesis H1: $p_1 \neq p_2$). Separate t-tests were run for the first recording (before) and second recording (after). Recognition between the test group samples

and control samples was interpreted to differ statistically significantly if the p-value of the test was $p<0.05$.

4.3.7 Friedman test

The Friedman test is the non-parametric alternative to the one-way repeated measures ANOVA test. It is used to determine if there is a statistically significant difference between the distributions of three or more related groups. We used the Friedman test to evaluate whether the acoustic parameter values from the short vowel samples differed across emotions. We decided to use the Friedman test instead of RM-ANOVA, because the assumption of normality was markedly violated due to the relatively small dataset. We ran the Friedman test separately for the test group and control group and the before and after conditions. Bonferroni corrections were used for multiple comparisons. The null hypothesis is exactly the same as in the RM-ANOVA and the univariate test, H_0: $\mu1 = \mu2 = \mu3 = \mu4 = \mu5$, and the alternative hypothesis is Ha: $\mu1 \neq \mu2 \neq \mu3 \neq \mu4 \neq \mu5$.

4.4 The parameter modulation technique used in training

The purpose of the parameter modulation technique was to introduce the student to the basic acoustic characteristics (and their perceptual correlates) typically observed in the expressions of joy, tenderness, sadness, and anger (Hakanpää et al., 2019, 2021a). The technique itself refers to the voluntary variation of these voice characteristics so that they result in a clearly recognizable emotional expression.

For the purposes of this study, we asked the students to try out the following voice quality manipulations:

Joy: Loud and well projecting voice, phonation balance (neither breathy nor pressed), bright timbre, and inclusion of vibrato acceptable.

Tenderness: Moderate loudness and projection, slightly breathy phonation, but clear articulation, bright timbre, and no perturbation.

Sadness: Soft voice with a few volume outbursts, breathy phonation, unclear articulation, dark vocal timbre, and a lot of vibrato (both *f0* & amplitude) and noise.

Anger: Loud volume with pressed phonation, very clear articulation, sharp vocal timbre, and no vibrato.

The starting point to the voice modulations was the participants' habitual neutral voice. In order to use the parameter modulation technique safely, students should

be aware of what their individual optimal (well-balanced effortless) voice use is like. The extent to which the parameter manipulation can be executed (i.e., how wide deviations from optimum can be introduced) needs to be scaled individually and also for the aesthetics of the singing style in use. As each exercise needs to be fitted individually to the students' conceptual understanding of the voice and to their individual way of using it, a specific description of the exercises one should use when working with the parameter modulation technique cannot be given. Instead, we give a general description of how the parameter modulation was taught.

In general, when working with the parameter modulation technique, it would be good to give a reference sound for the student either by modeling the target sound oneself or giving some other type of reference sound. Modeling the target sound in relation to one's own neutral sound will give the student a conceptual idea of the voice quality and will light up the mirror neuron system (Rizzolatti & Gallese, 1996), which will make it easier to grasp the idea of how the target sound might be produced. Discussing the sound with students and giving feedback after they have attempted the sound will further aid in the conceptualization of the sound and help to build motor programs for the sound. Directing the students' attention to the way that the voice sounds might help them to focus externally, which might be beneficial for the learning process. Allowing for enough repetition will help the motor schemas to develop and the neural network to consolidate (Guadagnoll & Lee, 2004; Wulf et al., 1999; Nolen-Hoeksma et al., 2009; Schmidt, 1975). Giving enough theoretical background so that students can start building their own knowledge base is required for deep learning to occur. The theoretical information provided needs to be scaled to the students' skill level (Bloom, 1956; Guadagnoll & Lee, 2004). Talking about the relevance of the exercises and allowing time for reflection can help to motivate training and reinforce learning (Maas et al., 2008; Guadagnoli & Kohl, 2001).

4.4.1 Volume control

For volume control, we used exercises exploring the loudness range of each individual student from the softest possible to the loudest. We discussed each participant's habitual loudness use, comfort loudness, air flow regulation, vocal fold adduction, and influence of the oral cavity and mouth opening on the perceived loudness.

Previous research has identified SPL as a prominent indicator of emotion expression. To put it simply, a high SPL is used for high energy emotions and a low

SPL for low-energy emotions (Hakanpää et al., 2021a, 2021b; Jansens et al., 1997; Livingstone et al., 2014; Scherer, Sundberg, et al., 2017; Scherer et al., 2015; Sundberg et al., 1995, 2021). The physiological characteristics for increasing loudness include increased P_{sub} and increased amplitude with which the vocal folds open and close the glottis (Echternach, Burk, Burdumy, Traser, & Richter, 2016; Herbst, Hess, Müller, Švec, & Sundberg, 2015; Sundberg, 2017; Zhang, 2015), increased medial compression of the vocal folds, increased grade of vocal fold adduction, more air pressure, possible modification of the lower vocal tract (Sundberg, 1987), raising of the first resonance frequency (Vurma, 2020), opening of the jaw and making the lip opening larger (Sundberg, 1987), increased pharyngeal volume, and the raising or lowering of the larynx (Sundberg, 1987; Echternach et al., 2016). The physiological characteristics of decreasing loudness works by decreasing the aforementioned elements. Therefore, suitable exercises for volume control include different kinds of breathing and breath management exercises; body alignment and muscle work for building support; experimenting with soft to moderate to strong vocal fold adduction and soft, simultaneous, and hard vocal attacks; practicing jaw opening, lip opening, and increasing pharyngeal volume while keeping the laryngeal position steady; and, most importantly, finding a way to support the voice that works for the student. Studies have continuously identified different breath management patterns in professional singers (Lam Tang, Boliek, & Rieger, 2008; Salomoni et al., 2016; Traser et al., 2016), which suggests that there is more than one way of supporting the voice correctly.

According to previous research on emotion expression, the following apply:
Joy: high SPL, loud
Tenderness: low SPL, soft/quiet
Neutral: moderate SPL, moderate loudness
Sadness: low SPL, soft/quiet
Anger: high SPL, loud.

4.4.2 Phonation

For phonation, we used phonation balance exercises fitted to the individual needs of the student (soft attack and general "hypofunction" for a habitually "hyperfunctional" voice, and vice versa). Phonation with barely abducted/barely adducted vocal folds is said to produce a resonant voice (Grillo & Verdolini, 2008; Myers & Finnegan, 2015; Verdolini et al., 1998). The goal of our exercises was to

establish this zone for the students so that they can then safely depart from it. We also drilled polar opposite exercises, ranging from a very breathy voice through an optimal sound balance to pressed phonation. The goal of these exercises was to clearly demonstrate the perceptual (both acoustic and tactile/sensory) differences between the different modes of phonation.

The phonation part of emotion expression plays both to the source and the filter domains of voice use, and therefore it can be seen as a part of loudness control or part of timbre control. Phonation affects the timbre domain depending on the adduction force of the vocal folds – the spectrum gets steeper or flatter. Loose adduction causes a large level difference between the lowest partials (H1-H2, i.e., H1 dominates the spectrum), and tight adduction causes a lowering of the level difference between the lowest partials (H1-H2) and results in stronger relative spectral energy above 500 Hz (or 1 kHz) (Fant, 1970; Pulkki & Karjalainen, 2015; Sundberg, 1987; Titze, 1994; Titze et al., 2015; Welch et al., 2000). Raising the formant frequencies makes the voice timbre brighter, while lowering them makes the timbre darker.

In the acoustic analysis, phonation mode can be detected by combining information about SPL and formant analyses. Typically, high energy emotions are expressed with a flatter spectrum so that the formants pack closer together, while in low-energy emotions, they tend to move further apart from each other. Low energy emotions usually have more energy at the lower part of the spectrum, while high energy emotions tend to have energy condensed at the higher frequencies. A higher SPL usually means stronger vocal fold adduction and a lower SPL softer phonation (Hakanpää, Waaramaa, & Laukkanen, 2021; Hakanpää, Waaramaa, & Laukkanen, 2019; Jansens et al., 1997; Livingstone et al., 2014; Scherer, Sundberg, et al., 2017; Scherer et al., 2015; Sundberg et al., 1995, 2021).

Suitable exercises for practicing phonation could be found for example from literature pertaining to fixing faulty vocal habits, such as adding an H-sound before the vowel for breathy phonation or giving a hard glottal attack for vowels in order to go towards a more pressed phonation (Behrman & Haskell, 2019).

According to previous research on emotion expression, the following apply:

Joy: More adductive tension and medial compression, and simultaneous vocal attack.

Tenderness: Less adductive tension and medial compression, soft vocal attack, and less muscular effort.

Neutral: Moderate adductive tension and medial compression, and simultaneous vocal attack.

Sadness: Less adductive tension and medial compression, soft vocal attack, and less muscular effort.

Anger: More adductive tension and medial compression, hard vocal attack, and much muscular effort.

4.4.3 Articulation

The articulation part of the voice is detectable from the different formant analyses in the acoustic analyses. Acoustically speaking, articulation refers to the way the vocal tract reinforces or attenuates its own resonances (Hakanpää, Waaramaa, & Laukkanen, 2021; Hakanpää, Waaramaa, & Laukkanen, 2019; Jansens et al., 1997; Livingstone et al., 2014; Scherer, Sundberg, et al., 2017; Scherer et al., 2015; Sundberg et al., 1995, 2021).

For resonance and articulation, we used exercises that shape the vocal tract in various ways. The articulatory exercises addressed the different possibilities of physiological positioning of the tongue, jaw, velum, and lips. Varying the type of articulation one uses can help create different sound colors. The positioning of the articulators also affects the resonances we are able to exploit for sound reinforcement. Moving the tongue along a front-back or up-down axis; lifting or lowering the larynx; working with the lips (opening, closing, stretching, puckering, pouting, pursing, etc.); varying the shape and size of the oral cavity, pharynx, and nasopharynx; moving the cheeks; or opening and closing the nasal port can all change the perceptual quality of sound (Seikel et al., 2014).

The lips have a role in lengthening and shortening the vocal tract, which effects the formant frequencies and makes the voice color sound darker or shriller (Sundberg, 1987; Tartter, 1980). Horizontal expansion of the interlabial space raises the formant frequencies and makes the timbre brighter. The latitudinal rounding of the lips has a similar effect as longitudinal protrusion; both tend to lower the formant frequencies. The different shapes of the lip-openings are created by the actions of different muscles or by different degrees of tension in the same muscle (Laver, 1980). We asked our students to experiment with elongation of the vocal tract by using exercises extending the lips outwards (protruding) and shortening the tract by retracting them. In addition to the longitudinal modification of the vocal tract, latitudinal lip movement can also be used for creating different interlabial spaces (Laver, 1980). In this study, we also experimented with different lip openings and with restricted lip movement.

Raising the tongue and giving a slight fronting to the mass of the tongue (e.g., like in the vowel /i/) gives a brighter timbre, raising the formant frequencies. In this setting, the second formant is maximally high and close to F3. The distance between the first and the second formant is usually large. A raising of the larynx (as a consequence of lifting and moving the tongue root forward) results in a shortening of the pharynx and the narrowing of the lower part of it as the wall tissues pile up and fill part of the lower pharynx. When the vocal tract shortens, the formants tend to rise (Sundberg, 1987). Retracting the tongue, moving the center of the mass of the tongue backwards and slightly downwards, gives a darker timbre, lowering the formant frequencies. In this type of setting, F1 is usually a little higher than in a neutral setting, and the second formant is typically lower than in neutral (Laver, 1980). The tongue tip does not play a major role in changing the habitual voice quality, but it is important in articulation, and using the tongue tip effectively in settings where the tongue's center of mass is in different locations is crucial for communication (Edmondson & Esling, 2006; Laver, 1980; Lindblom & Sundberg, 1972; Sundberg, 1987). In this study, we used exercises for moving the tongue forward and backward, flattening the tongue, pulling the tongue back and up, and working with the intrinsic muscles of the tongue. The point of these exercises was to acknowledge the major role that the tongue has in shaping the oral cavity and the resulting sound. The students were encouraged to look for sounds that would (in their opinion) fit the acoustic descriptions given to the target emotions.

F1 rises as the jaw opening becomes larger. In a closed jaw setting, the frequency of the first formant tends to drop and its range decreases. Higher formants are proportionally less affected by the jaw opening, but they too tend to rise with the degree of jaw opening (Lindblom & Sundberg, 1971). The relationship between jaw and lip positions is such that each can magnify or diminish the other's effect. That is why it is important to check the lip effect before specifying acoustic phenomena to any particular mandibular setting (Laver, 1980). Exercises of jaw movement were presented on an open-closed continuum. The students were given different exercises addressing the relaxed, open, and closed jaw positions, and they were instructed to experiment with different jaw openings as well as a fixed jaw position (with a bite block). The aim of these exercises was to demonstrate the full range of jaw movement as well as the possibility to hold the jaw in place and still sound intelligible. Again, the students were encouraged to explore and pick different sounds to be used in emotion expression.

The action of the velopharyngeal system is a mixture of valvular movement on the part of the soft palate and sphincter movement by the superior constrictor and

its related fibers (Laver, 1980; Seikel et al., 2014). The most obvious perceptual marker of velopharyngeal activity is the nasalized/denasalized voice. There are a lot of misconceptions about what constitutes a nasal sound, however (Laver, 1980), and furthermore, it is customary in voice pedagogy to promote firm closure of the nasal port and elevation of the velopharyngeal area (Bunch, 1997) to produce a better transfer function for the voice. The velum exercises we used were either velum up or velum down, and the way we approached them was through mind imagery instructions such as "smell the flower" and "like you are just about to cry" (Chapman, 2006; Miller, 1996; Sadolin, 2008; Steinhauer, McDonald Klimek, & Estill, 2017). We felt that a simple instruction of up and down would better enable the students to focus on the way the vocal sound changes in changing the velopharyngeal settings without getting too caught up on what is a nasal/denasal voice or having to think about the concept of transfer function and its acoustic effects.

Suitable exercises for practicing different cavity shapes include different types of articulation exercises (enunciation, speed, extent) and different types of isolation exercises for the articulators (e.g., practicing raising and fronting the tongue as a muscle movement, practicing raising and fronting with a vowel sound, practicing raising and fronting in a song, etc.).

According to previous research on emotion expression, the following apply:

Joy: Retraction of the lips and advancement of the root of the tongue. Opening the jaw and lifting the velum to prevent nasality.

Tenderness: Less horizontal expansion of the interlabial space and smaller opening of the jaw. Advancement of the root of the tongue and lifting the velum to prevent nasality.

Neutral: Moderate horizontal expansion of the interlabial space and moderate opening of the jaw. Neutral tongue position. Lifting the velum to prevent nasality.

Sadness: Less horizontal expansion of the interlabial space, latitudinal rounding and longitudinal protrusion of the lips. Smaller opening of the jaw. Pulling the tongue back and lowering the velum to promote nasality.

Anger: Horizontal expansion of the interlabial space, possible latitudinal rounding and longitudinal protrusion while opening the jaw maximally. (Although in subdued/pent-up rage. clenching the teeth might also work.) Advancement or retracting of the tongue. Lifting the velum to prevent nasality.

4.4.4 Perturbation element

Jitter and shimmer are seen as the result of 1) small asymmetries or variations in cricothyroid muscle tension, 2) fluctuations in subglottal pressure, 3) perturbation in the mucous of the vocal folds, or 4) a combination of these elements (van Puyvelde et al., 2018). Lower jitter and shimmer values have been found to be indicative of a perceptually clearer sounding voice (Warhurst et al., 2012). Increasing the SPL decreases jitter and shimmer (Brockmann-Bauser et al., 2018). The *fo* vibrato is characterized by an undulation of the fundamental frequency, and the amplitude vibrato is characterized by a pulsation of subglottal pressure, the frequency variation of formants, or the frequency variation itself (Sundberg, 1994). The *fo* vibrato is thought to be produced by the pulsating contractions of the cricothyroid muscle. The other, perceptually different type of vibrato that is characterized by a pulsation of subglottal pressure is sometimes referred to as the "hammer vibrato." This type of vibrato is produced with the tremolo mechanism by rapidly alternating the abduction and adduction of the vocal folds (Bunch, 1997; Sundberg, 1987, 1994). The perturbation and the vibrato have similar production elements in them: namely the contractions of the cricothyroid muscle and the pulsation of subglottic pressure. Perturbation is usually less voluntary than vibrato, which is why we thought that exaggerating vibrato could lead to a perceptually perturberant effect in the voice. An overly large vibrato extent can leave the voice sounding wobbly, too slow a vibrato rate – especially in a hammer type of vibrato – can leave audible gaps in the voice, and too fast a vibrato rate can sound nervous. Adding a vocal tract-induced noise component to a steady vocal fold vibration cycle, such as singers do in dist sounds (Borch, Sundberg, Lindestad, & Thalén, 2004) is also a possibility, but it often takes a considerable amount of practice to be done safely, and for the purposes of this study we felt that the unbalanced vibrato was the better option. Turbulence noise may also be added to the voice by leaving a gap in the glottis and using sufficient subglottal pressure. The perceived voice contains a hissing component. Both perturbation and turbulence noise may contribute to a decrease in HNR.

In this study, we used atypical support, altering between lax and pressed, both at the level of the diaphragm/abdomen (as in panting) and the vocal folds (alternating between pressed and breathy). We also drilled an overly large *fo* vibrato, alternating between several semitones, and we experimented with cutting the sound altogether as in tremolo. We experimented with insufficient support and forceful support to affect the sound production. Some of our test subjects also experimented with ventricular phonation (when it came naturally – we did not demand it ourselves).

Suggested exercises for perturbation effects could include long tones with *f0* undulation and long tones with tremolo. Atypical use of support and, for the more advanced students, the use of more demanding vocal effects such as distortion, growls, and screams could also be proposed.

According to previous research on emotion expression, the following apply:

Joy: Low perturbation variation, high perturbation level, and less jitter and shimmer due to increased SPL. Greater HNR, suggesting a clearer sound.

Tenderness: Less waveform irregularity, but still more jitter and shimmer due to decreased SPL in comparison to high-activity emotions. Greater HNR, suggesting a clearer sound.

Neutral: -

Sadness: More jitter and shimmer due to decreased SPL, high perturbation variation, but low perturbation level. Smaller HNR, suggesting a noisy sound.

Anger: High on the perturbation level, less jitter due to increased SPL, but also a smaller HNR possibly due to excessive laryngeal tension and distortion in the signal.

4.5 Ethical statement and distribution of work

The design and implementation of the study followed the principles of good scientific practice described by the Finnish Research Ethics Board (2012, p. 30). The data collection was made between 2016 and 2019. Subjects were told at the recruitment stage the general purpose of the study, and a more detailed study design was revealed immediately after the study. The study did not cause any physical or mental harm to the subjects, and they had the right to discontinue the study at any time. In the research situation, the participants were informed about anonymity, data protection, and data management.

The voice sample files in this study were coded, and the data were randomized. All the analyses were run with numerical coding. The results of this study are published with average values, so no personal data can be inferred from the results. In addition, we have asked the singers for consent to play, analyze, and possibly reuse the samples. The sound samples are stored in wav. files according to the instructions of the National Digital Library (2016, p. 4). The data are stored and handled according to the GDPR guidelines (2016/679, General Data Protection Regulation).

In Study I, I was the first author and responsible for designing the experiment; collecting, organizing, and analyzing the data; and writing the report.

In Study II, I was the first author and responsible for designing the experiment; collecting, organizing, and analyzing the data (except for the univariate analysis); and writing the report.

In Study III, I was the first author and responsible for designing the experiment; teaching the parameter modulation technique; collecting, organizing, and analyzing the data; and writing the report.

5 RESULTS

The study investigated the acoustic features and perceptual recognition of emotion expressions for four different emotions: joy, tenderness, sadness, and anger. We wanted to know whether it was possible to teach emotion expression in singing by using vocological information about voice quality and implementing this information to form a training pattern (the parameter modulation technique).

The logic of our study was to first establish whether it was possible to perceive emotion from the singing voice (Study I), then to take a deeper look into the acoustic compilation of the voices to find out what kinds of elements might account for the recognition (Study II), and finally to come up with a training pattern that would drill the acoustic elements found typical for expressing the aforementioned emotions to see if training in this way would help to make expressing these emotions easier (Study III). The general result of this study was that the communication of emotion (using the singing voice) became easier after incorporating vocological information into the regular singing training.

5.1 Experiment 1

The first experiment was designed to gain information about emotion recognition from vowel samples sung at three different pitches. We also investigated if there were differences in the recognition of emotions when listening to samples sung using the CCM technique or Classical technique. Our results indicated that at 30%, recognition of emotion expression exceeded random guessing (H0: $p = 1/5$: z-value 24, p-value $\leq .001$). According to the evaluation using Pearson's Chi-square test of homogeneity, the recognition of emotion expressions differed statistically significantly in favor of the samples sung using the CCM technique in comparison to the samples sung using the Classical singing technique (z-value 120.2, p-value $\leq .001$) in the female samples. Recognition of emotion from the male samples showed no difference between the genres.

The three different pitches affected the recognition of emotion expressions in such a way that overall, the low frequency samples were recognized more poorly

than the high frequency samples (Table 3, Study I). Sadness was more easily recognized from a low pitch and less easily recognized from a high pitch. The recognition of joy was better from a high pitch and poorer from a low pitch. The recognition of tenderness was slightly easier at a middle pitch. Anger was best recognized from high frequencies in both Classical and CCM samples (Table 4).

For the female singers' samples, recognition of emotion was consistently easier from the CCM singing in all other emotions except for sadness, which was better recognized from samples sung using the Classical singing technique. From the male singers' samples, anger and sadness were better recognized from the Classical singing technique, and joy, tenderness, and neutral were better recognized from the CCM singing technique (Table 4, Study I).

5.1.1 Appraisals of valence and activation in the first experiment

For this experiment, we derived valence and activation from the listeners' answers. In these data, activation was more accurately perceived from all pitches in comparison to valence. At a low pitch, valence was more accurately perceived for joy and anger, whereas activation was more accurately perceived for tenderness and sadness. At a middle pitch, the tendency was similar with the female samples, but with the male samples, the valence was more accurately perceived only for joy. Activity was perceived more accurately at high pitches for all other emotions except for female tenderness (for which valence was perceived more accurately). The perception of valence and activation from the female samples was most accurate at a high pitch. From the male samples, valence was correctly perceived from a middle pitch most accurately, whereas activation was most accurately perceived from a high pitch.

When we compared the perceived accuracy of valence and activation between the CCM and Classical style, we found that in the female samples, both valence and activation were more accurately perceived from the CCM samples, whereas with the males, it was the other way around (Table 5, Study I).

Table 4. Correctly recognized emotions, differences in recognition between CCM and Classical singing in three different pitches, and the internal consistency of the answers (statistical significance level $\alpha < .05$).

				%	z-value	p-value	Cronbach's alpha
Female	CCM	Joy		24.3%	3.36	0.001	0.90
			low pitch	8.2%	-5.29	<0.001	0.17
			medium pitch	16.6%	-1.51	0.131	0.20
			high pitch	48.3%	12.63	<0.001	0.86
		Tenderness		33.1%	10.15	<0.001	0.78
			low pitch	24.5%	2.02	0.044	0.82
			medium pitch	39.5%	8.71	<0.001	0.73
			high pitch	35.4%	6.89	<0.001	0.77
		Neutral		29.5%	7.03	<0.001	0.73
			low pitch	32.8%	5.43	<0.001	0.28
			medium pitch	31.7%	4.99	<0.001	0.85
			high pitch	24.1%	1.76	0.078	0.74
		Sadness		34.5%	11.2	<0.001	0.89
			low pitch	58%	16.96	<0.001	0.64
			medium pitch	31%	4.93	<0.001	0.78
			high pitch	14.4%	-2.49	0.013	0.72
		Anger		53.9%	26.23	<0.001	0.95
			low pitch	49.8%	13.28	<0.001	0.97
			medium pitch	56.1%	16.12	<0.001	0.95
			high pitch	56.1%	16.12	<0.001	0.87

				14.5%	-3.62	<0.001	0.89
		Joy	low pitch	1.7%	-6.96	<0.001	0.61
			medium pitch	8.2%	-4.5	<0.001	-0.37
			high pitch	33.6%	5.19	<0.001	0.77
		Tenderness		13.4%	-4.380	<0.001	0.56
			low pitch	9.5%	-4	<0.001	0.60
			medium pitch	20.7%	0.26	0.793	0.32
			high pitch	9.9%	-3.84	<0.001	0.31
	Classical	Neutral		28.7%	5.76	<0.001	0.39
			low pitch	31.6%	4.41	<0.001	0.49
			medium pitch	31.5%	4.37	<0.001	0.46
			high pitch	23.3%	1.25	0.212	-0.34
		Sadness		36.2%	10.69	<0.001	0.86
			low pitch	53%	12.57	<0.001	0.86
			medium pitch	38.4%	6.99	<0.001	0.43
			high pitch	17.2%	-1.05	0.294	0.71
		Anger		26.7%	4.43	<0.001	0.95
			low pitch	15.6%	-1.68	0.093	0.96
			medium pitch	22.4%	0.92	0.358	0.97
			high pitch	42.2%	8.47	<0.001	0.87

				%	z-value	p-value	Cronbach's alpha
Male	CCM	Joy		12.6%	-1.72	0.086	0.73
			low pitch	0%	-2.69	0.007	
			medium pitch	13.8%	-0.84	0.403	
			high pitch	24.1%	0.56	0.577	
		Tenderness		31%	2.57	0.01	0.78
			low pitch	37.9%	2.41	0.016	
			medium pitch	41.4%	2.88	0.004	
			high pitch	13.8%	-0.84	0.403	
		Neutral		40.2%	4.75	<0.001	0.66
			low pitch	31%	1.49	0.137	
			medium pitch	31%	1.49	0.137	
			high pitch	58.6%	5.2	<0.001	
		Sadness		34.5%	3.38	<0.001	0.81
			low pitch	41.4%	2.88	0.004	
			medium pitch	48.3%	3.81	<0.001	
			high pitch	13.8%	-0.84	0.403	
		Anger		18.4%	-0.38	0.707	0.97
			low pitch	0%	-2.69	<0.001	
			medium pitch	0%	-2.69	0.007	
			high pitch	55.2%	4.74	<0.001	

				11.5%	-1.98	0.047	0.91
		Joy	low pitch	0%	-2.69	0.007	
			medium pitch	3.4%	-2.23	0.026	
			high pitch	31%	1.49	0.137	
		Tenderness		27.6%	1.770	0.077	0.79
			low pitch	10.3%	-1.3	0.194	
			medium pitch	41.4%	2.88	0.004	
			high pitch	31%	1.49	0.137	
	Classical	Neutral		36.8%	3.91	<0.001	-0.69
			low pitch	41.4%	2.88	0.004	
			medium pitch	34.5%	1.95	0.051	
			high pitch	34.5%	1.95	0.051	
		Sadness		50.6%	7.13	<0.001	0.76
			low pitch	69%	6.59	<0.001	
			medium pitch	48.3%	3.81	0	
			high pitch	34.5%	1.95	0.051	
		Anger		20.7%	0.16	0.872	-0.09
			low pitch	13.8%	-0.84	0.403	
			medium pitch	20.7%	0.09	0.926	
			high pitch	27.6%	1.02	0.307	

5.2 Experiment 2

The second experiment was designed to gain information about typical sound structures of emotional expression in singing. We investigated short vowel samples and compared Classical and CCM singing voice qualities.

There were no clear indications of emotion-related differences in the fine tuning of fundamental frequency, CCM: Greenhouse & Geisser correction $F(1.845,49.812) = 1.155$, $p = 0.32$, Classical: $F(4.84) = 1.983$, $p = 0.105$.

Pitch, however, had a strong effect on the overall sound pressure level, which increased as the fundamental frequency rose. According to our data, this tendency did not affect the use of SPL adjustment as a means for communicating emotion. Emotions of lower activity were sung with a lower SPL than emotions with higher activity. In the Classical singing style, the singers sang slightly louder and varied the loudness more than in the CCM style (Table 5). The SPL differed significantly between emotional expression for both CCM and Classical singers; CCM: $F(2.533,68.386) = 54.9$, $p = <0.001$, Classical: $F(2.732, 57.367) = 18.278$, $p \leq .001$.

In almost all of the cases in this study, the alpha ratio increased (smaller negative absolute value) when singing higher in pitch. The samples of high activity emotions (joy and anger) were characterized by a larger alpha ratio than the low activity emotions (sadness and tenderness). The alpha ratio was larger in the CCM samples than in the Classical ones (Table 5). According to the RM-ANOVA results, the alpha ratio was a significant differentiator of emotions for both CCM ($F(4.108) = 23.105$, $p \leq .001$) and Classical ($F(1.934,40.612) = 5.10$, $p = 0.011$) singers.

HNR increased with pitch. The Classical samples had a slightly higher HNR (suggesting a clearer sounding voice quality) in all emotions except in anger, in which the CCM samples had a larger HNR. The RM-ANOVA showed a statistically significant difference between emotions for both CCM ($F(3.098,83.638) = 0.844$, $p = 0.001$) and Classical ($F(4.84) = 2.944$, $p = 0.025$) samples.

There were a few distinctive formant structure patterns concerning emotion expression in our data, the formants packing tighter together in high activity emotions and scattering in low activity emotions. In expressions of sadness, the formants F1 and F2 adopted a lower position, and in anger formants F1-F3 were in a high position compared to other expressions of emotion. Vice versa, the upper formants F4-F5 adopted a relatively high position in sadness and a relatively low position in anger (Table 2, Study II). The RM-ANOVA analysis found F1 to be statistically significant in differentiating the emotional expression in both the CCM ($F(2.455,66.273) = 21.382$, $p \leq 0.001$) and Classical ($F(4.84) = 6.242$ $p \leq .001$)

samples. In the CCM samples, also F2 ($F(3.047,82.268) = 7.06$, $p \leq 0.001$) and F4 ($F(2.898,78.255) = 3.602$, $p = 0.018$) was found to be statistically significant in differentiating emotional expressions.

Tenderness and sadness contained more aperiodic variation of *fo* (jitter) than the rest of the emotions (Table 4). The *fo* vibrato was slightly slower for the Classical samples than the CCM ones. RM-ANOVA analysis found jitter to be a statistically significant difference between emotions for the CCM samples (jitter rap $F(2.531,68.343) = 5.208$, $p = 0.004$, Jitter ppq $F(2.589,69.898) = 4.781$, $p = 0.006$). No significant differences between emotions were found for vibrato in either of the genres.

The irregular variation of the period amplitude (shimmer) was larger in the low activity emotions (tenderness, sadness) and smaller in the high activity emotions (joy, anger) (Table 5). Shimmer yielded statistically significant results for the CCM samples (Shimmer apq3 $F(4.108) = 10.056$, $p \leq 0.001$, Shimmer apq5 $F(2.949,79.626) = 10.153$, $p < 0.001$) according to the RM-ANOVA. It also showed that the amplitude vibrato rate differed significantly between emotions for both the CCM ($F4.108) = 5.335$, $p = 0.001$) and Classical ($F(4.84) = 4.827$, $p = 0.001$) samples.

We wanted to investigate if the amplitude contour revealed some sort of pattern in these short vocal samples as well. The nominal durations for sustained sounds in "Summertime" according to the notation and selected tempo of the music piece were 2.25 sec for a (220 Hz), 4.5 sec for e1 (330 Hz), and 2.25 sec for a1 (440 Hz). In Table 5, we can see the tendency of the CCM singers to cut the notes somewhat shorter in comparison to the Classical singers. We can also see that the higher the pitch, the shorter is the note value in comparison to the nominal value. The onset of the vowel sound is very similar in all of the samples, but the offset is slightly different depending on the emotion. The sustain time for CCM samples was found to differ significantly between emotions (Greenhouse & Geisser correction $F (2.430, 65.608) = 3.557$, $p = 0.026$).

Table 5. Mean values of parameters that distinguished significantly between emotions in the RM-ANOVA analysis.

		Mean Values																		
		SPL (dB)		HNR (dB)		F1 (Hz)		F2 (Hz)	F4 (Hz)	Alpha ratio (dB)		Jitter rap (%)	Jitter ppq5 (%)	Shimmer appq3 (dB)	Shimmer apq5 (dB)	f_o vibrato rate (Hz)	dB vibrato rate (dB)		Sustain time (sec)	
		CCM	Classical	CCM	Classical	CCM	Classical	CCM	CCM	CCM	Classical	CCM	CCM	CCM	CCM	CCM	CCM	Classical	CCM	Classical
Joy	low pitch	62	63	15	16	787	632	1453	3754	-13	-22	0.003	0.003	0.020	0.032	6	8	8	2.4	2.4
	medium pitch	70	69	22	22	839	693	1525	3717	-13	-18	0.002	0.002	0.016	0.018	6	7	7	3.4	3.5
	high pitch	73	77	22	24	916	774	1482	3674	-12	-17	0.002	0.002	0.014	0.016	6	6	7	1.8	1.9
Tenderness	low pitch	58	59	14	14	733	599	1433	3887	-16	-22	0.003	0.003	0.028	0.041	7	9	9	2.7	2.4
	medium pitch	65	67	21	22	753	644	1450	3854	-17	-22	0.002	0.002	0.015	0.020	6	7	7	3.6	3.8
	high pitch	66	73	22	25	779	744	1385	3683	-18	-21	0.002	0.002	0.015	0.014	4	7	6	1.8	2.0

Neutral	low pitch	59	62	15	17	707	641	1391	3809	-18	-23	0.003	0.003	0.022	0.036	5	8	8	2.4	2.4
	medium pitch	66	68	22	23	757	659	1431	3791	-17	-22	0.002	0.002	0.014	0.017	6	7	6	3.3	3.3
	high pitch	70	74	24	25	825	765	1419	3669	-17	-20	0.002	0.002	0.018	0.020	6	6	6	1.8	2.0
Sadness	low pitch	59	60	14	15	714	584	1409	3894	-19	-23	0.004	0.004	0.026	0.045	6	9	8	2.5	2.4
	medium pitch	67	68	22	24	719	663	1412	3709	-18	-23	0.002	0.002	0.014	0.017	6	7	7	3.4	3.8
	high pitch	69	73	23	25	773	716	1362	3646	-18	-20	0.002	0.002	0.015	0.018	6	7	7	1.7	1.9
Anger	low pitch	70	64	17	17	809	654	1351	3694	-15	-24	0.002	0.002	0.015	0.023	4	7	7	2.3	2.4
	medium pitch	75	72	23	21	873	750	1454	3617	-14	-18	0.001	0.001	0.011	0.014	5	6	6	2.9	3.2
	high pitch	77	78	24	24	1003	791	1533	3650	-11	-16	0.002	0.002	0.008	0.010	5	6	6	1.5	1.9

Repeated measures ANOVA found a significant effect of emotion for 11 of the 20 parameters measured for the CCM samples and four of the parameters measured for the Classical samples. The common statistically significant parameters found in both genres were SPL, HNR, alpha ratio, and F1. When the effects of *fo* and SPL were taken into account in the univariate analysis, significance remained for CCM in alpha ratio, sustain time, shimmer, F1, F2, and amplitude vibrato rate. For the Classical samples, none of the parameters remained significant differentiators between emotions after *fo* and SPL had been set as co-variates (Table 5, Study II).

5.3 Experiment 3

This experiment investigated the effects of a seven-week training regime using the parameter manipulation technique in vocal emotional expression. We did listening tests to validate the efficiency of the method and acoustic analysis to gain more information about emotion-specific voice quality. We derived the analysis from short vowels and phrases for the listening tests and from short vowel samples for the acoustic analyses.

5.3.1 Recognition of emotions

In this study, we were mainly interested in the short vowel samples, as they are seen as a carrier of information about the voice quality and as such are reflective of the usefulness of the practice regime used in the teaching intervention. Our hypothesis was that by teaching different voice qualities, we could improve the recognition of emotion from the singing voice.

We ran unrelated t-tests to determine if the recognition differed between the test group samples and control samples. There were more correct recognitions of intended emotions in the test group in both the before (M = 15, SD = 10) and after (M = 17, SD = 10) conditions. For the control group, the correct recognition of intended emotions was (M = 14, SD = 8) in the before condition and (M = 11, SD = 8) in the after condition (Table 3, Study III). The mean difference in the correct answers given in response to the heard samples was 1.50 (95% CI, -3.28 to 6.28) higher in the test group in the before condition in comparison to the control group,

and 5.97 (95% CI, 1.31 to 10.62) higher in the test group in the after condition in comparison to the control group. There was a statistically significant difference in mean correct recognition of emotion between the test group samples and the control group samples after the teaching intervention: $t(58) = 2.565$, $p = 0.013$.

The results indicate that for the test group samples, recognition of emotion increased in all emotion portrayals in the after condition. The recognition of neutral samples decreased in the after condition. For the control group samples, the recognition of emotion decreased for the after condition in all other emotion portrayals except joy. The recognition of neutral also increased.

We ran the Pearson's Chi-squared test to see if there was a statistically significant difference between the answers given for samples recorded before and after the seven-week training period. The Pearson's Chi-squared test showed a significant difference in the answers given for the neutral and anger portrayals in the test group. The recognition of anger increased by 28.4% units from before to after training. The recognition of neutral decreased by 20.6% units.

Recognition from phrases seemed to be easier than recognition from vowel samples, but there were no statistically significant differences in recognition between the test group and the control group (Tables 5 & 6, Study III).

The results indicate that for the test group samples, recognition of emotion from phrases increased in all emotion portrayals in the after condition. The recognition of neutral samples decreased in the after condition. For the control group samples, the recognition of emotion increased for the after condition in tenderness and neutral and decreased in joy and sadness. The recognition of anger was similar in both conditions (Table 6, Study III).

5.3.2 Acoustic analysis results

A Friedman test was run to determine if there were differences in the usage of different sound parameters in different emotion expressions. Pairwise comparisons were performed (SPSS, 2019) with a Bonferroni correction for multiple comparisons. The acoustic parameters F1, SPL, and HNR differed significantly between expressions of the emotions in both the before and after conditions (Tables 7 & 8, Study III). The alpha ratio was found to differ statistically significantly between the emotions in the after condition for the test group.

Post hoc analysis of the test group parameter values of the before condition revealed statistically significant differences **in F1** between sadness (Mdn = 620 Hz)

and joy (Mdn = 801 Hz) (p = 0.010) and between sadness and anger (Mdn = 807 Hz) (p = 0.001); **for SPL** between tenderness (Mdn = 76 dB) and anger (Mdn = 86 dB) (p = 0.035), between tenderness and joy (Mdn = 86 dB) (p = 0.019), and between sadness (Mdn = 75 dB) and joy (p = 0.035); and **for HNR** between sadness (Mdn = 17 dB) and joy (Mdn = 21 dB) (p = 0.019)

In the after condition, post hoc analysis showed statistically significant differences **in F1** between tenderness (Mdn = 658 Hz) and anger (Mdn = 852 Hz) (p = 0.019), between sadness (Mdn = 658 Hz) and anger (p = 0.019), and between neutral (Mdn = 658 Hz) and anger (p = 0.019); **in SPL** between sadness (Mdn = 73 dB) and joy (Mdn = 82 dB) (p = 0.019) and between sadness and anger (Mdn = 87 dB) (p = 0.001); and **in HNR** between sadness (Mdn = 17 dB) and joy (Mdn = 21 dB) (p = 0.019) in the test group.

In addition to these, in the after condition post hoc tests revealed statistically significant differences **in the alpha ratio** between neutral (Mdn = -27 dB) and joy (Mdn = -20 dB) (p = 0.035) and between tenderness (Mdn = -27 dB) and joy (p = 0.035) in the test group.

In the test group, SPL increased on the continuum of tenderness → sadness → neutral → joy → anger. Low intensity emotions (sadness, tenderness) were sung with a lower volume than the emotions with higher activity. The effect of the exercise routine can be seen in the wider variance in SPL control in the after condition as opposed to the before condition in the test group. In comparison to the control group, the test group showed a more consistent SPL between the recordings. In the control group, singers sang slightly louder the first time and softer the last time (Figure 1, Study III). SPL differed significantly between emotional expressions for both groups.

All of the samples for the high activity emotions (joy and anger) were characterized by a larger alpha ratio than the low activity emotions (sadness and tenderness). The effects of the training routine can be seen again in the wider variety of alpha ratio usage in the test group when comparing the before to after samples. The difference in the test group after samples in comparison to the before samples indicates that the test group started to vary their sound balance more after the teaching intervention. In this study, HNR did show a statistically significant difference between the emotions in the test group in both the before and after conditions. The HNR decreased in the second recording for the test group, suggesting the use of more noise components in the signal. The most harmonic content was found in joy and the least in sadness. For the control group, HNR increased for the second recording, suggesting a more sonorous sound. The effects

of the training routine was again observed in the increased variety in the use of the HNR component in the test group.

The overall formant structure of the samples revealed a few distinctive patterns in regard to emotion expression. In sadness, formant F1 was in a low position and F2, F3, and F4 in a relatively high position in comparison to other emotions and neutral, suggesting a more diffuse formant structure. In anger, the opposite was true: F1 was in a high position and F2, F3, and F4 packed lower in both groups. A similar but less pronounced formant structure could be found in joy. In tenderness, the first formant was positioned slightly higher than in sadness, but still relatively low; other formants were positioned fairly high. In neutral expression, the first formant was neither high nor low, while the second and the third formants were in a relatively low position. The formant structure was most compact in anger, and then it scattered in a sequence of anger → joy → neutral → tenderness → sadness. F1 was found to be statistically significant in differentiating the emotional expressions in both groups.

6 DISCUSSION

This dissertation investigates the possibility of teaching acoustic emotion expression to singers using the parameter modulation technique. The technique uses acoustic research-based elements to build emotionally expressive components to the singing voice. These elements include acoustic parameters, such as sound pressure level, formant frequencies 1-5, alpha ratio, HNR, and different types of vibrato and/or perturbation. These parameters can be heard in the voice as loudness, timbre, clarity of sound, and fluctuation of *fo*.

We first investigated if it was indeed possible to recognize emotion expression in the singing voice (Study I). We used CCM and Classical singing techniques in this study, as they are known to differ from each other in vocal technique. We also used three different pitches, as previous research has indicated that different emotion expressions tend to happen at different pitches in "real life" in speech, where the pitch is not pre-determined as in singing songs (Murray & Arnott, 1993) and thus the perception of emotion might be effected by the pitch. Pitch may also affect the ability to realize timbral changes or it may impose such requirements of the singing technique that restrict the free variability of voice quality. Secondly, we looked at the acoustic parameters of the intended emotional expressions of Classical and CCM singers at different pitches to determine if there was any specific pattern of acoustic cues in different emotion expressions (Study II). Finally, on the bases of Studies I and II and on previous studies conducted on this research area, we investigated whether it was possible to teach these parameter modulations to acting students. The aim of the teaching intervention was to enhance the student's capability of acoustic emotion expression in singing (Study III). The results of this study give support to our hypothesis that including research-based elements of the acoustic characteristics of sound to regular voice training will enhance the communicative power of the singing voice. This information and the technique model have direct applicability to voice training with regards to both the singing and speaking voice. As the model seems to make the communication of emotion clearer, the technique could have applicability also in different kinds of therapeutic, care, and education settings.

6.1 The vowel /a/

The value of using the simple /a/ vowel as an object of study may be questioned, as there is naturally so much more to emotional singing. However, the focus of the present study is the extent to which the expression of emotion is mediated by voice quality. Examining a steady vowel reduces the variables needed to be accounted for in the analysis. From an acoustic point of view, the transmission properties of the vocal tract are due to the cross-sectional area function of the tract (Fant, 1970). Acoustic impedance is inversely proportional to the area (Titze & Verdolini Abbot, 2012). The soundwave travels in the tract without modification as long as the area remains constant, but if it changes, as it does when articulating, a part of the wave is reflected back and the rest propagates forward. The simplest tract shape is where the cross-sectional area in the glottis-lip dimension is constant (Pulkki & Karjalainen, 2015). I use the vowel [a:] for ease of measurement, on one hand, and for focusing on the sound quality, on the other. I believe that the simple and neutral vocal tract setting allows the most variation in voice quality to emerge. This assumption is backed by Waaramaa et al. (2008), who found that the open back vowel [a:] conveyed the expression of anger better than the vowels [i:] and [u:] (Waaramaa, Laukkanen, Alku, & Väyrynen, 2008). Furthermore, the first formant frequency in the open back vowel [a:] is quite high, so it can more easily be differentiated from the fundamental frequency in formant analyses.

Longer utterances and melody lines contain a lot of additional information, which is likely to be more strongly affected by musical tastes and cultural effects. Musical features such as tempo, melody, and the overall structure of the music piece all affect emotion appraisals. In the future, it would be interesting to investigate the correlational effects of voice quality and other musical features.

6.2 Recognizing emotion in the singing voice

Previous research suggests that the expression of emotions in the singing and speaking voice are related and that the same methods of emotion recognition apply to both (Eyben et al., 2015; Scherer et al., 2015). The percentage of correctly recognized emotion samples in Study I was relatively low (30.2%) compared to earlier studies concerning speech. Our second listening test (Study III) yielded a slightly better recognition of emotion expression from the short vowel samples (42.6%) and even better overall recognition from phrases (52.7%). Most of the

studies examining emotion recognition from the speaking voice reach recognition percentages ≥ 50% (Banse & Scherer, 1996; Iida, Campbell, Higuchi, & Yasumura, 2003; Scherer, 2003; Scherer, Banse, & Wallbott, 2001; Waaramaa, Laukkanen, Alku, & Väyrynen, 2008). Thus, according to our results, it seems to be harder to recognize emotions from the singing voice, at least when the samples are short. The increased recognition percentage in phrases is probably due to increased information that exceeds the quality aspect of the voice. In other words, the longer melody line and the added consonants that we had in our phrase samples carried more information about the state of the voice instrument and thus allowed for a more precise evaluation concerning expression.

Concerning our main study question of whether or not it is possible to train emotion expression in voice quality, the results indicated that the recognition of expressed emotion increased and recognition of neutral expression decreased after the teaching intervention. The "neutral" answer is the most likely selected when one does not have a clear indication of emotion expression, therefore the decrease in neutral answers can be interpreted as emotion expression becoming more precise. In the control group, the recognition of sadness, anger, and tenderness decreased and the recognition of joy and neutral increased. There was not a statistically significant difference in mean correct recognition of emotion between the test group samples and the control group samples before the teaching intervention, but after it, a significant difference was found (Study III). This result can be seen as indicative of training emotion expressions through changes in voice quality having a clarifying effect on expression.

6.2.1 Differences between recognition in the CCM and Classical singing styles

In our study, emotion expression samples sung using the CCM technique were correctly recognized 11% more often than samples sung using the Classical singing technique. It is known that there is a considerable variability in the individual ability to express emotion by singing. Listeners also report some voices as being generally more expressive than others (Siegwart & Scherer, 1995). There is no comprehensive explanation to date as to why this is so, but speculations have been made about idiosyncratic interpretations of different emotions, a greater affinity with certain emotion expressions, a restricted range of expressivity, or even limitations due to the singer's vocal range or voice type. Another possible explanation is the listener's preconceptions concerning the affective nature of certain voice types (Scherer,

Trznadel, et al., 2017). This postulation could offer support to our findings, as the CCM style singing might feel more familiar to the listener. In general, most people are exposed to far more CCM than Classical singing. Therefore, they may be more attuned to emotion expression in these genres.

The speech-like qualities (such as the creak-like sounds at the end of tone and the amount of noise in the signal) of the CCM style of singing might explain why it seems to be easier to recognize emotion portrayals from it. Another possibility as to why the CCM style of singing is more recognizable could be that it uses the "chest voice" (or the modal register) more often, whereas Classical singing operates more with the "head voice" (falsetto). In the chest voice, the mass of the vocal folds vibrates more vertically, making a more robust impact on air pressure, and the formants appear easier (Titze & Martin, 1998). Henrich et al. (2014) found that in the head voice, the glottal contact quotients are usually lower and the vocal tract resonances lower (f_{R1} by 65Hz, f_{R2} by 90Hz) than in the chest voice sung at the same pitch using the same vowel (Henrich Bernardoni, Smith, & Wolfe, 2014). The slope of the sound spectrum is more gradual, as the relative amplitude of the upper partials is more pronounced in the chest voice. In the head voice, the slope is steeper (Sundberg, 1987). Regarding the expression of sadness in our study, we found that while all the other emotions were systematically better recognized from the CCM samples, sadness was better recognized from the samples using the Classical singing technique. The steeper spectral slope and lower resonance frequencies in the Classical style may explain why sadness was better recognized from the samples using the Classical singing technique. It is also plausible that speakers and CCM-style singers use more variation in phonation type along the axis from breathy to pressed (Peterson et al., 1994) while Classical singers keep the voice source more stable. This is related to both aesthetic and technical demands. Another possible explanation could be the use of vibrato in the Classical technique, which might correspond to the slow vibrato quality of sad music (Juslin & Laukka, 2004).

It was also possible to see in Study II that the alpha ratio was larger in the CCM style samples than in the Classical style samples (see Table 5). Hallqvist et al. (2016) found a stronger subglottal pressure and maximum flow declination rate in a soul singing style when compared to a musical theater style. However, the SPL was lower, which suggests a higher glottal resistance (Hallqvist, Lã, & Sundberg, 2017). Bourne et al. (2012, 2016) also found differences of technique within a specific singing style (musical theater) concerning phonation and resonant properties of the tract (Bourne & Garnier, 2012; Bourne et al., 2016). Classical singers typically have a louder and more omnidirectional sound radiation when singing the vowel [a:] in comparison to

CCM singers (Podzimková & Fric, 2019). They also have been found to have a consistently stronger sound radiation in the frequency bands 2 and 4 kHz (Podzimková & Fric, 2019). These findings lend further support to the notion that there are genre-typical ways of using the singing voice. The voice technique(s) can also change inside the umbrella genres of Classical and CCM. As our investigation indicated that the recognition of emotion might be genre-specific, more research is needed on how emotion is recognized between these sub-genres under the umbrella terms of Classical and CCM.

6.2.2 The effects of pitch on emotion recognition and expressive singing

The pitch also affects the recognition of emotion from the singing voice quality. In Study I, the low frequency samples were recognized less often than the high frequency samples. As in the speaking voice, the higher the pitch, the more often the listeners choose an emotion that represents a high overall activity level. This is understandable, since a higher pitch is typically produced with higher subglottic pressure and thus intensity (Sundberg, 1987; Titze, 1994). Fundamental frequency has been found to be a strong indicator of emotion expression in speech. Expressions of anger and joy tend to raise the *fo* of the speaking voice, while expressions of sadness and tenderness tend to lower it (Murray & Arnott, 1993).

In Study I, the tendency not to recognize joy but to recognize sadness was very pronounced at a low pitch (220/110 Hz) for both the female and male samples. At a high pitch (440 Hz), the phenomenon was reversed. A possible explanation relates to the functionality of the Classical singing technique at low pitches. The lowest [a:] was sung at 220 Hz, which is quite low for female singers accustomed to using only the Classical technique. A lack of control of the vocal technique at lower fundamental frequencies (singing in the head voice all the way down) could lead to a lack of power (low SPL), increase of noise, lack of control at the vocal fold level (very breathy phonation), and difficulties in controlling the vibrato, which all summed could lead to a perception of sadness.

The recognition of tenderness was slightly easier at the middle pitch in comparison to the low and the high pitch. Tenderness has the same activity level as sadness, and they share almost all of the acoustic parameter directions. Therefore, the slight increase in phonation frequency (from low to middle) might explain the increase in recognition. Earlier research on the speaking voice has indicated that it is the brighter voice quality in tenderness in comparison to sadness that sets them apart

from each other (Waaramaa, Alku, & Laukkanen, 2006). The higher pitch would need a little more P_{sub} and vocal fold adduction, which would make the formants more pronounced and thus would brighten the sound color. This could also be an explanation for the increased recognition.

Anger was most consistently recognized at a high fundamental frequency in the CCM and Classical samples. This is congruent with earlier findings stating that vocal expressions of anger are characterized by an increase both in SPL and *fo* (Laukkanen et al., 1996; Murray & Arnott, 1993; Williams & Stevens, 1972). The physical activity state in anger tends to be tense (Nummenmaa, 2019), and expressions of anger are typically characterized by a more hyperfunctional voice production with a high alpha ratio. As phonation at a relatively high pitch may induce stronger vocal effort and increase the alpha ratio, this may also result in the false recognition of anger in expressions of other emotions.

We originally thought that micro-tuning of the fundamental frequency might be a strategy used for coding emotions to the singing voice. In their study about the influence of vocal training and acting experience on voice quality and emotional genuineness, Livingstone et al. (2014) found that decreased pitch accuracy in sung emotion portrayals correlated with more years of acting experience and increased appraisals of genuineness in the listening tests. They postulated that the decreased pitch accuracy was a deliberate technique that the actors adapted for the emotion expression task (Livingstone et al., 2014). There was no emotion effect on the fine tuning of fundamental frequency in our study.

As expected, pitch had a strong effect on the overall SPL (Study II), but this effect did not affect the use of SPL adjustment as a tool for emotion expression according to the univariate analysis (Study II). Pitch also effected the note value in comparison to the nominal (note) value: the higher the pitch was, the shorter the note (Study II). This result most likely just reflects the increased technical difficulty in sustaining a pitch at a higher frequency, as there were no emotion-specific differences in the sustain time.

The cricothyroid muscle is primarily required for increasing *fo*, whereas the thyroarotenoid can increase or decrease *fo* and SPL. The lateral cricoarytenoid/interarytenoid activation is likely to maintain the vocal fold adduction during increased P_{sub}, and this is thought to improve vocal efficiency. The same target *fo* and SPL may be achieved with multiple combinations of intralaryngeal muscle activation (Chhetri & Park, 2016). According to the findings of Björkner et al. (2006) concerning female musical theatre singers, P_{sub} was slightly higher in the

chest than in the head register, and the MFDR[7] values were clearly greater in chest register. Furthermore, the glottal adduction was firmer in the chest register (Björkner, Sundberg, Cleveland, & Stone, 2006). At high frequencies, it is necessary to employ a different vocal strategy to keep the quality of the voice consistent. The typical vocal tract modification in trying to modulate the formant frequencies to keep the "ring" in the voice and increase loudness at high pitches is opening the jaw (Sundberg, 1975). In another study, Björkner (2008) found that both opera and musical theater singers varied their P_{sub} systematically, approximately doubling P_{sub} for a doubling of *fo*. However, the musical theatre singers had weaker fundamentals and higher formant frequencies, and they lacked the opera singers' characteristic clustering of F3-F5 (Björkner, 2008). Podzimkova and Fric (2019) found that groups of Classical and pop singers were differentiated by the maximal values of SPL and the amount of energy in the spectral bands of the first formant and the 3-4 kHz band where the Classical singers were louder than the pop singers. The pop singers, on the contrary, had more energy in the areas above 4 kHz (Podzimková & Fric, 2019). These "preconditions" of typical (and genre typical) voice use at higher pitches make it hard to express different types of vocal qualities at high pitches.

6.2.3 Valence and activation appraisals

Research papers have continuously recognized the difficulty of inferring valence from vocal expressions (Cunningham, Weinel, & Picking, 2018; Eyben et al., 2015; Scherer, Sundberg, et al., 2017; Scherer, Trznadel, et al., 2017). The valence dimension has been previously linked to formant structure and vocal tract settings in the speaking voice (Erickson, Shochi, Menezes, Kawahara, & Sakakibara, 2008; Laukkanen, Vilkman, Alku, & Oksanen, 1997; Mori & Kasuya, 2007; Waaramaa et al., 2006). In their recent study, Li et al. (2018) suggested that F1 positioning would contribute to the perception of valence (Li, Li, & Akagi, 2018). The arousal/activation axis has been previously linked to glottal characteristics such as utterance duration, intensity, *fo*, spectral flatness, and HNR (Juslin & Laukka, 2001; Juslin & Scherer, 2008; Laukkanen et al., 1997; Li et al., 2018; Scherer, 2003). Eyben et al. (2015) found in their study about automatic recognition of emotions in the singing voice that very few vocal features are required for arousal/activation classification, but for valence, adding more features generally improves the

[7] maximum flow declination rate

performance. Our findings support the notion of valence being hard to detect from the voice, as joy was the least recognized emotion in our investigations whereas anger was the most recognized one, and furthermore tenderness and sadness got mixed up easily.

In Study I, valence and activation were derived from the listeners' answers. Valence and activity were perceived with 41.6% and 45.8% accuracy, respectively, from the answers given. High activity was perceived with 41.5% accuracy and low activity with 57.5% accuracy. Positive valence was perceived with 38.6% accuracy and negative valence with 50.2% accuracy (Table 5, Study I). In our data, activation was more accurately perceived from all pitches in comparison to valence. The accuracy of perceived valence and activity in the listening test answers may suggest that it is easier to make assessments of valence and activity than to recognize emotions *per se*. This corresponds to the earlier findings for speech. Similarly, we found in the female samples that samples with a negative valence and high activity were more easily recognized than those with a positive valence and low activity. It is understandable, since it is important for survival to be able to quickly recognize signs of potentially dangerous situations.

6.3 Singing with an emotional voice quality

Study II indicated several differences between Classical and CCM styles in the use of voice quality as a carrier of emotional messages in singing. The repeated measures ANOVA showed a significant effect of emotion for 11 of the 20 parameters measured for the CCM samples and four of the parameters measured for the Classical samples. The common statistically significant parameters found in both genres were SPL, HNR, alpha ratio, and F1. Similar results have been found in previous studies concerning the singing voice. Scherer et al. found that loudness, vocal dynamics, and high low-frequency energy and formant amplitude, spectral balance, and low partial dominance were indicative of emotion expression/recognition (Scherer, Sundberg, et al., 2017; Scherer et al., 2015; Scherer, Trznadel, et al., 2017). In Study III, the test group vowel samples after the teaching intervention eventually yielded significance for SPL, HNR, alpha ratio, and F1 as well. For the control group, F1, SPL, and F4 showed statistical significance in differentiating the emotions in the samples collected in the former recording, but only F1 and SPL remained statistically significant in the samples collected in the latter recording. When comparing our two studies (Study II & Study III) SPL, HNR, alpha

ratio, and F1 were the common denominators with regards to sound parameters in emotion expression.

Eyben et al. (2015) recognized pitch, jitter/shimmer-based indicators, spectral band descriptors, and MFCCs[8] to be indicative in separating emotion expressions in an automated recognition test (original sets of 205-6373 parameters)(Eyben et al., 2015). Patel et al. (2011) in their investigation of emotion from sustained [a:] vowels, recognized phonatory effort, phonatory perturbation, and phonation frequency to be indicative of different emotion expressions. Perturbation and perturbation variation were also found to be indicative of emotion expression in singing (Scherer, Sundberg, et al., 2017; Scherer et al., 2015; Scherer, Trznadel, et al., 2017). Jitter, shimmer, and vibrato were not found to be statistically significant indicators of emotion expression in this study. However, jitter and shimmer were detected by the RM-ANOVA as statistically significant differentiators of emotion expression in CCM singing samples in Study II and in the test group samples recorded before the teaching intervention in Study III. In Study II, when factoring in the effects of *fo* and SPL, the effect vanished and in Study III the effect vanished after the teaching intervention. The sudden disappearance of jitter and shimmer may be attributed to the improvement of vocal technique in acting students as a result of increased skill in vocal control in Study III. In Study II, the CCM singers' tendency to use jitter and shimmer is probably due to the more lenient approach to the vocal technique.

In Study III, the number of parameters showing statistical significance were reduced after training. The result could be interpreted as showing that rehearsing the stereotypical use of voice parameters narrows down the choice of parameters. On the other hand, it may reflect the more focused and conscious use of certain key parameters and the reduction of parameters that may not be related to expression as such, or that may not have any relevance to emotion recognition. The highly reduced parameter set does not allow for a comprehensive explanation of acoustic emotion expression in the singing voice, however. The parameters seem to play more into the activity/arousal dimension of emotion than the valence dimension. The listening test results nevertheless indicate that it is possible to distinguish both dimensions from short samples reflecting voice quality. So, there is still something missing in the analysis of our data. Whether it is a missing parameter, some sort of threshold value,

[8] MFCC Mel-frequency cepstral coefficient quantifies the gross shape of the spectral envelope and removes the micro-level spectral structure, which is often less important for recognition. It converts the conventional Hz frequency to the Mel scale and is therefore compatible with human perception for sensitivity at appropriate frequencies.

or the exact proportions of different sound parameters that ultimately constitute the emotional voice remains to be investigated.

6.3.1 SPL variation in expressing emotions in singing

SPL is the most important parameter in emotional coding and decoding in the singing voice (Eyben et al., 2015; Scherer, Sundberg, et al., 2017; Scherer et al., 2015). It can be recognized as a very clear differentiator between emotions (Study II, Study III). High activity emotions (joy, anger) were characterized by a larger SPL than low activity emotions (sadness, tenderness) in all of our sample sets. The Classical singers sang slightly louder and varied the loudness more than the CCM singers (Study II). This result may reflect the overall tendency of Classical singers to sing loudly. Another possible explanation is that as the aesthetic demands of the Classical singing style are more rigid than those of CCM singers, so the Classical singers have fewer opportunities for variation in the other sound parameters and therefore have to maximize their opportunities for SPL variation.

The difficult part of acoustic analysis is that basically every measurement component is tied to SPL values, and it is very difficult to discern whether the parameter values change because of the emotion expression or simply because of volume change. When the effects of *fo* and SPL were taken into account in the univariate analysis in Study II, significance remained for CCM in alpha ratio, sustain time, shimmer, F1, F2, and amplitude vibrato rate. For the Classical samples, none of the parameters remained significant differentiators between emotions.

6.3.2 Alpha ratio in expressing emotions in singing

The alpha ratio varied across the emotion portrayals on an activity dimension – increasing in high activity emotions and reducing in low activity emotions (Study II, Table 3/Study III, Figure 3). The alpha measure provides information about the spectral slope declination, and our results indicate that in sadness the declination rate is steep. In tenderness and the neutral expression, the slope is less steep, and in joy and anger expressions, the slope is less pronounced.

The alpha ratio increased when singing at a higher pitch (Study II), which complies with previous studies using the alpha ratio measurement. Larger alpha values indicate a flatter spectrum (less difference between the energy of the lower and higher harmonics), which is often coupled with a louder voice (Guzman et al.,

2015; Nordenberg & Sundberg, 2004; Sundberg & Nordenberg, 2006). An increase in loudness affects the sound spectrum by amplifying frequencies between about 1500 and 3000 Hz more than the lower frequencies (Nordenberg & Sundberg, 2004). Lower alpha values indicate a steeper slope where lower harmonics dominate the spectrum.

In Study II, the alpha ratio values were consistently larger for the CCM samples, even though the Classical singers sang louder overall. One possible explanation for this result is that the alpha ratio values reflect a phonation style. Previous research has also coupled alpha ratio with type of phonation. Kitzing (1986) investigated the balance between high and low frequency partials in relation to phonation in healthy voices faking breathy, pressed, modal, and soft voice qualities. One of the measurements he found indicative of glottal function was alpha ratio, although it yielded significant results only in differentiating breathy and modal voices from each other (Kitzing, 1986). Bourne and Garnier (2012 & 2016) have investigated differences in legit and belt musical singing styles and found that the belt quality was more likely perceived when f_{R1}, SPL, alpha ratio, and CQ_{EGG}[9] increased. Conversely, the legit quality was perceived more often when they decreased. The increased contact quotient measures in belt are consistent with greater vocal effort and enhanced energy above 1 kHz. The belt qualities in their studies were always produced with higher alpha ratios compared to legit (Bourne & Garnier, 2012; Bourne et al., 2016). In light of these studies, the lower alpha values in the Classical samples in comparison to CCM samples in our study most likely reflect the stylistic demand for a more stable voice source quality in a Classical singing style.

6.3.3 HNR in expressing emotions in singing

HNR showed a statistically significant difference between the emotions when assessed by the RM-ANOVA and the Friedman test (Studies II & III).

The univariate analysis, however, did not yield statistically significant values for HNR independently, as *fo* and SPL both influenced its effect (Study II, Tables 4 and 5). Significant effects of *fo* and SPL on HNR have been found also in previous studies concerning the speaking voice. Higher SPL and *fo* are related to higher HNR (Brockmann-Bauser, Bohlender, & Mehta, 2018; Sampaio, Masson, Soares, Bohlender, & Brockmann-Bauser, 2020).

[9] Electroglottographic contact quotient

High activity emotions tend to have a larger HNR value than lower activity emotions. Warhurst et al. (2012) found that a perceptually clear voice was associated with high HNR values and lower jitter and shimmer, suggesting that acoustic noise and perturbation contribute to the perceptual assessment of vocal clarity. Lower HNR values have been found in imitated dysphonic voices indicative of a lower degree of harmonicity when mimicking a noisy voice quality (Schiller, Remacle, & Morsomme, 2020). In our study, anger was portrayed with a larger HNR than other emotions in Study II, but in Study III, the most harmonic content was found in joy and the least in sadness (Study II, Table 3/Study III, Figure 4). The singers in our study seemed to vary HNR systematically according to emotion, increasing it when expressing joy and anger and decreasing it when expressing sadness and tenderness.

In Study II, HNR seemed to vary strongly according to pitch. Low pitches were portrayed with a smaller HNR and high pitches with a larger HNR. The result is consistent with earlier findings on a higher *fo* being related to a higher HNR (Brockmann-Bauser et al., 2018; Sampaio et al., 2020). The Classical samples had a slightly higher HNR in all other emotions except anger, in which the CMM samples had a higher HNR. This might refer to the differences in aesthetics in these two genres, while in the Classical singing technique projection is one of the key elements, so it does not allow too much laryngeal compression. Then again, CCM techniques are a little more liberal regarding effort in voice production and it might be aesthetically more acceptable to "let it rip" in anger expressions. When the effects of *fo* and SPL were taken into account in the univariate analysis in Study II, HNR did not remain a significant differentiator of emotions.

6.3.4 Formants as a means of the expression of emotion in singing

In Studies II & III, F1 was found to be statistically significant in differentiating emotion expressions. In speech, F1 has been reported to be higher in expressions of joy and anger and lower in sadness (Waaramaa et al., 2008). Similar trends could be observed in our study as well; participants lifted the F1 for the high activity emotions and lowered it for low activity emotions (Study II, Table 3, Figure 1/Study III, Figure 5). Vos et al. (2018) found that at *fo* values below the first resonance, tuning f_{R2}, no tuning at all, or tuning both f_{R1} and f_{R2} to the closest harmonic increases vowel identification. Using only the f_{R1} tuning strategy (f_{R1} to *fo*), the vowel is misidentified more easily (Vos et al., 2017). Vurma (2020) found in his study of male operatic voices that the first resonance of the vocal tract was consistently higher in samples

sung louder, indicating a significant correlation with SPL and F1. However, there are also other possibilities: Koenig & Fuchs (2019) investigated how F1 and F2 differ between normal and loud speech in healthy female speakers. They found limited formant variation as a function of loudness in high, tense vowels. Loudness changes for F1 were systematic in direction but variable in extent. Their data indicated that loud speech in typical speakers does not always lead to changes in vowel formant frequencies (Koenig & Fuchs, 2019).

We know from previous research that opening the jaw raises the first resonance (Sundberg & Skoog, 1997), the second resonance is controlled mainly by the position of the tongue (Fant, 1970; Laver, 1980; Sundberg, 1987), and that shortening the vocal tract slightly by smiling raises all the resonance values (Tartter, 1980). We found that in expressions of sadness, F1 and F2 were at a low position and F4-F5 in a high position in both singing styles (Table 2, Study II). In Study III, F1 was in a low position in expressions of sadness, while F2-F4 were quite diffused and in a high position. In anger expressions, formants F1-F3 were in a high position and F4-F5 in a low position in both singing styles (Table 2, Study II). In Study III, F1 was in a high position in expressions of anger, and F2-F4 were packed tighter in a low position.

When reflecting on the information presented above, it seems plausible that modifying the position of F1 can also be an emotion expression strategy, as it can affect vowel intelligibility and spectral energy. The center of gravity in the voice spectrum has been found to be an indicative parameter of voice quality perception, as sounds with a higher center of gravity are perceived to be brighter and sharper (Vurma, 2020). Furthermore, the statistical significance of F1 as a differentiator of emotion expressions for CCM samples held in Study II after performing the univariate test for effects of the *fo* and SPL.

6.3.5 Vibrato and perturbation in expressing emotions in singing

Tenderness and sadness contained more jitter and shimmer than joy and anger. Jitter and shimmer were found to be statistically significant differentiators between the emotion expressions in CCM samples (Study II). Lower jitter and shimmer values have been found to be indicative of a perceptually clearer sounding voice (Warhurst, Madill, McCabe, Heard, & Yiu, 2012). Increasing the SPL decreases jitter and shimmer (Brockmann-Bauser et al., 2018). Higher activity emotions are therefore

inclined to sound clearer and to have less aperiodic variation of the *fo* and period amplitude.

There are similarities between vibrato and essential vocal tremor (Ramig & Shipp, 1987). The perceptual difference between perturbation and vibrato is that vibrato has a clearly distinguishable pitch variation (fluctuating around the target pitch), while jitter produces more of a noisy component to the voice. Physiologically speaking, shimmer and jitter are the result of 1) small asymmetries or variations in the cricothyroid muscle tension, 2) fluctuations in subglottal pressure, 3) perturbation of the vocal folds, or 4) a combination of these elements (van Puyvelde et al., 2018). The perceptual correlate is a rough, more or less hoarse voice quality. Turbulence noise may be increased by leaving a gap in the glottis and using sufficient subglottal pressure. The perceived voice contains a hissing component. Both perturbation and turbulence noise may contribute to a decrease in HNR (or the clarity of voice). Regular vibrato is an integral part of a healthy singing voice, but excessive use of vibrato can sound unacceptable. Too great an extent can leave the voice sounding wobbly, while too slow a vibrato rate – especially in a hammer type of vibrato – can leave audible gaps in the voice and too fast a vibrato rate can sound nervous.

In our study, the *fo* vibrato was slightly slower for the Classical samples than for the CCM samples. This result complies with previous research, as some vibrato characteristics (extent, regularity, and duration) have been found to very clearly differentiate the Western operatic singing style from popular singing styles (Manfredi et al., 2015). However, no significant effect was found on the *fo* vibrato in relation to emotion expression.

6.4 Teaching emotion expression using the parameter modulation technique

The purpose of the parameter modulation technique was to introduce basic acoustic characteristics (and their perceptual correlates) typically observed in the expressions of joy, tenderness, sadness, and anger to the student in a practical way (Hakanpää et al., 2019, 2021b, 2021a). The technique itself refers to the voluntary variation of these voice characteristics so that they result in a clearly recognizable emotional expression.

A similar system has been proposed by Coutinho, Scherer, and Dibben (2019) in the *Oxford Handbook of Singing*, where they postulated that the application of voice

science knowledge to singing practice could allow singers to practice the expression of certain emotions by systematically producing the configurations of voice parameters characteristic of particular expressions. They thought it was plausible to assume that the mechanical production of a vocal expression in singing could lead to physiologically measurable emotional experiences in the singer. Furthermore, they envisioned a real-time acoustic measurer of the singer's voice, which would provide feedback during practice or performance and allow the singer to fine-tune their technique during expressive singing (Coutinho, Scherer, & Dibben, 2019). My study confirmed that certain typical expressions of sung emotion can be trained and practiced in the vocal technique so that other people recognize the intended emotion more easily. The real-time acoustic measurer in my study was replaced by student-teacher interaction, but surely the use of new voice technology in teaching would yield new possibilities to perfect the work we do with voice quality.

The parameter modulation technique is a quite rigid, mechanical way of exercising the voice, and as such it should be used "with caution". The purpose of this technique is not to produce singing robots who mindlessly alter sound parameters in a scientifically pre-approved way, but rather to offer a tool for exploring one's own voice qualities with a template organized in this specific way.

As mentioned before, the singing voice is an instrument where the self and the voice intertwine, so the parameter modulation technique will never work completely free from cultural, emotional, and cognitive ties. This is true also for the usage of the technique in a teaching situation: what happens in the classroom will always be something more complex than merely the use of a template. That being said, I do believe that the approach can offer a way of delving into the world of emotions without going too deep into the students' emotional life.

Many practice trials provide multiple opportunities to consolidate relationships between the different types of information associated with each voice quality and the muscle movement that produces it. Multiple trials are thought to enhance the stability of recall and recognition schemas. Furthermore, it requires many instances of retrieval of the motor programs (essence of movement) to automatize the activation of generalized movement programs on future trials (Maas et al., 2008). In general, a large number of practice trials is beneficial for learning. However, if the movement pattern is done incorrectly, there is a risk that the wrong way will consolidate as a default. This is a matter of great concern in voice pedagogics in general but especially in teaching different voice qualities through the parameter modulation technique. The teachers' role in ensuring the safe execution of voice

quality changes is very important. This might advocate for a shorter than usual self-practice time for the student outside the classroom when using the method initially.

The exercise routine we used to train the parameters of emotional expression in the sound signal included exercises for volume control, phonation balance, articulation, and extreme vibrato. For volume control, we explored the whole dynamic range of the student, from the most quiet sound to the loudest. For phonation, we drilled polar opposite exercises ranging from a very breathy voice through optimal sound balance to pressed phonation. For resonance and articulation, we used exercises that shape the vocal tract in various ways. For perturbation we used extreme vibratos, both frequency (undulating between several semitones) and amplitude (crying-like volume changes) modulation, and even breaks in the voice to simulate perturbation in a voice-friendly way.

As explained in the theory part, the five bodily systems needed for singing are the nervous, breathing, phonation, resonance, and articulatory systems (Seikel et al., 2014). Already for the perceptually seemingly easy task of volume control, the students need to control these systems simultaneously. It was our pedagogical challenge to adjust the exercises of the "parameter modulation" model to correspond with the students' own ways of using and coordinating these systems. Our solution to this problem was to use the most simple instruction of moving clearly definable anatomical organs on the vertical, horizontal, and anterior/posterior axis. Previous research has indicated that singers can respond to kinematic directives and maintain behavioral changes throughout a song (Collyer, Kenny, & Archer, 2009). When kinematic direction was not possible, we used vocal sounds that could be easily segregated, demonstrated, and mimicked using the auditory feedback loop and/or visual information (Guadagnoli & Lee, 2004; Seikel et al., 2014; Tourville & Guenther, 2011). Previous research has also suggested that conscious attention to the mechanics of a motor task, such as a singing technique, can affect learning and performance negatively. Attention to task outcomes, on the other hand, can benefit performance and learning (Titze & Verdolini Abbot, 2012; Wulf et al., 1999). In using this dual approach of kinematic instruction and auditory modeling in the singing technique, we felt that the possibilities of misunderstandings due to a lack of conceptual knowledge or terminology can be avoided and the idea of the "parameter modulation" technique can be conveyed impartially to all participants.

The quickest way of testing whether this type of model works in real life is to prototype it. The aim of our experiment was to see whether the specific training improves the recognition of emotions from the singing voice and whether the acoustic differences between emotional expressions increase after the particular

training. We hypothesized that the recognition of emotions would increase in the test group and not change in the control group, and that the number of significantly differentiating parameters and the range of the parameters would increase after training.

The effects of the training routine could be seen in a wider variance in SPL control, alpha ratio usage, and HNR use in the after condition as opposed to the before condition in the test group (Figures 2, 3, & 4, Study III).

Study III's results suggest that training with the parameter modulation technique did increase the correct recognition of emotional expression from the short vowel and phrase samples. Our results show that for the test group samples, recognition of emotion increased in all emotion portrayals in the after condition. The recognition of neutral samples decreased in the after condition. It is fairly common to get a lot of "neutral" answers with the type of forced choice questionnaires that we used for the listening test. If the listeners are not sure what they heard, they are more likely to select "neutral" (Marcel & Eerola, 2012). The decrease in neutral answers in the test group after condition samples can be therefore interpreted as an increase in the expressivity of the singing voice. For the control group, vowel sample recognition of emotion decreased for the after condition in all other emotion portrayals except joy. The recognition of neutral, on the other hand, increased. This can be interpreted as a difficulty in arriving at a specific emotion appraisal on the listener's side. This lends support to our hypothesis that the systemic parameter manipulation of the singing voice can help in building expressivity in contrast to a situation where the students receive just regular voice tuition without rehearsing the use of voice quality in emotional expression.

6.5 General shortcomings of this study and suggestions for future studies

This study was limited by the small number of participants, and therefore the potential effect of individual factors cannot be excluded. In Studies I & II, we had 13 portfolio-type singers who performed at regional or local venues and 29 listeners. In Study III, we had 12 singing acting students and 32 listeners. As singers and singing styles and listeners and listening styles are many and various, it is important to keep in mind that the results obtained from this study are (in the strictest sense) valid only in the context of the data gathered for this study. More investigation is needed to further validate the usefulness of the parameter modulation technique.

In the listening tests, the listeners used their own devices for listening. The choice of listening device (e.g., headphones or a loudspeaker), whether they listened on their way home from work or while cooking supper for their children – these are all conditions out of the investigator's control. Nevertheless, we wanted to have a broader listener base (to not use only students) and thus the web-based listening test that one can take at home is ideal. We instructed the participants to use headphones for listening, and we further facilitated the listening experience by allowing the participants to quit and return to the test. This allowed the participants the opportunity to listen to the test at the most convenient time for them.

Both of our listening tests contained a lot of samples. In Study III, there were 300 samples, and in Study II, there were 246 samples to go through. Although it was possible to stop and continue listening on subsequent days, it is likely that at least some of the participants went through all the samples at once. It is likely that if this was the case, the samples listened to at the beginning of the sample-battery would have received more attention compared to the samples at the end of the sample-battery. We tried to tackle this problem by randomizing the sample-battery so that different samples would appear at the beginning of the battery for different listeners. The randomization was done through automatization with the idea that at least most of the time the Classical-CCM, male-female, and different pitches would be mixed in Study I, and the male-female and before-after expressions would be mixed in Study III. It is likely that the listener would start to form a ratio scale of perceived expressions and start comparing samples to each other, which might lead to misinterpretations (e.g., "This sample sounds kind of angry, but not really angry compared to the previous."). It is also likely that the listener would start to recognize individual voices and scale their appraisals accordingly (e.g., "This is the singer who sounds so sweet. I think it sounded angry in comparison to the sample this singer performed before."). It is also possible that the recognition of a singer would lead to attentional drift (e.g., "Ummm, this is the one who used a raspy voice before…. wait, what was I doing? ...I think it sounded like tenderness."). Of course, the listener's habituation to detect emotional undertone from vocal expression will also affect the appraisals, and in the case of detecting emotion from Classical vs. CCM styles of singing, it will inevitably be dependent upon what kind of music the listener is used to listening to. We did check the listening preferences of our listeners in Study I, but unfortunately the experiment was not initially designed with this particular question in mind, so we ended up with two people liking Classical music and 27 liking popular music with no mentionable differences in their recognition of emotion expression, so we did not report that. In future studies, it might be interesting to look at listener

preferences in music styles in relation to recognizing emotion expression from different vocal styles.

One possible factor pertaining to the results of this study can be the level of expertise of the sample givers. The bulk of the research conducted in this area has been done with professional world-renowned opera singers. There have been efforts to categorize singers on the basis of number of performances and size of the fanbase (Bunch & Chapman, 2000), but especially in small countries such as Finland, it is sometimes too hard to find enough singers to meet the demands of such categorizations. A lot of working singers in small countries use multiple singing styles, teach, and do some other extra work to make a living, so finding enough participants that fit the categorizations and do not overlap is simply impossible. Therefore, it is important that we also start to accumulate knowledge about "regular" singers and think about the ways we can standardize the methodology to allow for valid research also in situations where the level of singers varies.

The choice of keeping the pitches constant for both of our sample sets was made because we wanted to standardize the individual voice use as much as possible to make the analyses of our data more straightforward. The problem with standardizing the pitch is that it does not accurately reflect real-life conditions. For most music styles, it is quite possible to transpose the music to a key that is the best possible match for the singer. For the styles that use strict scores, the vocal lines are usually written with a certain voice type in mind, and therefore when they are performed (by the possessors of these voice types), the singing sounds effortless and right in the comfort zone – which it is – as it is written there. One possibility would be to try to standardize the target pitch individually by measuring each candidates' total pitch range and pinpointing the target pitch as a function of the total pitch range.

Sundberg et al. (2021) separates between the glottal function and tract function by dividing the parameters to those that can be measured from the LTAS (long-term-average spectrum) and those that can be measured with FLOGG (flow glottogram)(Sundberg et al., 2021). Perhaps in our study, the lack of clear separation of the glottal and tract parameter measurement technique-wise affected the results. Adding the glottal measurements to future research on this subject might give a more precise look at the parameters that in this study were insignificant after the effects of *fo* and SPL were accounted for in the statistical reduction.

The teaching intervention in this study was performed by the author due to resource-related restrictions. In future studies, it would be beneficial for the validity of the results obtained by using the parameter modulation technique if there were multiple teachers conducting the intervention. A neutral way of ensuring the validity

of this method would be to have a scientist designing it and a teacher teaching it. It could be said that practical tools – albeit theory based – should be developed in practical conditions by practitioners for practitioners. Therefore, I think it is beneficial in this kind of research that the scientist doubles as a teacher. However, to further ensure the validity of the method, it should be replicable also when taught by someone not involved in its development.

7 EPILOGUE

All teaching is a mix of science and art: Art in a sense that teachers are constantly faced with numerous changing variables, which require fast judgement and decision-making (improvisation). Science in a sense that skillful teaching is based on theory and research, which in turn is driven by epistemological differences about the nature of knowledge, and by different value systems. (Bates, 2015.) Teaching music has traditionally leaned towards the *art* of teaching and many teachers of music still shy away from research-based teaching methods (see Healey, 2005; Jenkins & Healey, 2012) as being too sterile and not in touch with the reality of learning to play music. It is fairly common for teachers to differ over what constitutes good teaching, depending on their epistemological standpoint, their priorities in terms of desirable learning outcomes and what they think matters most in learning (Bates, 2015).

Bloom (1979) has categorized the way we learn in his famous taxonomy. He states that we need to first gather up, memorize, and understand a bank of information before we can start applying and analyzing it, coming up with critiques and assessments, and creating something of our own (Bloom, 1979). The bulk of the work that a voice teacher does in a voice lesson is straight forward behaviorism. We repeat and reward (enforce operant conditioning). The central idea of behaviorism is that certain behavioral responses become associated (in a mechanistic and invariant way) with specific stimuli. The connection between a stimulus and a response will depend on reinforcement happening at the time of association. This depends on random behavior (trial and error) being appropriately reinforced as it occurs. Furthermore, we can withdraw reinforcement if we wish to extinguish inappropriate behavior. (Skinner, 1968; Bates, 2015.) This is basically teaching singing 101; you say "good" when the student sings correctly, you fathom a blank expression when they don't sound good.

In music education participation in a community of practice, with its social networks, roles and relationships is of utmost importance and in that way a lot of the work done in the studios fall under the theories of situated learning (Bloch, Lave, & Wenger, 1994). Another key concept in a modern voice studio is the humanistic approach. The approach emphasizes personal freedom, choice, and the validity of subjective experience of the student. The cognitive and emotional dimensions of

learning are given equal weight and the teacher is regarded as a facilitator of learning (Rogers & Freiberg, 1994). The Feather Newton value expectation theory is a good guiding principle to pupil-oriented teaching. It says that in order for the student to engage with their studies and in order for deep learning to occur the student must gain something useful (tangible) from the education and that the students must be able to perform the tasks they are given (Feather & Newton, 1982).

The critical understanding of teaching sees teaching as a political activity and promotes emancipation (Skelton, 2004) Teaching from a critical perspective means asking questions relating to authority and control over what counts as knowledge, how knowledge is organized and transmitted, who has access to knowledge and whose interests are served by the current system. The teachers aim, from the critical perspective, is to support a process of student emancipation. The role of the teacher is to act as a critical or transformative intellectual who disturbs the student's current epistemological understandings and interpretations of reality by offering new insights. (Skelton, 2004; Tennant et al., 2010.) Research based teaching and learning could be one way of promoting critical understanding of teaching (Jenkins & Healey, 2012). This is what my research efforts aim at.

My work seems to suggest an objectivist approach to teaching and learning. The whole methodology is based on presenting a body of knowledge to be learned, and the tool that I have created relies on the effective transmission of this body of knowledge. I want my teaching to be informative, organized, and clear, and I take pride in anchoring it to the vast knowledge base created by other thinkers. For my students, this means that I expect them to accurately comprehend the information, and I hope that they in turn add to the knowledge following the standards of empirical testing or some other form of research. (This approach to teaching can be described as the traditional liberal tradition of education (Tennant et al., 2010)).

As a practicing voice teacher, however, I have seen that the objectivist, traditional, one-way approach to teaching that relies on the students' strong inner motivation seldom works. Students have such different motivations to practice music and train their voice that lulling oneself into thinking that everyone will enjoy my research-led teaching style is the equivalent of digging a hole and falling into it. The amount that the students will engage with the theory part of the parameter modulation technique will inevitably be dependent on their skill level regarding the conceptual understanding of acoustics and the physiology of voice use.

The parameter modulation technique is in its essence a very simple model. It gives a taxonomy of acoustic parameters related to four different emotional voice qualities and suggests anatomically and physiologically justified actions to perform these

"acoustic emotions." At this level, it enables simple task allocation and fast formative assessment during training, which complies well with behaviorist ideas. As the developed model for parameter modulation is quite loose, it allows for a holistic approach of teaching as well. It is quite possible to personalize the technique by giving singer-specific instructions as to how to achieve the sound qualities under investigation. This way of using the parameter modulation technique is in fact the only possible (responsible) way of successfully working with this tool as the individualized exercises render the safe use of different sound qualities possible. Breaking down the sound parameters of a complex sound is a textbook example of how to use Bloom's taxonomy in teaching the singing voice. We give students information about the singing voice one conceptual area at a time: information is provided combining acoustic, tactile, and perceptual information, and the students combine the information, forming new ways of using their voices. The students gain instant value, as they are now equipped to use their voices in many different tone colors, and the students will be able to perform the tasks given to them, as it is possible to present the tasks at their own level. This ticks the box for value-expectancy theory and should motivate the students to further investigate the qualities of their sound. The critical understanding of teaching referring to the use of the parameter modulation technique comes from two directions: 1) the use of a scientific method in validating singing instruction brings to light the importance of open discussion, peer review, and the possibilities of inter-disciplinary work in arts and sciences; and 2) it questions the aesthetics of the teacher (and/or tradition), giving students agency over their own voice. The added benefit of the parameter modulation technique is that it really focuses on the expressive ***voice***. If it is not in the students' best interest to work on their interpretation of different emotions in singing by remembering or analyzing actual emotions, the parameter modulation technique provides a buffer that can alleviate the anxieties related to emotion regulation.

When integrating the parameter modulation technique into a curriculum, it is advisable to use constructive alignment to investigate its functionality. The parameter modulation technique provides many possibilities of aligning assessment with study objectives. With the help of acoustic analyses or real-time visual feedback technologies (see e.g. Welch, Howard, Himonides, & Brereton, 2005) it is possible to measure the extent of parameter modulations at the beginning and at the end of a teaching period. This could offer a different, more objective, route to evaluating development of expressivity in singing.

In our study, the test groups' understanding of voice quality, the terminology revolving around it, and the way they were able to perform tasks requiring changes in voice quality improved. I think this was mostly due to the parameter reduction: when working with one voice quality element at a time we (teacher + student) were able to form a common understanding of what that element means. This allowed the students to tackle the acoustic and physiological complexity of voice quality one task at a time, instead of trying to control all of the elements involved in it at once.

It would be interesting to see if the use of the parameter modulation technique would also aid conceptual learning in students of other instruments. If it would work, it could bring a new way of teaching musical expression to those students who are not naturally expressive in their playing. It would provide long awaited answers to those students who when asked to portray the feeling of the song through their instrument, answer:

"I can't picture myself in this situation, I have no imagination,"
"so what exactly do you want me to do?"
or simply,
"I don't know how."

8 REFERENCES

Alderson, R. (1979). *Complete handbook of voice training*. New York: Parker Publishing company.

Alipour, F., Scherer, R. C., & Finnegan, E. (1997). Pressure-flow relationships during phonation as a function of adduction. *Journal of Voice*, *11*(2), 187–194.

ANSI. (2020). ANSI/ASA S1.1 & S3.20 Standard Acoustical & Bioacoustical Terminology Database. Retrieved from https://asastandards.org/asa-standard-term-database/

Arnold, M. B. (1960a). *Emotion and personality. Volume I:Psychological aspects*. New York: Columbia University Press.

Arnold, M. B. (1960b). *Emotion and personality. Volume II: Neurological and physiological aspects*. New York: Columbia University Press.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *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), p.e0136675-e0136675. https://doi.org/10.1371/journal.pone.0136675

Barrichelo, V. M. O., Heuer, R. J., Dean, C. M., & Sataloff, R. T. (2001). Comparison of singer's formant, speaker's ring, and LTA spectrum among classical singers and untrained normal speakers. *Journal of Voice*, *15*(3), 344–350. https://doi.org/10.1016/S0892-1997(01)00036-4

Bates, A. W. (2015). Teaching in a Digital Age - Second Edition. *Quarterly Review of Distance Education*.

Behrman, A., & Haskell, J. (2019). *Exercises for Voice Therapy*. (J. Haskell & A.

Behrman, Eds.) (3rd ed.). Plural publishing.

Bericat, E. (2016). The sociology of emotions: Four decades of progress. *Current Sociology*, *64*(3), 491–513. https://doi.org/10.1177/0011392115588355

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. https://doi.org/10.1007/BF00138871

Biggs, J., & Tang, C. (2011). Setting the stage for effective teaching. In J. Biggs (Ed.), *Teaching for quality learning at university.:what the student does.* (pp. 34–57). Open University Press.

Birch, P., Gümoes, B., Prytz, S., Karle, A., Stavad, H., & Sundberg, J. (2002). *Effects of a Velopharyngeal opening on the sound transfer characteristics of the vowel [a]. TMH-QPSR* (Vol. 43).

Birch, P., Gümoes, B., Stavad, H., Prytz, S., Björkner, E., & Sundberg, J. (2002). Velum Behavior in Professional Classic Operatic Singing. *Journal of Voice*, *16*(1), 61–71.

Björkner, E. (2008). Musical Theater and Opera Singing-Why So Different? A Study of Subglottal Pressure, Voice Source, and Formant Frequency Characteristics. *Journal of Voice*, *22*(5), 533–540. https://doi.org/10.1016/j.jvoice.2006.12.007

Björkner, E., Sundberg, J., Cleveland, T., & Stone, E. (2006). Voice Source Differences Between Registers in Female Musical Theater Singers. *Journal of Voice*, *20*(2), 187–197. https://doi.org/10.1016/j.jvoice.2005.01.008

Bloch, M., Lave, J., & Wenger, E. (1994). Situated Learning: Legitimate Peripheral Participation. *Man*. https://doi.org/10.2307/2804509

Bloom, B. S. (1979). Taxonomy of Educational Objectives: The Classification of Educational Goals. In *Handbook I: Cognitive Domain*. London: Longman.

Boersma, P., & Weenink, D. (2014). Praat.

Borch, D. Z., Sundberg, J., Lindestad, P.-Å., & Thalén, M. (2004). Vocal fold vibration and voice source aperiodicity in "dist" tones: a study of a timbral ornament in rock singing. *Logopedics, Phoniatrics, Vocology*, *29*(6), 147–153. https://doi.org/10.1080/14015430410016073

Bottalico, P., Graetzer, S., & Hunter, E. J. (2016). Effect of Training and Level of External Auditory Feedback on the Singing Voice: Volume and Quality. *Journal of Voice*, *30*(4), 434–442. https://doi.org/10.1016/j.jvoice.2015.05.010

Bourne, T., & Garnier, M. (2012). Physiological and acoustic characteristics of the female music theater voice. *The Journal of the Acoustical Society of America*, *131*(2), 1586–1594. https://doi.org/10.1121/1.3675010

Bourne, T., Garnier, M., & Samson, A. (2016). Physiological and acoustic characteristics of the male music theatre voice. *The Journal of the Acoustical Society of America*, *140*(1), 610–621. https://doi.org/10.1121/1.4954751

Brockmann-Bauser, M., Bohlender, J. E., & Mehta, D. D. (2018). Acoustic Perturbation Measures Improve with Increasing Vocal Intensity in Individuals With and Without Voice Disorders. *Journal of Voice*, *32*(2), 162–168. https://doi.org/10.1016/j.jvoice.2017.04.008

Brockmann, M., Drinnan, M. J., Storck, C., & Carding, P. N. (2011). Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of Voice*, *25*(1), 44–53. https://doi.org/10.1016/j.jvoice.2009.07.002

Brockmann, M., Storck, C., Carding, P. N., & Drinnan, M. J. (2008). Voice loudness and gender effects on jitter and shimmer in healthy adults. *Journal of Speech, Language, and Hearing Research*, *51*(5), 1152–1160. https://doi.org/10.1044/1092-4388(2008/06-0208)

Brown, O. L. (2007). *Discover your voice*. New York: Delmar cenage learning.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments* (2nd ed.). University of California Press.

Bunch, M. (1997). *Dynamics of the singing voice* (4th ed.). Wien: Springer-Verlag.

Bunch, M., & Chapman, J. L. (2000). Taxonomy of singers used as subjects in scientific research. *Journal of Voice*, *14*(3), 363–369. https://doi.org/10.1016/S0892-1997(00)80081-8

Câmara, R., & Griessenauer, C. J. (2015). Anatomy of the Vagus Nerve. In R. S.

Tubbs, E. Rizk, M. M. Shoja, M. Loukas, N. Barbaro, & R. J. Spinner (Eds.), *Nerves and Nerve Injuries* (1st ed., pp. 385–396). Academic Press. https://doi.org/10.1016/B978-0-12-410390-0.00028-7

Chan, R. W., & Titze, I. (2006). Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics. *The Journal of the Acoustical Society of America*, *119*(4), 2351–2362.

Chandler, D., & Munday, R. (2011). *A Dictionary of Media and Communication*. *A Dictionary of Media and Communication*. Oxford university press. https://doi.org/10.1093/acref/9780199568758.001.0001

Chapman, J. L. (2006). *Singing and Teaching Singing A holistic approach to classical voice*. San DIego: Plural publishing.

Chhetri, D. K., & Park, S. J. (2016). Interactions of subglottal pressure and neuromuscular activation on fundamental frequency and intensity. *Laryngoscope*, *126*(5), 1123–1130. https://doi.org/10.1002/lary.25550

Collyer, S., Kenny, D. T., & Archer, M. (2009). The effect of abdominal kinematic directives on respiratory behaviour in female classical singing. *Logopedics Phoniatrics Vocology*, *34*(3), 100–110. https://doi.org/10.1080/14015430903008780

Collyer, S., Kenny, D. T., & Archer, M. (2011). Listener perception of the effect of abdominal kinematic directives on respiratory behavior in female classical singing. *Journal of Voice*, *25*(1), e15–e24.

Coolican, H. (2009). *Research Methods and Statistics in Psychology* (5th ed.). London: Hodder Education.

Coutinho, E., Scherer, K. R., & Dibben, N. (2019). Singing and Emotion. In *The Oxford Handbook of Singing* (pp. 297–314). Oxford: Oxford university press.

Cunningham, S., Weinel, J., & Picking, R. (2018). High-level analysis of audio features for identifying emotional valence in human singing. In *ACM International Conference Proceeding Series* (pp. 1–4). https://doi.org/10.1145/3243274.3243313

D'Angelo, E., Solinas, S., Garrido, J., Casellato, C., Pedrocchi, A., Mapelli, J., … Prestori, F. (2013). Realistic modeling of neurons and networks: Towards

brain simulation. *Functional Neurology*, *28*(3), 153–166. https://doi.org/10.11138/FNeur/2013.28.3.153

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, *12*(5), 1085–1101. https://doi.org/10.1037/a0025737

Darwin, C. (1873). *The expression of the emotions in man and animals*. *The expression of the emotions in man and animals*. New York: D. Appleton and company.

Davis, P. J., Zhang, S. P., Winkworth, A., & Bandler, R. (1996). Neural control of vocalization: Respiratory and emotional influences. *Journal of Voice*, *10*(1), 23–38. https://doi.org/10.1016/S0892-1997(96)80016-6

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research*, *91*(1), 176–180. https://doi.org/10.1007/BF00230027

Draper, M. H., Ladefoged, P., & Whitteridge, D. (1959). Respiratory muscles in speech. *Journal Od Speech & Hearing Research*, *2*, 16–27.

Dromey, C., Holmes, S. O., Hopkin, J. A., & Tanner, K. (2015). The effects of emotional expression on vibrato. *Journal of Voice*, *29*(2), 170–181. https://doi.org/10.1016/j.jvoice.2014.06.007

Echternach, M., Burk, F., Burdumy, M., Traser, L., & Richter, B. (2016). Morphometric differences of vocal tract articulators in different loudness conditions in singing. *PLoS ONE*, *11*(4), p.e0153792-e0153792. https://doi.org/10.1371/journal.pone.0153792

Echternach, M., Burk, F., Köberlein, M., Selamtzis, A., Döllinger, M., Burdumy, M., … Herbst, C. T. (2017). Laryngeal evidence for the first and second passaggio in professionally trained sopranos. *PLoS ONE*, *12*(5), e0175865 1-18. https://doi.org/https://doi.org/10.1371/ journal.pone.0175865

Echternach, M., Döllinger, M., Sundberg, J., Traser, L., & Richter, B. (2013). Vocal fold vibrations at high soprano fundamental frequencies. *The Journal of the Acoustical Society of America*, *133*(2), EL82–EL87.

Edmondson, J. A., & Esling, J. H. (2006). The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies.

Phonology, *23*(2), 157–191. https://doi.org/10.1017/S095267570600087X

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392. https://doi.org/10.1037/0003-066X.48.4.384

Erickson, D., Shochi, T., Menezes, C., Kawahara, H., & Sakakibara, K.-I. (2008). Some non-F0 cues to emotional speech: An experiment with morphing. In *Proceedings of the 4th International Conference on Speech Prosody* (pp. 677–680).

Estenne, M., Zocchi, L., Ward, M., & Macklem, P. T. (1990). Chest wall motion and expiratory muscle use during phonation in normal humans. *Journal of Applied Physiology*, *68*(5), 2075–2082. https://doi.org/10.1152/jappl.1990.68.5.2075

Eyben, F., Salomão, G. L., Sundberg, J., Scherer, K. R., & Schuller, B. W. (2015). Emotion in the singing voice—a deeperlook at acoustic features in the light ofautomatic classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1), 19. https://doi.org/10.1186/s13636-015-0057-6

Fant, G. (1970). *Acoustic theory of speech production description and analysis of contemporary standard russian*. (R. Jakobson & C. H. Schooneveld, Eds.) (2nd ed.). The Hague: De Gruyter, Inc. (Mouton). https://doi.org/10.2307/304731

Fant, G. (1986). Glottal flow: models and interaction. *Journal of Phonetics*. https://doi.org/10.1016/s0095-4470(19)30714-4

Farrús, M., & Hernando, J. (2009). Using Jitter and Shimmer in speaker verification. *IET Signal Processing*, *3*(4), 247–257. https://doi.org/10.1049/iet-spr.2008.0147

Feather, N. T., & Newton, J. W. (1982). Values, expectations, and the prediction of social action: An expectancy-valence analysis. *Motivation and Emotion*, *6*(3), 217–244. https://doi.org/10.1007/BF00992246

Fisher, J., Kayes, G., & Popeil, L. (2019). Pedagogy of different sung genres. In G. Welch, D. Howard, & J. Nix (Eds.), *The Oxford Handbook of Singing* (pp. 707–727). Oxford: Oxford university press.

Frøkjaer-Jensen, B., & Prytz, S. (1976). Registration of voice quality. *Bruel-Kjaer Technology Review*, *3*, 3–17.

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

Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, *5*(1), 123–147. https://doi.org/10.1177/10298649020050s105

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(2), 593–609. https://doi.org/10.1093/brain/119.2.593

Garnier, M., Henrich, N., & Crevier-Buchman, L. (2012). Glottal behavior in the high soprano range and the transition to the whistle register. *The Journal of the Acoustical Society of America*, *131*(1), 951–962. https://doi.org/https://doi.org/10.1121/1.3664008

Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the Somatotopic Auditory Mirror System in Humans. *Current Biology*, *16*(18), 1824–1829. https://doi.org/10.1016/j.cub.2006.07.072

Gill, B. P., Lee, J., Lã, F. M. B., & Sundberg, J. (2020). Spectrum Effects of a Velopharyngeal Opening in Singing. *Journal of Voice*, *34*(3), 346–351.

Gobl, C., & Ní Chasaide, A. (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

Griffin, B., Woo, P., Colton, R., Casper, J., & Brewer, D. (1995). Physiological characteristics of the supported singing voice. A preliminary study. *Journal of Voice*, *9*(1), 45–56.

Grillo, E. U., & Verdolini, K. (2008). Evidence for Distinguishing Pressed, Normal, Resonant, and Breathy Voice Qualities by Laryngeal Resistance and Vocal Efficiency in Vocally Trained Subjects. *Journal of Voice*, *22*(5), 546–552. https://doi.org/10.1016/j.jvoice.2006.12.008

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge Point: A Framework for

Conceptualizing the Effects of Various Practice Conditions in Motor Learning. *Journal of Motor Behavior*, *36*(2), 212–224. https://doi.org/10.3200/JMBR.36.2.212-224

Guzman, M., Dowdall, J., Rubin, A. D., Maki, A., Levin, S., Mayerhoff, R., & Jackson-Menaldi, M. C. (2012). Influence of emotional expression, loudness, and gender on the acoustic parameters of vibrato in classical singers. *Journal of Voice*, *26*(5), 675.e5-675.e11. https://doi.org/10.1016/j.jvoice.2012.02.006

Guzman, M., Lanas, A., Olavarria, C., Azocar, M. J., Muñoz, D., Madrid, S., … Mayerhoff, R. (2015). Laryngoscopic and spectral analysis of laryngeal and pharyngeal configuration in non-classical singing styles. *Journal of Voice*, *29*(1), p.130.e21-130.e28. https://doi.org/10.1016/j.jvoice.2014.05.004

Hakanpää, T., Waaramaa, T., & Laukkanen, A.-M. (2019). Emotion Recognition From Singing Voices Using Contemporary Commercial Music and Classical Styles. *Journal of Voice*, *33*(4), 501–509. https://doi.org/https://doi.org/10.1016/j.jvoice.2018.01.012

Hakanpää, T., Waaramaa, T., & Laukkanen, A.-M. (2021a). Comparing Contemporary Commercial and Classical Styles: Emotion Expression in Singing. *Journal of Voice*, *35*(4), 570–580. https://doi.org/https://doi.org/10.1016/j.jvoice.2019.10.002

Hakanpää, T., Waaramaa, T., & Laukkanen, A.-M. (2021b). Training the Vocal Expression of Emotions in Singing: Effects of Including Acoustic Research-Based Elements in the Regular Singing Training of Acting Students. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.12.032

Hallqvist, H., Lã, F. M. B., & Sundberg, J. (2017). Soul and Musical Theater: A Comparison of Two Vocal Styles. *Journal of Voice*, *31*(2), 229–235. https://doi.org/10.1016/j.jvoice.2016.05.020

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-Laryngologica*, *90*(1–6), 441–451. https://doi.org/10.3109/00016488009131746

Harrison, P. T. (2006). *The human nature of the singing voice exploring a holistic basis for sound teaching and learning*. Edinburgh: Dunedin Academic Press Ltd.

Hawk, S. T., Fischer, A. H., & Van Kleef, G. A. (2012). Face the noise: Embodied responses to nonverbal vocalizations of discrete emotions. *Journal of Personality and Social Psychology*, *102*(4), 796–814. https://doi.org/10.1037/a0026234

Healey, M. (2005). Linking research and teaching to benefit student learning. *Journal of Geography in Higher Education*, *29*(2), 183–201. https://doi.org/10.1080/03098260500130387

Helou, L. B., Jennings, R. J., Rosen, C. A., Wang, W., & Verdolini Abbot, K. (2020). Intrinsic Laryngeal Muscle Response to a Public Speech Preparation Stressor: Personality and Autonomic Predictors. *Journal of Speech, Language, and Hearing Research*, *63*, 2940–2951.

Helou, L. B., Rosen, C. A., Wang, W., & Verdolini Abbot, K. (2018). Intrinsic Laryngeal Muscle Response to a Public Speech Preparation Stressor. *Journal of Speech, Language, and Hearing Research*, *61*, 1525–1543.

Henrich Bernardoni, N., Smith, J., & Wolfe, J. (2014). Vocal tract resonances in singing: Variation with laryngeal mechanism for male operatic singers in chest and falsetto registers. *The Journal of the Acoustical Society of America*, *135*(1), 491–501. https://doi.org/10.1121/1.4836255

Henrich, N. (2006). Mirroring the voice from Garcia to the present day: some insights into singing voice registers. *Logopedics, Phoniatrics, Vocology*, *31*(1), 3–14.

Herbst, C. T. (2017). A review of the singing voice subsystem interactions - toward an extended physiological model of "support." *Journal of Voice*, *31*(2), 249.e13-249.e19.

Herbst, C. T., Hess, M., Müller, F., Švec, J. G., & Sundberg, J. (2015). Glottal Adduction and Subglottal Pressure in Singing. *Journal of Voice*, *29*(4), 391–402. https://doi.org/10.1016/j.jvoice.2014.08.009

Herbst, C. T., Howard, D. M., & Svec, J. (2019). The Sound Source in Singing: Basic Principles and Muscular Adjustments for Fine-tuning Vocal Timbre. In *The Oxford Handbook of Singing* (pp. 109–144).

Herbst, C. T., Schutte, H. K., & Švec, J. G. (2011). Membranous and cartilaginous vocal fold adduction in singing. *The Journal of the Acoustical Society of America*,

129(4), 2253–2262.

Herbst, C. T., Ternström, S., & Švec, J. G. (2009). Investigation of four distinct glottal configurations in classical singing-a pilot study. *The Journal of the Acoustical Society of America*, *125*(3), EL104–EL109.

Hirano, M. (1974). Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica et Logopaedica*, *26*(2), 89–94.

Hirano, M., Kakita, M., Kawasaki, H., Gould, W., & Lambiase, A. (1981). Data from high-speed motion picture studies. In K. N. Stevens & M. Hirano (Eds.), *Vocal fold physiology* (pp. 85–93). Tokyo: University of Tokio Press.

Hirano, M., Kakita, Y., Ohmaru, K., & Kurita, S. (1982). Structure and Mechanical Properties of the Vocal Fold. *Speech and Language*, *7*, 271–298.

Hodges, D. a. (2010). Psychophysiological Measures. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and Emotion, Theory research applications* (pp. 279–312). Oxford university press.

Hollien, H. (1974). On vocal registers. *Journal of Phonetics1*, *2*, 125–143.

Hutchins, S., & Moreno, S. (2013). The linked dual representation model of vocal perception and production. *Frontiers in Psychology*, *4*, 825–825. https://doi.org/10.3389/fpsyg.2013.00825

Iida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, *40*(1–2), 161–187. https://doi.org/10.1016/S0167-6393(02)00081-X

Iwarsson, J., & Sundberg, J. (1998). Effects of Lung Volume on Vertical Larynx Position During Phonation. *Journal of Voice*, *12*(2), 159–165.

Iwarsson, J., Thomasson, M., & Sundberg, J. (1998). Effects of lung volume on the glottal voice source. *Journal of Voice*, *12*(4), 424–433.

Izard, C. . (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, *99*(3), 561–565. https://doi.org/10.1037//0033-295X.99.3.561

Izard, C. . (2007). Basic Emotions, Natural Kinds, Emotion Schemas, and a New

Paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280. https://doi.org/10.1111/j.1745-6916.2007.00044.x

Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage*, *34*(4), 1744–1753. https://doi.org/10.1016/j.neuroimage.2006.10.032

Jansens, S., Bloothooft, G., & de Krom, G. (1997). Perception And Acoustics Of Emotions In Singing. *In Proceedings of the Fifth European Conference on Speech Communication and Technology*, *0*, 0–3. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=9747D0A838F2790BD0161DCF94739C2E?doi=10.1.1.56.8871

Jenkins, A., & Healey, M. (2012). Research–led or research–based undergraduate curricula. In *University Teaching in Focus: A Learning-Centred Approach*. https://doi.org/10.4324/9780203079690

Johnstone, T. (2001). *The effect of emotion on voice production and speech acoustics.*

Juslin, P. N. (1997). Emotional Communication in Music Performance: A Functionalist Perspective and Some Data. *Music Perception: An Interdisciplinary Journal*, *14*(4), 383–418. https://doi.org/10.2307/40285731

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2013.05.008

Juslin, P. N., & Laukka, P. (2001). Impact of Intended Emotion Intensity on Cue Utilization and Decoding Accuracy in Vocal Expression of Emotion. *Emotion*, *1*(4), 381–412. https://doi.org/10.1037/1528-3542.1.4.381

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

Juslin, P. N., & Laukka, P. (2004). Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research*, *33*(3), 217–238. https://doi.org/10.1080/0929821042000317813

Juslin, P. N., & Scherer, K. R. (2008). Vocal Expression of Affect. In J. Harrigan, R. Rosenthal, & K. Scherer (Eds.), *The New Handbook of Methods in Nonverbal Behavior Research* (pp. 65–136). Oxford: Oxford university press. https://doi.org/10.1093/acprof:oso/9780198529620.001.0001

Juslin, P. N., & Sloboda, J. A. (2010). *Handbook of Music and Emotion: Theory, Research, Applications*. *Handbook of music and emotion Theory research applications*. https://doi.org/10.1093/acprof

Juslin, P. N., & Sloboda, J. A. (2011). Introduction: aims, organization and terminology. In J. Sloboda & P. Juslin (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 3–14). Oxford: Oxford university press. https://doi.org/10.1093/acprof:oso/9780199230143.003.0001

Kempton, W., Scherer, K. R., & Giles, H. (1981). *Social Markers in Speech*. *Man*. https://doi.org/10.2307/2801995

Keysers, C., & Fadiga, L. (2008). The mirror neuron system: New frontiers. *Social Neuroscience*, *3*(3–4), 193–198. https://doi.org/10.1080/17470910802408513

Keysers, C., Kohler, E., Umiltà, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research*, *153*(4), 628–636. https://doi.org/10.1007/s00221-003-1603-5

Kitzing, P. (1986). LTAS criteria pertinent to the measurement of voice quality. *Journal of Phonetics*, *14*(3–4), 477–482. https://doi.org/10.1016/s0095-4470(19)30693-x

Koenig, L. L., & Fuchs, S. (2019). Vowel formants in normal and loud speech. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1278–1295. https://doi.org/10.1044/2018_JSLHR-S-18-0043

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, *297*(5582), 846–848. https://doi.org/10.1126/science.1070311

Kotlyard, G., & Morozov, V. (1976). Acoustical correlates of emotional content on vocalized speech. *Sov Phys Acoust*, *22*, 208–2011.

Kumar, S. P., Phadke, K. V., Vydrova, J., Novozamsky, A., Zitova, B., & Švec, J. (2020). Visual and automatic evaluation of vocal fold mucosal waves through sharpness of lateral peaks in high-speed videokymographic images. *Journal of Voice*, *34*(2), 170–178. https://doi.org/https://doi.org/10.1016/j.jvoice.2018.08.022

Lã, Filipa M.B., & Gill, B. P. (2019). Physiology and its impact on the performance of singing. In G. Welch, D. Howard, & J. Nix (Eds.), *The Oxford Handbook of Singing* (pp. 67–83). Oxford: Oxford university press.

Lã, Filipa Martins Baptista, & Sundberg, J. (2012). Pregnancy and the singing voice: Reports from a case study. *Journal of Voice*, *26*(4), 431–439. https://doi.org/10.1016/j.jvoice.2010.10.010

Lam Tang, J. A., Boliek, C. A., & Rieger, J. M. (2008). Laryngeal and Respiratory Behavior During Pitch Change in Professional Singers. *Journal of Voice*, *22*(6), 622–633. https://doi.org/10.1016/j.jvoice.2007.04.002

Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, *19*(5), 633–653. https://doi.org/10.1080/02699930441000445

Laukkanen, A.-M. (1995). *On Speaking Voice Exercises*. Acta Universitatis Tamperensis.

Laukkanen, A.-M., & Leino, T. (2001). *Ihmeellinen ihmisääni*. Gaudeamus.

Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: A preliminary study. *Journal of Phonetics*, *24*(3), 313–335. https://doi.org/10.1006/jpho.1996.0017

Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1997). On the perception of emotions in speech: the role of voice quality. *Logopedics Phonatrics Vocology*, *22*(4), 157–168. https://doi.org/10.3109/14015439709075330

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Lazarus, R. S. (1991). *Emotion & Adaptation. Oxford University Press.*

Lazarus, R. S., Averill, J. R., & Opton, E. M. j. (1970). Toward a cognitive theory of emotion. In M. B. Arnold (Ed.), *Feelings and emotions: The Loyola symposium* (pp. 207–232). New York: Academic Press.

Leanderson, R., Sundberg, J., & Von Euler, C. (1987). Role of diaphragmatic activity during singing: A study of transdiaphragmatic pressures. *Journal of Applied Physiology*, *62*(1), 259–270. https://doi.org/10.1152/jappl.1987.62.1.259

Lee, S.-H., Kwon, H.-J., Choi, H.-J., Lee, N.-H., Lee, S.-J., & Jin, S.-M. (2008). The Singer's Formant and Speaker's Ring Resonance: A Long-Term Average Spectrum Analysis. *Clinical and Experimental Otorhinolaryngology*, *1*(2), 92–96. https://doi.org/10.3342/ceo.2008.1.2.92

Leino, T., & Laukkanen, A.-M. (1993). Äänitysetäisyyden vaikutus puheäänen keskiarvospektriin. (Effects of microphone distance on the long-term-average spectrum of speech). In A. Iivonen & R. Aulanko (Eds.), *Papers from the 17th meeting of Finnish Phoneticians, Helsinki 1992* (pp. 117–129). Helsinki: Publications of the department of phonetics, University of Helsinki.

Levenson, R. W., Ekman, P., Heider, K., & Friesen, W. V. (1992). Emotion and Autonomic Nervous System Activity in the Minangkabau of West Sumatra. *Journal of Personality and Social Psychology*, *62*(6), 972–988. https://doi.org/10.1037/0022-3514.62.6.972

Lewis, M., Haviland-Jones, J., & Feldman Barrett, L. (2010). *Handbook of emotions* (3rd ed.). New York: The Guilford Press.

Li, N. Y. K., Verdolini, K., Clermont, G., Mi, Q., Rubinstein, E. N., Hebda, P. A., & Vodovotz, Y. (2008). A patient-specific in silico model of inflammation and healing tested in acute vocal fold injury. *PLoS ONE*, *3*(7), p.e2789-e2789. https://doi.org/10.1371/journal.pone.0002789

Li, Y., Li, J., & Akagi, M. (2018). Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space. *The Journal of the Acoustical Society of America*, *144*(2), 908–916.

Lindblom, B., & Sundberg, J. (1971). Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. *The Journal of the Acoustical Society of America*, *50*(4), 1166–1179. https://doi.org/10.1121/1.1912750

Lindblom, B., & Sundberg, J. (1972). Observations on tongue contour length in spoken and sung vowels. *STL-QPSR*, *13*(4), 001–005.

Livingstone, S. R., Choi, D. H., & Russo, F. A. (2014). The influence of vocal training and acting experience on measures of voice quality and emotional genuineness. *Frontiers in Psychology*, *5*(156), 1–13. https://doi.org/10.3389/fpsyg.2014.00156

Maas, E., Robin, D. A., Hula, S. N. A., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology*, *17*(3), 277–298. https://doi.org/10.1044/1058-0360(2008/025)

Macdonald, I., Rubin, J. S., Blake, E., Hirani, S., & Epstein, R. (2012). An Investigation of Abdominal Muscle Recruitment for Sustained Phonation in 25 Healthy Singers. *Journal of Voice*, *26*(6), 815.e9-815.e16.

Mainka, A., Poznyakovskiy, A., Platzek, I., Fleischer, M., Sundberg, J., & Mürbe, D. (2015). Lower vocal tract morphologic adjustments are relevant for voice timbre in singing. *PLoS ONE*, *10*(7), p.e0132241-e0132241. https://doi.org/10.1371/journal.pone.0132241

Manfredi, C., Barbagallo, D., Baracca, G., Orlandi, S., Bandini, A., & Dejonckere, P. H. (2015). Automatic Assessment of Acoustic Parameters of the Singing Voice: Application to Professional Western Operatic and Jazz Singers. *Journal of Voice*, *29*(4), 517.e1-517.e9. https://doi.org/10.1016/j.jvoice.2014.09.014

Marcel, Z., & Eerola, T. (2012). Self-Report Measures and Models. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion Theory, research, applications* (pp. 187–221). Oxford university press.

Maxwell, S. E., & Delaney, H. (2004). An introduction to Multilevel Hierarchical Mixed Models: Nested Designs. In *Designing experiments and analyzing data: A model comparison perspective.* (2nd ed., pp. 828–866). New York: Psychology Press.

Miller Richard. (1996). *The Structure of Singing, system and art in vocal technique*. Belmont CA: Wadsworth Group.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal

theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Mori, H., & Kasuya, H. (2007). Voice source and vocal tract variations as cues to emotional states perceived from expressive conversational speech. In *Interspeech* (pp. 102–105). https://doi.org/10.21437/Interspeech.2007-49

Mürbe, D., Pabst, F., Hoffmann, G., & Sundberg, J. (2002). Significance of Auditory and Kinesthetic Feedback to Singers' Pitch Control. *Journal of Voice*, *16*(1), 44–51.

Murphy, P. J., & Akande, O. O. (2007). Noise estimation in voice signals using short-term cepstral analysis. *The Journal of the Acoustical Society of America*, *121*(3), 1679–1690. https://doi.org/10.1121/1.2427123

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, *93*(2), 1097–1108. https://doi.org/10.1121/1.405558

Myers, B. R., & Finnegan, E. M. (2015). The effects of articulation on the perceived loudness of the projected voice. *Journal of Voice*, *29*(3), 390.e9-390.e15. https://doi.org/10.1016/j.jvoice.2014.07.022

Nair, D., & Large, E. W. (2002). Perceiving emotion in expressive piano performance: A functional MRI study. In *Proceedings of the 7th International Conference on Music Perception and Cognition*.

Neubauer, J., Edgerton, M., & Herzel, H. (2004). Nonlinear phenomena in contemporary vocal music. *Journal of Voice*, *18*(1), 1–12. https://doi.org/10.1016/S0892-1997(03)00073-0

Neumann, R., & Strack, F. (2000). "Mood contagion": The automatic transfer of mood between persons. *Journal of Personality and Social Psychology*, *79*(2), 211–223. https://doi.org/10.1037/0022-3514.79.2.211

Niedenthal, P. M. (2010). Emotion Concepts. In M. Lewis, J. M. Haviland-Jones, & L. Feldman Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 587–600). New York: The Guilford Press.

Niedenthal, P. M., Krauth-Gruber, S., & Ric, F. (2006). *Psychology of Emotion:*

Interpersonal, Experiential and Cognitive Approaches. New York: Psychology Press.

Niedenthal, P. M., & Ric, F. (2017). *Psychology of Emotion*. *Psychology of Emotion*. https://doi.org/10.4324/9781315276229

Niiniluoto, I. (2017). Optimistic realism about scientific progress. *Synthese*, *194*(9), 3291–3309. https://doi.org/10.1007/s11229-015-0974-z

Niiniluoto, I. (2018). Social aspects of scientific knowledge. *Synthese*, *197*(1), 1–22. https://doi.org/10.1007/s11229-018-1868-7

Nolen-Hoeksema, S., Fredrickson, B. L., Loftus, G. R., & Wagenaar, W. A. (2009). *Atkinson & Hilgard's Introduction to Psychology (15th Edition)*. *Wadsworth Cengage Learning*.

Nordenberg, M., & Sundberg, J. (2004). Effect on LTAS of vocal loudness variation. *Logopedics Phoniatrics Vocology*, *29*(4), 183–191. https://doi.org/10.1080/14015430410004689

Nummenmaa, L. (2019). *Tunnekartasto*. EU: Tammi.

Orlikoff, R. F., & Baken, R. J. (1989). The effect of the heartbeat on vocal fundamental frequency perturbation. *Journal of Speech and Hearing Research*, *32*(2), 576–582. https://doi.org/10.1044/jshr.3203.576

Orón Semper, J. V., & Blasco, M. (2018). Revealing the Hidden Curriculum in Higher Education. *Studies in Philosophy and Education*, *37*(5), 481–498. https://doi.org/10.1007/s11217-018-9608-5

Oxenham, A. J. (2012). Pitch perception. *Journal of Neuroscience*, *32*(39), 13335–13338. https://doi.org/10.1523/JNEUROSCI.3815-12.2012

Panksepp, J. (2008). The Affective Brain and The Core Consciousness. In M. Lewis, J. M. Haviland, & L. Feldman Barrett (Eds.), *Handbook of Emotions* (3rd ed., pp. 47–67). New York: The Guilford Press.

Parada-Cabaleiro, E., Baird, A., Batliner, A., Cummins, N., Hantke, S., & Schuller, B. W. (2017). The Perception of Emotion in the Singing Voice The Understanding of Music Mood for Music Organisation. In *Proceedings of 4th International Workshop on Digital Libraries for Musicology* (p. 8). Shanghai.

Parada-Cabaleiro, E., Costantini, G., Batliner, A., Schmitt, M., & Schuller, B. W. (2019). DEMoS: an italian emotional speech corpus. Elicitation methods, machine learning and perception. *Lang Resources & Evaluation*, *54*(2), 341–383. https://doi.org/https://doi.org/10.1007/s10579-019-09450-y

Park, Y., Yun, S., & Yoo, C. D. (2010). Parametric emotional singing voice synthesis. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2010.5495137

Peterson, L., Verdolini-Marston, K., Barkmeier, J. M., & Hoffman, H. T. (1994). Comparison of aerodynamic and electroglottographic parameters in evaluating clinically relevant voicing patterns. *Annals of Otology, Rhinology & Laryngology*, *103*(5 Pt 1), 335–346. https://doi.org/10.1177/000348949410300501

Pettersen, V., & Westgaard, R. H. (2004). Muscle activity in professional classical singing: A study on muscles in the shoulder, neck and trunk. *Logopedics Phoniatrics Vocology*, *29*(2), 56–65. https://doi.org/10.1080/14015430410031661

Pichon, S., de Gelder, B., & Grèzes, J. (2008). Emotional modulation of visual and motor areas by dynamic body expressions of anger. *Social Neuroscience*, *3*(3–4), 199–212. https://doi.org/10.1080/17470910701394368

Plack, C. J. (2014). *The sense of hearing. The Sense of Hearing* (2nd ed.). New York: Taylor & Francis. https://doi.org/10.4324/9781315881522

Podzimková, I., & Fric, M. (2019). Comparison of sound radiation between pop and classical singers. In *Models and Analysis of Vocal Emissions for Biomedical Applications - 11th International Workshop, MAVEBA 2019*. https://doi.org/10.1016/j.bspc.2021.102426

Pulkki, V., & Karjalainen, M. (2015). *Communication Acoustics an introduction to speech, audio and psychoacoustics*. John Wiley & Sons, Ltd.

Purves, D., Cabeza, R., Huettel, S., LaBar, K., Platt, M., & Woldorff, M. (2013). *Principles of cognitive neuroscience, second edition*.

Quirin, M., Kazén, M., & Kuhl, J. (2009). When Nonsense Sounds Happy or Helpless: The Implicit Positive and Negative Affect Test (IPANAT). *Journal*

of Personality and Social Psychology, *97*(3), 500–516. https://doi.org/10.1037/a0016063

Ramig, L., & Shipp, T. (1987). Comparative measures of vocal tremor and vocal vibrato. *Journal of Voice*, *1*(2), 162–167. https://doi.org/10.1016/S0892-1997(87)80040-1

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131–141. https://doi.org/10.1016/0926-6410(95)00038-0

Robbins, J. (2014). Practitioner inquiry. In C. Conway (Ed.), *The Oxford Hndbook of Qualitatieve Research in American Music Education* (pp. 186–208). New York City: Oxford university press.

Rogers, C. R., & Freiberg, H. J. (1994). *Freedom to learn (3rd ed.). Freedom to learn (3rd ed.).*

Rothenberg, M. (1981). An interactive model for the voice source. *STL-QPSR*, *22*(4), 001–017.

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

Sadolin, C. (2008). *complete vocal technique* (2nd ed.). CVI Publications.

Salomoni, S., Van Den Hoorn, W., & Hodges, P. (2016). Breathing and singing: Objective characterization of breathing patterns in classical singers. *PLoS ONE*, *11*(5), e0155084. https://doi.org/10.1371/journal.pone.0155084

Sampaio, M., Masson, M. L. V., Soares, M. F. de P., Bohlender, J. E., & Brockmann-Bauser, M. (2020). Effects of fundamental frequency, vocal intensity, sample duration, and vowel context in cepstral and spectral measures of dysphonic voices. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1326–1339. https://doi.org/10.1044/2020_JSLHR-19-00049

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Scherer, K. R. (1984a). Emotion as a multicomponent process: A model and

some cross-cultural data. *Review of Personality & Social Psychology*, *5*, 37–63.

Scherer, K. R. (1984b). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 295–318). Psychology Press. https://doi.org/https://doi.org/10.4324/9781315798806

Scherer, K. R. (1986). Vocal Affect Expression. A Review and a Model for Future Research. *Psychological Bulletin*, *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. https://doi.org/10.1016/S0892-1997(05)80231-0

Scherer, K. R. (2001). Appraisal Considered as a Process of Multilevel Sequential Checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, Methods, Research* (pp. 92–120). Oxford: Oxford university press.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R. (2004). Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them? *Journal of New Music Research*, *33*(3), 239–251. https://doi.org/10.1080/0929821042000317822

Scherer, K. R. (2005). What are emotion? And how can they be measured? *Social Science Information Sur Les Sciences Sociales*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech and Language*, *27*(1), 40–58. https://doi.org/10.1016/j.csl.2011.11.003

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). 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

Scherer, K. R., Dan, E., & Flykt, A. (2006). What determines a feeling's position in affective space? A case for appraisal. *Cognition & Emotion*, *20*(1), 92–113. https://doi.org/10.1080/02699930500305016

Scherer, K. R., & Moors, A. (2019). The Emotion Process: Event Appraisal and Component Differentiation. *Annual Review of Psychology*, *70*(1), 719–745. https://doi.org/10.1146/annurev-psych-122216-011854

Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., & Eyben, F. (2017). The expression of emotion in the singing voice: Acoustic patterns in vocal performance. *The Journal of the Acoustical Society of America*, *142*(4), 1805–1815. https://doi.org/10.1121/1.5002886

Scherer, K. R., Sundberg, J., Tamarit, L., & Salomão, G. L. (2015). Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech and Language*, *29*(1), 218–235. https://doi.org/10.1016/j.csl.2013.10.002

Scherer, K. R., Trznadel, S., Fantini, B., & Sundberg, J. (2017). Recognizing emotions in the singing voice. *Psychomusicology: Music, Mind, and Brain*, *27*(4), 244–255. https://doi.org/10.1037/pmu0000193

Schiller, I. S., Remacle, A., & Morsomme, D. (2020). Imitating dysphonic voice: a suitable technique to create speech stimuli for spoken language processing tasks? *Logopedics Phoniatrics Vocology*, *45*(4), 143–150. https://doi.org/10.1080/14015439.2019.1659410

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, *82*(4), 225–260. https://doi.org/10.1037/h0076770

Schmidt, R. A. (2003). Motor schema theory after 27 years: Reflections and implications for a new theory. *Research Quarterly for Exercise and Sport*, *74*(4), 366–375. https://doi.org/10.1080/02701367.2003.10609106

Sears, T. A., & Davis, J. N. (1968). Control of respiratory muscles during voluntary breathing. *Annals of the New York Academy of Sciences*, *155*(A1), 183–190.

Seashore, C. E. (1923). Measurements on the Expression of Emotion in Music. *Proceedings of the National Academy of Sciences*, *9*(9), 323–325. https://doi.org/10.1073/pnas.9.9.323

Seashore, C. E. (1931). The Natural History of the Vibrato. *Proceedings of the National Academy of Sciences*, *17*(12), 623–626. https://doi.org/10.1073/pnas.17.12.623

Seikel, J. A., King, D. W., & Drumright, D. G. (2014). *Anatomy and physiology for speech, language, and hearing* (5th ed.). Cengage Learning.

Sell, K. (2005). *The diciplines of vocal pedagogy: Towards a Holistic Approach*. Ashgate Publishing Limited.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication (Urbana, IL. University of Illinois Press*.

Siegwart, H., & Scherer, K. R. (1995). Acoustic concomitants of emotional expression in operatic singing: The case of lucia in Ardi gli incensi. *Journal of Voice*, *9*(3), 249–260. https://doi.org/10.1016/S0892-1997(05)80232-2

Simpson, E. A., Oliver, W. T., & Fragaszy, D. (2008). Super-expressive voices: Music to my ears? *Behavioral and Brain Sciences*, *31*(5), 596–597. https://doi.org/10.1017/S0140525X08005517

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for Pain Involves the Affective but not Sensory Components of Pain. *Science*, *303*(5661), 1157–1162. https://doi.org/10.1126/science.1093535

Skelton, A. (2004). Understanding "teaching excellence" in higher education: A critical evaluation of the national teaching fellowships scheme. *Studies in Higher Education*, *29*(4), 451–468. https://doi.org/10.1080/0307507042000236362

Skinner, B. F. (1968). *The technology of Teaching*. (J. s. Vargas, Ed.). B.F. Skinner Foundation Reprint Series.

Smith-Lovin, L., Lewis, M., & Haviland, J. M. (1995). *Handbook of Emotions*. (Lewis, Haviland-Jones, & Barret Feldman, Eds.), *Contemporary Sociology* (3rd ed., Vol. 24). New York: The Guilford Press.

https://doi.org/10.2307/2076468

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131. https://doi.org/10.1207/S15327957PSPR0402_01

Steinhauer, K., McDonald Klimek, M., & Estill, J. (2017). *The Estill voice model Theory & Translation*. Pittsburgh, Pennsylvania: Estill Voice International.

Story, B. H. (2016). The Vocal Tract in Singing. In G. Welch, D. M. Howard, & J. Nix (Eds.), *The Oxford Handbook of Singing* (pp. 1–21). https://doi.org/10.1093/oxfordhb/9780199660773.013.012

Strand, E. A., & Debertine, P. (2000). The efficacy of integral stimulation intervention with developmental apraxia of speech. *Journal of Medical Speech-Language Pathology*, *8*(4), 295–300.

Sundberg, J. (1974). Articulatory interpretation of the "singing formant." *The Journal of the Acoustical Society of America*, *55*(4), 838–844. https://doi.org/10.1121/1.1914609

Sundberg, J. (1975). Formant technique in a professional female singer. *Acustica*, *32*(2), 89–96.

Sundberg, J. (1987). *The science of the singing voice*. DeKalb, Illinois: Northen Illinois University Press.

Sundberg, J. (1994). Acoustic and psychoacoustic aspects of vocal vibrato. In *STL-QPSR*.

Sundberg, J. (1998). Expressivity in singing. A review of some recent investigations. *Logopedics Phoniatrics Vocology*, *23*(3), 121–127. https://doi.org/10.1080/140154398434130

Sundberg, J. (2000). Emotive Transforms: acoustic patterning of speech Its Linguistic and Physiological Bases. *Phonetica*, (57), 95–112.

Sundberg, J. (2001). Level and center frequency of the singer's formant. *Journal of Voice*, *15*(2), 176–186. https://doi.org/10.1016/S0892-1997(01)00019-4

Sundberg, J. (2017). Flow Glottogram and Subglottal Pressure Relationship in Singers and Untrained Voices. *Journal of Voice*, *32*(1), 23–31. https://doi.org/10.1016/j.jvoice.2017.03.024

Sundberg, J., Birch, P., Gümoes, B., Stavad, H., Prytz, S., & Karle, A. (2007). Experimental Findings on the Nasal Tract Resonator in Singing. *Journal of Voice*, *21*(2), 127–137.

Sundberg, J., Iwarsson, J., & Hagegård, H. (1995). A Singer's Expression of Emotions in Sung Performance. In O. Fujimura & M. Hirano (Eds.), *Vocal Fold Physiology: Voice Quality Control* (pp. 81–92). Singular Pub Group.

Sundberg, J., Lã, F. M. B., & Himonides, E. (2013). Intonation and Expressivity: A Single Case Study of Classical Western Singing. *Journal of Voice*, *27*(3), 391.e1-391.e8.

Sundberg, J., Leanderson, R., & von Euler, C. (1986). Voice source effects of diaphragmatic activity in singing. *Journal of Phonetics*, *14*(3–4), 351–357. https://doi.org/10.1016/s0095-4470(19)30700-4

Sundberg, J., & Nordenberg, M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The Journal of the Acoustical Society of America*, *120*(1), 435–457. https://doi.org/10.1121/1.2208451

Sundberg, J., Salomão, G. L., & Scherer, K. R. (2021). Analyzing Emotion Expression in Singing via Flow Glottograms, Long-Term-Average Spectra, and Expert Listener Evaluation. *Journal of Voice*, *35*(1), 52–60. https://doi.org/10.1016/j.jvoice.2019.08.007

Sundberg, J., & Skoog, J. (1997). Dependence of jaw opening on pitch and vowel in singers. *Journal of Voice*, *11*(3), 301–306. https://doi.org/10.1016/S0892-1997(97)80008-2

Sundberg, J., & Thalén, M. (2010). What is "twang"? *Journal of Voice*, *24*(6), 654–660. https://doi.org/10.1016/j.jvoice.2009.03.003

Suomi, K. (1990). *Johdatusta puheen akustiikkaan*. Oulun yliopisto.

Svec, J. (2018). Tutorial and guidelines on measurement of sound pressure level in voice and speech. *Journal of Speech, Language, and Hearing Research*, *61*(3),

441–461. https://doi.org/10.1044/2017_JSLHR-S-17-0095

Švec, J. G., Sundberg, J., & Hertegard, S. (2008). Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography. *The Journal of the Acoustical Society of America*, *123*(1), 347–353.

Szego, C. K. (2002). Music transmission and learning a conspectus of ethnographic research in ethnomusicology and music education. In R. Colwell & C. Richardson (Eds.), *The New Handbook of Research on Music Teaching and Learning : A Project of the Music Educators National Conference* (2nd ed., pp. 707–729). Oxford: Oxford university press.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, *27*(1), 24–27. https://doi.org/10.3758/BF03199901

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology*, *9*, 1112–1122. https://doi.org/10.1016/j.protcy.2013.12.124

Tennant, M., McMullen, C., & Kaczynski, D. (2010). Chapter 2: Perspectives on quality teaching. *Teaching, Learning and Research in Higher Education: A Critical Approach*.

Thomasson, M. (2003). Belly-in or belly-out? Effects of inhalatory behavior and lung volume on voice function in male opera singers. *TMH-QPSR*, *45*, 61–73.

Thomasson, Monica, & Sundberg, J. (1999). Consistency of Phonatory Breathing Patterns in Professional Operatic Singers. *Journal of Voice*, *13*(4), 529–541.

Thorpe, C. W., Cala, S. J., Chapman, J. L., & Davis, P. J. (2001). Patterns of breath support in projection of the singing voice. *Journal of Voice*, *15*(1), 86–104.

Timcke, R., Von Leden, H., & Moore, G. P. (1958). Laryngeal vibrations: Measurements of the glottic wave: Part 1. The normal vibratory cycle. *AMA Archives of Otolaryngology*, *68*, 1–19.

Timcke, R., Von Leden, H., & Moore, G. P. (1959). Laryngeal vibrations: Measurements of the glottic wave: Part 2. physiologic variations. *AMA Archives of Otolaryngology*, *69*, 238–444.

Titze, I. (1988a). A framework for the study of vocal registers. *Journal of Voice*, *2*(3), 183–194.

Titze, I. (1988b). The physics of small-amplitude oscillation of the vocal folds. *The Journal of the Acoustical Society of America*, *83*(4), 1536–1552.

Titze, I. (1991). A model for neurologic sources of aperiodicity in vocal fold vibration. *Journal of Speech and Hearing Research*, *34*(3), 460–472. https://doi.org/10.1044/jshr.3403.460

Titze, I. (2000). *Principles of Voice Production* (2nd ed.). Iowa City: National Center for Voice and Speech.

Titze, I. (2008a). Nonlinear source-filter coupling in phonation: theory. *The Journal of the Acoustical Society of America*, *123*(5), 2733–2749.

Titze, I. (2008b). Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, *123*(5), 2733–2749. https://doi.org/10.1121/1.2832337

Titze, I. (2015). On flow phonation and airflow management. *Journal of Singing*, *72*(1), 57–58.

Titze, I., & Alipour, F. (2006). *The myoelastic aerodynamic theory of phonation*. Denver Colorado: National Center for Voice and Speech.

Titze, I., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., … Wolfe, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America*, *137*(5), 3005–3007. https://doi.org/10.1121/1.4919349

Titze, I., Bergan, C. C., Hunter, E. J., & Story, B. H. (2003). Source and filter adjustments affecting the perception of the vocal qualities twang and yawn. *Logopedics Phoniatrics Vocology*, *28*(4), 147–155. https://doi.org/10.1080/14015430310018874

Titze, I., & Martin, D. W. (1998). Principles of Voice Production. *The Journal of the Acoustical Society of America*, *104*(3), 1148. https://doi.org/10.1121/1.424266

Titze, I., Riede, T., & Popolo, P. (2008). Nonlinear source–filter coupling in phonation: Vocal exercises. *The Journal of the Acoustical Society of America2*, *123*(4), 1902–1915. https://doi.org/10.1121/1.2832339

Titze, I., & Verdolini Abbot, K. (2012). *Vocology the science and practice of voice habilitation*. NCVS.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981. https://doi.org/10.1080/01690960903498424

Traser, L., Özen, A. C., Burk, F., Burdumy, M., Bock, M., Richter, B., & Echternach, M. (2016). Respiratory dynamics in phonation and breathing—A real-time MRI study. *Respiratory Physiology and Neurobiology*, *236*, 69–77. https://doi.org/10.1016/j.resp.2016.11.007

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing: A neurophysiological study. *Neuron*, *31*(1), 155.165. https://doi.org/10.1016/S0896-6273(01)00337-3

Vampola, Tomáš, Horácek, J., & Laukkanen, A.-M. (2021). Finite element modeling of the effects of velopharyngeal opening on vocal tract reactance in female voice. *The Journal of the Acoustical Society of America*, *150*(3), 2154–2162.

Vampola, Tomas, Horacek, J., Radolf, V., Švec, J. G., & Laukkanen, A.-M. (2020). Influence of nasal cavities on voice quality: Computer simulations and experiments. *The Journal of the Acoustical Society of America*, *148*(5), 3218–3231.

Van den Berg, J. (1958). Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research*, *3*, 227–244.

van der Gaag, C., Minderaa, R. B., & Keysers, C. (2007). Facial expressions: What the mirror neuron system can and cannot tell us. *Social Neuroscience*, *2*(3–4), 179–222. https://doi.org/10.1080/17470910701376878

van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, *9*, 1994–1994.

https://doi.org/10.3389/fpsyg.2018.01994

Vennard, W., & Minoru, H. (1970). Physiological basis for vocal registers. *The Journal of the Acoustical Society of America*, *47*(1A), 120.

Verdolini, K., Druker, D. G., Palmer, P. M., & Samawi, H. (1998). Laryngeal adduction in resonant voice. *Journal of Voice*, *12*(3), 315–327. https://doi.org/10.1016/S0892-1997(98)80021-0

Vilkman, E., Alku, P., & Vintturi, J. (2002). Dynamic extremes of voice in the light of time domain parameters extracted from the amplitude features of glottal flow and its derivative. *Folia Phoniatrica et Logopaedica*, *54*(3), 144–157.

Vos, R. R., Murphy, D. T., Howard, D. M., & Daffern, H. (2017). The Perception of Formant Tuning in Soprano Voices. *Journal of Voice*, *32*(1), 126.e1-126.e10. https://doi.org/10.1016/j.jvoice.2017.03.017

Vurma, A. (2020). Amplitude Effects of Vocal Tract Resonance Adjustments When Singing Louder. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.05.020

Vurma, A., & Ross, J. (2006). Production and perception of musical intervals. *Music Perception*, *23*(4), 331–344. https://doi.org/10.1525/mp.2006.23.4.331

Waaramaa, T., Alku, P., & Laukkanen, A.-M. (2006). The role of F3 in the vocal expression of emotions. *Logopedics Phoniatrics Vocology*, *31*(4), 153–156. https://doi.org/10.1080/14015430500456739

Waaramaa, T., Laukkanen, A.-M., Alku, P., & Väyrynen, E. (2008). Monopitched expression of emotions in different vowels. *Folia Phoniatrica et Logopaedica*, *60*(5), 249–255. https://doi.org/10.1159/000151762

Warhurst, S., Madill, C., McCabe, P., Heard, R., & Yiu, E. (2012). The vocal clarity of female speech-language pathology students: An exploratory study. *Journal of Voice*, *26*(1), 63–68. https://doi.org/10.1016/j.jvoice.2010.10.008

Watson, A. (2019). Breathing in singing. In G. Welch, D. Howard, & J. Nix (Eds.), *The Oxford Handbook of Singing* (pp. 87–107). Oxford: Oxford university press.

Watson, A., Williams, C., & James, B. V. (2011). Activity Patterns in Latissimus

Dorsi and Sternocleidomastoid in Classical Singers. *Journal of Voice*, *26*(3), e95–e105.

Watson, P. J., & Hixon, T. J. (1985). Respiratory kinematics in classical (opera) singers. *Journal of Speech and Hearing Research*, *28*(1), 104–122. https://doi.org/10.1044/jshr.2801.104

Watson, P. J., Hoit, J. D., Lansing, R. W., & Hixon, T. J. (1989). Abdominal muscle activity during classical singing. *Journal of Voice*, *3*(1), 24–31. https://doi.org/10.1016/S0892-1997(89)80118-3

Welch, G., Howard, D., Himonides, E., & Brereton, J. (2005). Real-time feedback in the singing studio: an innovatory action-research project using new voice technology. *Music Education Research*, *7*(2), 225–249.

Welch, G., Thurman, L., Theimer, A., Grefsheim, E., & Feit, P. (2000). how your vocal tract contributes to basic voice qualities. In *Bodymind and voice* (pp. 449–469).

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: The common neural basis of seeing and feeling disgust. *Neuron*, *40*(3), 655–664. https://doi.org/10.1016/S0896-6273(03)00679-2

Wiener, N. (1971). *Cybernetics or control and communication in the animal and the machine* (2. ed., re). Cambridge: Cambridge, MA:MIT.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America*, *52*(4), 1238–1250. https://doi.org/10.1121/1.1913238

Woody, R. H., Sloboda, J. A., & Lehmann, A. C. (2007). *Psychology for musicians: understanding and acquiring skills*. New York: Oxford university press.

Wulf, G., Lauterbach, B., & Toole, T. (1999). The learning advantages of an external focus of attention in golf. *Research Quarterly for Exercise and Sport*, *70*(2), 120–126. https://doi.org/10.1080/02701367.1999.10608029

Wulf, G., & Prinz, W. (2001). Directing attention to movement effects enhances learning: A review. *Psychonomic Bulletin and Review*, *8*(4), 648–660. https://doi.org/10.3758/BF03196201

Yanagisawa, E., Estill, J., Kmucha, S. T., & Leder, S. B. (1989). The contribution of aryepiglottic constriction to "ringing" voice quality-A videolaryngoscopic study with acoustic analysis. *Journal of Voice*, *3*(4), 342–350. https://doi.org/10.1016/S0892-1997(89)80057-8

Zhang, Zaoyan. (2016). Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, *140*(4), 2614–2635. https://doi.org/10.1121/1.4964509

Zhang, Zhaoyan. (2015). Regulation of glottal closure and airflow in a three-dimensional phonation model: Implications for vocal intensity control. *The Journal of the Acoustical Society of America*, *137*(2), 898–910. https://doi.org/10.1121/1.4906272

Zhou, G., Hansen, J. H. L., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *9*(3), 201–216. https://doi.org/10.1109/89.905995

PUBLICATIONS

PUBLICATION I

Emotion Recognition from Singing Voices Using Contemporary Commercial Music and Classical Styles

Tua Hakanpää, Teija Waaramaa, and Anne-Maria Laukkanen

Original publication: Journal of Voice 2019, 33(4)501-509
https://doi.org/10.1016/j.jvoice.2018.01.012

Emotion Recognition From Singing Voices Using Contemporary Commercial Music and Classical Styles

***Tua Hakanpää, *,†Teija Waaramaa, and *Anne-Maria Laukkanen,** *†*Tampere, Finland*

Summary: Objectives: This study examines the recognition of emotion in contemporary commercial music (CCM) and classical styles of singing. This information may be useful in improving the training of interpretation in singing.

Study design: This is an experimental comparative study.

Methods: Thirteen singers (11 female, 2 male) with a minimum of 3 years' professional-level singing studies (in CCM or classical technique or both) participated. They sang at three pitches (females: a, e1, a1, males: one octave lower) expressing anger, sadness, joy, tenderness, and a neutral state. Twenty-nine listeners listened to 312 short (0.63- to 4.8-second) voice samples, 135 of which were sung using a classical singing technique and 165 of which were sung in a CCM style. The listeners were asked which emotion they heard. Activity and valence were derived from the chosen emotions.

Results: The percentage of correct recognitions out of all the answers in the listening test (N = 9048) was 30.2%. The recognition percentage for the CCM-style singing technique was higher (34.5%) than for the classical-style technique (24.5%). Valence and activation were better perceived than the emotions themselves, and activity was better recognized than valence. A higher pitch was more likely to be perceived as joy or anger, and a lower pitch as sorrow. Both valence and activation were better recognized in the female CCM samples than in the other samples.

Conclusions: There are statistically significant differences in the recognition of emotions between classical and CCM styles of singing. Furthermore, in the singing voice, pitch affects the perception of emotions, and valence and activity are more easily recognized than emotions.

Key Words: Voice quality−Emotion expression−Perception−Song genre−Singing style.

INTRODUCTION

Singers need to express emotions vocally with great passion, but with sufficient control that the audience can identify with the portrayal of emotion and at the same time enjoy the brilliance of the musical sound. We know from speaking voice research that emotions are reflected, for example, in voice quality.[1−3] For singers, the act of emotional expression is particularly demanding because the voice apparatus is already working at full capacity for singing alone.

The optimal function of the voice in singing is usually achieved through the physical activity of the body, the breathing mechanism, the laryngeal muscles, and the articulators, which we refer to as "the singing technique." The continuously changing optimal alignment of bony and cartilaginous structures along with exactly the right amount and distribution of muscle function are required to be able to achieve each pitch in a piece of music with a stylistically relevant timbre of voice.[4−6]

Emotions change the voice tone,[7] often deteriorating it from optimally balanced phonation; this is true for both the singing and the speaking voice. In the world of singing, there are genre-specific esthetic quality standards to which a singer must adjust. The stylistic differences also manifest themselves under the umbrella concepts of "classical" and "popular" music. There are distinct technical differences in singing baroque versus opera or musical theater versus soul.[4] When expressing emotions, singers need to be aware of the effects that emotional expression exert on the voice so they can send their acoustic message without compromising the sonority of the voice too greatly.

Voice quality in singing

Much of the emotional information in the speaking voice is perceived from pitch, tempo, loudness, and rhythm, the use of which is restricted in musical performance.[8,9] If a singer follows the written music very strictly, the voice quality is really the only parameter that can be freely varied. This is true for both contemporary commercial music (CCM) and classical styles of singing. In speech and singing, the term "voice quality" refers to "the long-term auditory coloring of a person's voice."[10] In a broad sense, it is the combined product of both laryngeal (phonation-related) and supralaryngeal (articulation-related) factors.[10]

Different styles of singing require the use of different voice qualities.[11−13] One needs to configure the phonation and articulatory settings differently for almost every musical style. For example, in bossa nova, the phonation settings are often breathy, while articulatory settings function at full throttle to make rhythmical distinctions. One of the key technical elements in opera is to be loud enough to be heard

Accepted for publication January 16, 2018.

From the *Speech and Voice Research Laboratory, Faculty of Education, University of Tampere, Tampere, Finland; and the †Faculty of Communication Sciences, University of Tampere, Tampere, Finland.

Address correspondence and reprint requests to Tua Hakanpää, Speech and Voice Research Laboratory, Faculty of Education, University of Tampere, MA, Vanhan-Mankkaan tie 16 e 14, 02180, Espoo, Finland. E-mail: Hakanpaa.tua.s@student.uta.fi

Journal of Voice, Vol. 33, No. 4, pp. 501−509

0892-1997

https://doi.org/10.1016/j.jvoice.2018.01.012

over the orchestra without electric sound amplification, so the singer configures the apparatus to take maximum advantage of vocal tract resonances.[14] In some styles of the heavy metal, singers need to adorn the voice with constant distortion, making the underlining voice quality sound harsh.[15] On top of this rather stable "stylistic voice quality," a singer makes another layer of smaller changes that mark the rendition of the emotional content of a song. Regardless of the genre or emotional content of the song, the singer needs basic skills to control the vocal apparatus to match pitches, produce dynamic variation, and deliver efficient articulation and various side sounds (like sighs, grunts, etc) where needed.

Research questions

The recognition of emotion from the singing voice has been traditionally studied using short samples and listener group ratings.[16,17] Previous research has shown that in music, it is often the case that general categories of emotion are well recognized, but nuances (such as hot anger vs. cold anger) within these categories are not.[9] To our knowledge, the recognition of emotions between different styles of singing has not yet been studied. In the present paper, we study the recognition of four emotions (anger, sadness, tenderness, and joy). These emotions have been selected because they can be placed on a fourfold table of valence and activation. Anger, sadness, and joy are regarded as basic emotions and should by definition be easy to recognize.[18,19] Tenderness is included because an emotion with a positive valence and low activity level was needed to complete the fourfold table. All of these emotions occur frequently in song interpretations in both the classical and contemporary commercial worlds, and are thus familiar performance tasks for most singers.

In this preliminary study, we use short vowel samples as test material to investigate the role of voice quality in conveyance of emotions. Short vowels are used because they do not contain semantic or prosodic information. Therefore, they carry voice quality in its purest sense. Although the recognition percentage is not supposed to be high in short samples lacking prosodic information, earlier research on both speaking and singing voice suggests that emotional information can be received also from short samples[2,20,21]

The study is an experimental comparative design using singing voice samples and listener evaluations. The specific research questions of this study are:

1. Are listeners able to recognize emotions in a singing voice from short vowel samples?
2. Is there a difference in the recognition of emotions when they are sung using a classical singing technique compared to when they are sung using a CCM style of singing?
3. Does pitch affect the recognition of emotion in the classical- or CCM-style singing voice samples?
4. Are valence and the activation of the emotions perceptible in the sung samples?

METHODS

Thirteen professionally trained singers sang a small excerpt of a song expressing four different emotions. A listening test was created to determine whether the listeners' appraisals of the sung emotions matched the singers' intended expressions.

Participants and recording

The voice samples were gathered from 13 singers (2 males, 11 females) with a minimum of 3 years of singing lessons at a professional level. The mean age of the subjects was 32 years (range: 20−44 years). The mean number of years' singing experience was 10. Singers were either classically trained (n = 7) or CCM singers (N = 6), and all of the subjects were native Finnish-speakers. Six of the singers worked in both classical and CCM genres, and these subjects gave voice samples in both styles.

The singers were instructed to perform an eight-bar excerpt from a song expressing the emotions of joy, tenderness, sadness, or anger using either a CCM or classical technique. The song was Gershwin's *Summertime* with Finnish lyrics by Sauvo Puhtila. This song was chosen because it has been composed as an aria, but it is widely popular among CCM singers as well, so it fits both the classical and the CCM repertoire. The Finnish lyrics depict a nature scene that contains no particular emotional information as such. An effort was made to make the experiment as lifelike as possible. Therefore, the singers used a backing track with a neutral accompaniment suitable for both classical and CCM style singing that was played to them via an S-LOGIC ULTRASONE Signature PRO headset (ULTRASONE AG, Wielenbach, Bavaria, Germany). The studio setup also featured a Shure SM58 vocal microphone (Shure Incorporated, Niles, IL), which allowed the singers' singing voice to be mixed in with the backing track as they were singing.

Because pitch is known to vary in the expression of emotions in speech, and it could thus be expected to affect the perception of emotions, all subjects were instructed to use the same pitch (with the males singing one octave lower) regardless of genre.

The key of the song was D minor, and the tempo was 80 beats per minute for all test subjects and every emotional portrayal. The emotion samples were performed in a randomized order and repeated three times. The singers also gave a neutral voice sample without expressing any emotion. This sample was also repeated three times.

All recordings were made at the recording studio of Tampere University Speech and Voice Research Laboratory using a Brüel & Kjær Mediator 2238 microphone (Brüel & Kjær Sound & Vibration Measurement A/S [HQ], Nærum, Denmark). The distance between the microphone and test subjects' lips was 40 cm. The samples were recorded with *Sound Forge 7 digital audio editing software* Sony Europe Limited, Suomen suvuliike, Vantaa, Suomi) with a 44.1-kHz sampling rate using a 16-bit external soundcard (Quad-Capture Roland). All samples were saved as wav. files for

further analysis with *Praat* (Paul Boersma and David Weenink, Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands).[22]

Voice samples

The vowel [ɑ:] was extracted from three different pitches in each sample for further analyses. The pitches were, for the females, a, e1, a1 (A3, E4, A4 according to the American system), and A, e, and a for the males (A2, E3, A3). The [ɑ:] samples were extracted from the Finnish words *aikaa* ['ɑikɑ:] (time), *hiljaa* ['çiljɑ:] (softly), and *saa* [sɑ:] (to be allowed). The nominal duration of the extracted vowels (including the preceding consonant) were 2.25 seconds for a, 4.5 seconds for for e1, and 2.25 seconds for a1, according to the notation and tempo of the song.

The [ɑ:] vowels were extracted from the sung excerpts using *Sound Forge 7 audio editing software*. The samples were cut right after the preceding consonant. The duration of the sample varied between 0.6267 seconds and 4.8063 seconds, depending on how the test subject had interpreted the time value of the notation. The tail end of the vowel was left as the singer interpreted it (the nominal note durations 2.25 seconds or 4.5 seconds), as previous studies have indicated that micromanaging the durations of written notes is one way of expressing emotions in the singing voice.[23]

From a total number of 900 samples, 300 samples were chosen for the listening test (Table 1).

Listening test

The listening task was an Internet-based test with 300 randomized [ɑ:] vowel samples and 12 control samples. As the samples were numerous, we constructed the listening test so that it was possible to stop and continue the test as needed. The test was accessible through a browser by logging in with a password. Participants completed the test using their own equipment. The voice samples were played in a randomized order and it was possible to play the samples as many times as needed. The Finnish questionnaire was translated for the one listener who was not Finnish-speaking. We used Pearson chi-squared test of homogeneity to determine if the probability of recognition was the same for the native Finns and the non-native participant in order to check for the possible language-related or cultural differences in emotion recognition. The zero hypothesis that recognition percentage for group 1 (28 listeners) equals that of group 2 (1 listener) was to be accepted (z-value 2.0, P value 0.160) at a statistical significance level of $\alpha = 0.05$. We also tested the possible effects of listening preferences to recognition by comparing the answers of participants who predominantly listen to classical singing to those who mostly listen to CCM. The zero hypothesis that recognition percentage for group 1 (those who mostly listen to CCM music, 26 listeners) equals that of group 2 (those who mainly listen to classical music, three listeners) was to be accepted (z-value 0.3, P value 0.579) at a statistical significance level of $\alpha = 0.05$. The listening test took approximately 60 minutes to complete, and the participants were offered either study credits or voice lessons in exchange for the completed test. The number of people who completed the test was 29 (22 females, 7 males, no reported hearing defects), and they were all selected for further analysis. The listeners completed a multiple-choice questionnaire on which emotion they perceived (anger, sadness, joy, tenderness, neutral) for each sample. Eight of the listeners were professionally involved in assessing the human voice (singing teachers and vocologists) and 21 were laypeople. Seventeen of the listeners were singers (14 amateur and 3 professionals).

The number of samples used

The total number of samples listened by each listener was 312. There were 100 samples from each pitch, 60 samples + 3 control samples (repetitions of the same samples in a randomized order) for each emotion category, and 60 neutral samples. There were thus 20 samples depicting the same emotion category and pitch in the whole data sample. Of the samples, 135 were sung using a classical singing

TABLE 1.
Numbers of Voice Samples in the Listening Test (Total Number of Samples N = 300)

	Joy		Tenderness		Sadness		Anger		Neutral	
	Classical	CCM	Classical	CCM	Classical	CCM	Classical	CCM	Classical	CCM
High pitch	9	11	9	11	9	11	9	11	9	11
Male 220 Hz	1	1	1	1	1	1	1	1	1	1
Female 440 Hz	8	10	8	10	8	10	8	10	8	10
Medium pitch	9	11	9	11	9	11	9	11	9	11
Male 165 Hz	1	1	1	1	1	1	1	1	1	1
Female 330 Hz	8	10	8	10	8	10	8	10	8	10
Low pitch	9	11	9	11	9	11	9	11	9	11
Male 110 Hz	1	1	1	1	1	1	1	1	1	1
Female 220 Hz	8	10	8	10	8	10	8	10	8	10

TABLE 2.
Numbers of Answers Given in the Listening Test (Total Number of Samples N = 9048)

	Joy		Tenderness		Sadness		Anger		Neutral	
	Classical	CCM	Classical	CCM	Classical	CCM	Classical	CCM	Classical	CCM
High pitch	261	348	261	348	261	348	261	348	261	319
Male 220 Hz	29	29	29	29	29	29	29	29	29	29
Female 440 Hz	232	319	232	319	232	319	232	319	232	290
Medium pitch	261	348	261	348	261	348	261	348	261	319
Male 165 Hz	29	29	29	29	29	29	29	29	29	29
Female 330 Hz	232	319	232	319	232	319	232	319	232	290
Low pitch	261	348	261	348	261	348	261	348	261	319
Male 110 Hz	29	29	29	29	29	29	29	29	29	29
Female 220 Hz	232	319	232	319	232	319	232	319	232	290

technique, and 165 samples were sung using a CCM-style singing technique.

The total number of answers in the listening test was 9048. From each pitch, we gathered 3016 answers. From each emotion category, we drew 1827 answers, while 1740 answers were drawn for the neutral portrayals. The samples where a classical style singing technique was used resulted in 3915 answers, and the samples where a CCM-style singing technique was used resulted in 5133 answers (Table 2).

Statistical analyses

The results of the listening test were coded numerically for statistical analysis. Both the intended and perceived emotions were given numbers (1 = joy, 2 = tenderness, 3 = neutral, 4 = sadness, 5 = anger). The valence and activation of the emotions expressed and perceived were given arbitrary numbers based on the emotions chosen in the listening test. Positive valence (emotion that is regarded as pleasant) was coded as 2, negative valence as 1, and neutrality as 0. Activity (the energy level typically inherent in an emotion) was coded as low = 1, high = 2, or medium = 0. The pitch levels were coded as 1 = low, 2 = middle, or 3 = high pitch. The samples sung with a classical technique were marked as 1, and those sung in a CCM style were marked as 2.

The number of the correct (intended = perceived) answers for emotion, valence, and activity are given as percentages.

Furthermore, the results of the listening test were analyzed using three different statistical tests.

The first statistical test used was a binomial test (one-proportion *z*-test) to evaluate the probability that the observed percentage of the correctly recognized emotions could have resulted from random guessing. The listening test contained five different emotions, which meant that the expected percentage of correctly recognized emotions in case of random guessing would be 20%. The observed percentage of correctly recognized emotions differs statistically significantly from random guessing if the P value is <0.05.

The second statistical test, Pearson chi-squared test of homogeneity, was used to evaluate the probability that two groups of results have the same percentage of correctly recognized emotions. The percentage of correctly recognized emotions is statistically different in two groups of results if the P value of the test is <0.05.

The third statistical test, Cronbach alpha, was used to evaluate the reliability of the internal consistency of listener evaluations. Alpha values >0.7 indicate an acceptable internal consistency of the data.

Intra-rater reliability was estimated using Cohen kappa coefficient.

All statistical analyses were performed using *Microsoft Excel* (Microsoft, Redmond, Washington).

RESULTS

Accuracy of emotion recognition

The percentage of correct recognitions out of all the answers in the listening test (N = 9048) was 30.2%. According to the one-proportion *z*-test, the recognition exceeded random guessing (H0: $p = 1/5$: *z*-value 24, P value $<<0.001$). The internal consistency of the listeners' evaluation was good (Cronbach alpha 0.89). The intra-rater reliability was moderate (mean Cohen kappa 0.48) according to the Landis and Koch benchmark.[24] In females, the emotions were recognized significantly better from the CCM samples than from the classical samples. Recognition from the female samples sung using a CCM style was higher (1653 correct answers from 4698 answers given) than from the samples sung in a classical style (832 correct answers from the total number of 3480 answers given). The statistical significance of the 11.3% difference in recognition was evaluated using the Pearson chi-square test of homogeneity, and the difference was found to be significant (*z*-value 120.2, P value $<<0.001$). Recognition from the male singers' CCM- and classical-style samples was not statistically significantly different (Pearson chi-square *z*-value 0.5 and P value 0.5). Correct recognition occurred in 119 answers from a total of 435 given for a CCM style and 128 answers from a total of 435 given for the classical style.

The discrepancy between the number of CCM and classical female samples was considered a possible factor in information distortion. We performed Pearson chi-squared test of homogeneity to test if the probability of recognition was statistically the same for these two groups. To validate the use of a larger sample number in one group, we excluded the two best recognized female CCM singers from the sample battery, and compared the correctly recognized samples of the nine least well-recognized female CCM singers to the female classical singers. Pearson chi-squared test of homogeneity indicated that on the statistical significance level $\alpha \leq 0.05$, the zero hypothesis that the recognition percentage for group 1 (11 CCM singers) equals that of group 2 (nine classical singers) had to be discarded. The same was true when comparing group 1 (the nine least well-recognized CCM singers) to group 2 (nine classical singers). Thus, even after excluding the two best recognized sample batteries among the female CCM samples, the recognition of the CCM samples remained significantly better than that of the classical samples. Therefore, the discrepancy between the number of CCM and classical samples has not affected the results.

There was a median 3.9% difference in the overall recognition of emotions from the low frequency to the high frequency in such a way that the low frequency samples were systematically recognized more poorly than the high frequency samples in all sample groups (Table 3).

Table 4 indicates that in the case of all other emotions except sadness, recognition of the emotional content from the sung [a:] vowels from the female singers was easier when the samples were sung using a CCM-style technique. Sadness, on the other hand, was better recognized from the samples sung with a classical-style technique. From the male singers' samples, joy, tenderness, and neutral portrayals were recognized more accurately from the CCM samples, whereas sadness and anger were recognized more accurately from the classical singing technique. The correct perception of anger seemed to be clearly easier from female CCM samples than from any other samples.

Pitch played a role in emotion recognition. Sadness was more easily recognized from a low pitch (female voice 55.9%, male voice 55.2%) and less easily recognized from a high pitch (females 15.6%, males 24.1%). The recognition of joy was better from a high pitch (females 42%, males 28.1%) and more poorly from a low pitch (females 5.4%, males 0%). The recognition of tenderness was slightly better at a middle pitch (females 31.6%, males 41.4%). Anger was best recognized from high frequencies in all sample groups (Table 5).

We tested the internal consistency of the answers in the female samples at different pitches with Cronbach alpha and it showed a mean consistency of 0.60. However, the fluctuation of listener agreement between emotions was considerable (Cronbach −0.37 to 0.97). Anger yielded the most consistent answers, whereas joy yielded the least consistent answers.

When comparing the recognition of emotions from different pitches in the classical and CCM styles of singing, the most prominent difference was seen in the recognition of anger in the female classical and male CCM samples. Anger was perceived 26.6% units better at a high pitch than at a low pitch from the female classical samples and 55.2% units better at a high pitch than at a low pitch from the male CCM samples (Table 5).

TABLE 3.
Correctly Recognized Samples, Differences in Recognition Between CCM Styles of Singing and Classical Singing at Three Different Pitches, and the Internal Consistency of the Answers (Statistical Significance Level $\alpha \leq 0.05$)

			%	*z*-Value	*P* Value	Cronbach Alpha
Female	CCM	Overall recognition	35.2%	26.02	0.000	0.90
		Low pitch	34.7%	14.51	0.000	0.92
		Medium pitch	35.1%	14.90	0.000	0.88
		High pitch	35.9%	15.72	0.000	0.90
	Classical	Overall recognition	23.9%	5.76	0.000	0.88
		Low pitch	22.3%	1.94	0.052	0.93
		Medium pitch	24.2%	3.60	0.000	0.85
		High pitch	25.3%	4.48	0.000	0.82
Male	CCM	Overall recognition	27.4%	3.84	0.000	0.85
		Low pitch	22.1%	0.62	0.533	0.87
		Medium pitch	26.9%	2.08	0.000	0.85
		High pitch	33.1%	3.94	0.000	–
	Classical	Overall recognition	29.4%	4.91	0.000	0.80
		Low pitch	26.9%	2.08	0.038	0.93
		Medium pitch	29.7%	2.91	0.004	0.78
		High pitch	31.7%	3.53	0.000	–

–, Not enough answers to calculate Cronbach Alpha.

TABLE 4.
Correctly Recognized Emotions, Differences in Recognition Between CCM and Classical Singing, and the Internal Consistency of the Answers (Statistical Significance Level $\alpha \leq 0.05$)

			%	*z*-Value	*P* Value	Cronbach Alpha
Female	CCM	Joy	24.3%	3.36	0.001	0.90
		Tenderness	33.1%	10.15	0.000	0.78
		Neutral	29.5%	7.03	0.000	0.73
		Sadness	34.5%	11.2	0.000	0.89
		Anger	53.9%	26.23	0.000	0.95
	Classical	Joy	14.5%	−3.62	0.000	0.89
		Tenderness	13.4%	−4.38	0.000	0.56
		Neutral	28.7%	5.76	0.000	0.39
		Sadness	36.2%	10.69	0.000	0.86
		Anger	26.7%	4.43	0.000	0.95
Male	CCM	Joy	12.6%	−1.72	0.086	0.73
		Tenderness	31%	2.57	0.01	0.78
		Neutral	40.2%	4.75	0.000	0.66
		Sadness	34.5%	3.38	0.000	0.81
		Anger	18.4%	−0.38	0.707	0.97
	Classical	Joy	11.5%	−1.98	0.047	0.91
		Tenderness	27.6%	1.77	0.077	0.79
		Neutral	36.8%	3.91	0.000	−0.69
		Sadness	50.6%	7.13	0.000	0.76
		Anger	20.7%	0.16	0.872	−0.09

Valence and activation appraisals

Valence and activation were derived from the listeners' answers. Of the samples produced by female CCM singers, valence was correctly perceived as positive 887 times (46.3%) and as negative 997 times (52.1%). From the female classical samples, valence was correctly perceived as positive 402 times (28.9%) and as negative 698 times (50.1%). From the male CCM samples, valence was correctly perceived as positive 63 times (36.4%) and as negative 63 times (36.2%). From male classical samples, valence was correctly perceived as positive 57 times (32.8%) and as negative 75 times (43.1%).

Of the samples produced by female CCM singers, activation was correctly perceived as high 967 times (50.5%) and as low 1145 times (59.8%). From the female classical samples, activation was correctly perceived as high 583 times (41.9%) and as low 733 times (52.7%). From the male CCM samples, activation was correctly perceived as high 48 times (27.7%) and as low 102 times (58.6%). From the male classical samples, activation was correctly perceived as high 47 times (27%) and as low 120 times (69%).

In the answers given, valence was perceived with 41.6% accuracy and activity with 45.8% accuracy. High activity was perceived with 41.5% accuracy and low activity with 57.5% accuracy. Positive valence was perceived with 38.6% accuracy and negative valence with 50.2% accuracy.

When comparing the perceived accuracy of valence and activation between the CCM style and classical style, we can see that in the female samples, both valence and activation are more accurately perceived from the CCM-style samples, where, as with the male samples, they were better perceived compared to the classical-style samples (Table 5).

In these data, activation was more accurately perceived from all pitches in comparison to valence. At a low pitch, valence was more accurately perceived for joy and anger, whereas activation was more accurately perceived for tenderness and sadness. At a middle pitch, the tendency was similar to the female samples, but with the male samples, the valence was more accurately perceived only for joy. At a high pitch, activity was perceived more accurately for all other emotions except for female tenderness, in which valence was more accurately perceived. The perception of valence and activation from the female samples was most accurate at a high pitch. From the male samples, valence was correctly perceived from a middle pitch most accurately, whereas activation was most accurately perceived from a high pitch.

DISCUSSION

The percentage of correctly recognized emotion samples in this study was relatively low (30.2%) compared to earlier studies concerning speech. Most of the studies examining emotion recognition from the speaking voice reach recognition percentages above 50%.[1,20,25–27] Thus, it seems to be harder to recognize emotions from singing samples, at least when they are short.

Previous research suggests that the expression of emotions in the singing and speaking voice are related,[28] and that the same methods of emotion recognition apply to

TABLE 5.
Correct Recognition of Emotion, Valence, and Activation at Different Pitches

	Females		Males	
	CCM	Classical	CCM	Classical
Low pitch (220 Hz/110 Hz)				
Joy	8.20%	1.70%	0.00%	0.00%
Valence	29%	9.50%	17%	14%
Activation	17.20%	16.80%	7%	0%
Tenderness	24.50%	9.50%	37.90%	10.30%
Valence	26.30%	11.20%	38%	14%
Activation	68.00%	65.50%	72.40%	69%
Sadness	58.00%	53.00%	41.00%	69.00%
Valence	59.90%	60.30%	44.80%	59%
Activation	70.50%	63.40%	69.00%	86%
Anger	49.80%	15.60%	0.00%	13.80%
Valence	69.00%	63.40%	41%	31.00%
Activation	53.90%	17%	3%	24.10%
Middle pitch (330 Hz/165 Hz)				
Joy	16.60%	8.20%	13.80%	3.40%
Valence	40.80%	30.60%	65.50%	41.40%
Activation	30.70%	14.20%	14%	3%
Tenderness	39.50%	20.70%	41.40%	41.40%
Valence	52.40%	27.20%	41%	48.30%
Activation	64.90%	66.40%	75.90%	69%
Sadness	31%	38%	48%	48%
Valence	32.60%	40.10%	48%	51.70%
Activation	64.30%	61.20%	82.20%	65.50%
Anger	56.10%	22.40%	0%	20.70%
Valence	65.80%	48.30%	10%	28%
Activation	64.30%	30.20%	14%	35%
High pitch (440 Hz/220 Hz)				
Joy	48.30%	33.60%	24.10%	31%
Valence	58.10%	40.90%	32.10%	37.90%
Activation	63.80%	56.50%	60.70%	37.90%
Tenderness	35.40%	9.90%	13.80%	31%
Valence	71.20%	53.90%	24.10%	41%
Activation	48.60%	28.90%	41.40%	58.60%
Sadness	14%	17%	14%	35%
Valence	18.20%	30.20%	17%	37.90%
Activation	42.60%	30.60%	20.70%	65.50%
Anger	56.10%	42.20%	56%	27.60%
Valence	67.10%	58.60%	55%	41.40%
Activation	73.00%	61%	69%	62.10%

both.[29] As with the speaking voice, the voice quality in anger is easiest to recognize. This phenomenon might have an evolutionary underpinning, as it continues to be a useful skill to recognize potentially hazardous situations.

In these data, emotional content from the CCM-style samples was correctly recognized 11.3% more often than from the classical-style samples. The recognition of sadness, however, was 3.3% units higher from the samples sung using the classical techniques. It could be postulated that the darker timbre typical in classical singing makes it easier to be interpreted as related to sadness. The dark timbre in classical singing is due to lower resonance frequencies related to the lowering of the larynx and, thus, the lengthening of the vocal tract.[30] According to earlier studies, strong, fundamental, relatively weak overtones near 3 kHz and lack of vibrato have been found to be indicative of typical expressions of sadness in the Western classical singing style.[17,31,32] However, many samples that were recognized as expressions of sadness in the present study had a very clear vibrato. Another possible explanation for the results of the present investigation (made in Finland) could be cultural. The Russian lament uses a simultaneous amplitude and frequency modulation that is reminiscent of vibrato. Another factor could be the distinct distribution of energy during one vowel, where intensity increases toward the end: this technique is also used in laments.[33] Further investigation is needed to examine the acoustic structure of the voice samples in this study.

The speech-like qualities of the CCM style of singing are one possible explanation as to why it seems to be easier to recognize emotion portrayals from it. Another possibility as to why the CCM style of singing is more recognizable is that it uses the "chest voice" more often, whereas classical singing operates more with the "head voice." In the chest voice, the mass of the vocal folds vibrates more vertically, making a more robust impact on air pressure, and the formants appear easier.[34] The slope of the sound spectrum is more gradual, as the relative amplitude of the upper partials is more pronounced. In the head voice, the slope is steeper.[35] It is also plausible that speakers and CCM-style singers use more variation in phonation type along the axis from breathy to pressed,[36] whereas classical singers keep the voice source more stable. This is related to both esthetic and technical demands. For instance, pressed phonation may make it more difficult or even impossible to reach the high pitches required.

Another possible explanation for the better identification of emotion in CCM-style singing is familiarity. In general, most people are exposed to far more CCM than classical singing. Therefore, they may be more attuned to emotion in these genres.

Previous research has shown that there is a considerable variability in the individual ability to express emotion by singing.[16] Mirroring this fact, the two male singers who participated in this study were too few to properly represent male CCM and classical singers. However, we wanted to keep them in the study because despite their individuality, the listening test appraisals were mostly very similar to those given for the female samples. The individual ability to express emotions was also seen in the female samples. When we excluded the two best recognized CCM female singers from the sample size, the groups of all CCM singers ($N = 11$) and CCM singers -2 ($N = 9$) were no longer statistically a part of the same group (Pearson chi-squared test of homogeneity). This suggests that in future studies, the sample size should be fairly large to increase validity.

Pitch seemed to affect the assessment of emotions. As in the speaking voice, the higher the pitch, the more often the

listeners chose an emotion that represents a high overall activity level (Tables 5). This is understandable, because a higher pitch is typically produced with higher subglottic pressure and thus intensity.[6,35] This phenomenon is potentially counterproductive for singers needing to portray nonactive negative emotions (like sorrow) at a high frequency or high-activation positive emotions (like joy) at a low frequency. The tendency not to recognize joy but to recognize sadness was very pronounced at a low pitch (220/110 Hz) for both the female and male samples. At a high pitch (440 Hz), the phenomenon was reversed.

The pitches were selected with the female singers in mind from a pitch range that would allow singing the whole eight-bar song in either the chest or the head register, should the singer choose to express it so. This was done to accommodate various singing styles and make as much room as possible for emotional expression while gaining data from the same pitches. It is possible that the choice to use the same pitches for all subjects and both singing styles may have somewhat interfered with the results, because the pitch range was somewhat low for classical singing. On the other hand, the participants were trained singers who ought to be well able to sing at these pitches.

The accuracy of perceived valence and activity in the listening test answers may suggest that it is easier to make assessments of valence and activity than to recognize emotions *per se*. This corresponds to the earlier findings for speech. Similarly, we found in the female samples that samples with a negative valence and high activity were more easily recognized than those with a positive valence and low activity. This is understandable, because it is important for survival to be able to quickly recognize signs of potentially dangerous situations.

CONCLUSIONS

Our study offers the following conclusions. Emotions were recognized above the level of chance in short vowel extracts from singing, and emotions were recognized statistically significantly better in the samples with a CCM style of singing compared to the samples featuring classical singing. Furthermore, pitch also plays a role in emotion recognition in the singing voice, and the valence and activation levels of the voice also play a role in emotion recognition in singing.

Acknowledgments

This research was supported by the Eemil Aaltonen Foundation through a grant (160036 N1) and by the Oskar Öflunds Stiftelse Foundation through a grant to Tua Hakanpää. The authors would like to thank Antti Poteri, D.Sc. (Tech), for help with the statistical analysis.

REFERENCES

1. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*. 1996;70:614–636. https://doi.org/10.1037/0022-3514.70.3.614.
2. Laukkanen A-M, Vilkman E, Alku P, et al. On the perception of emotions in speech: the role of voice quality. *Logop Phoniatr Vocol*. 1997;22:157–168. https://doi.org/10.3109/14015439709075330.
3. Tartter VC. Happy talk: perceptual and acoustic effects of smiling on speech. *Percept Psychophys*. 1980;27:24–27. https://doi.org/10.3758/BF03199901.
4. Hallqvist H, La FMB, Sundberg J. Soul and musical theater: a comparison of two vocal styles. *J Voice*. 2016. https://doi.org/10.1016/j.jvoice.2016.05.020.
5. Björkner E. Musical theater and opera singing—why so different? A study of subglottal pressure, voice source, and formant frequency characteristics. *J Voice*. 2008;22:533–540. https://doi.org/10.1016/j.jvoice.2006.12.007.
6. Titze IR. *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice-Hall; 1994.
7. Williams CE, Stevens KN. Emotions and speech: some acoustical correlates. *J Acoust Soc Am*. 1972;52:1238–1250. https://doi.org/10.1121/1.1913238.
8. Chua G, Chang QC, Park YW, et al. The expression of singing emotion—contradicting the constraints of song. In: *Proceedings of 2015 International Conference on Asian Language Processing, IALP 2015*. 201698–102. https://doi.org/10.1109/IALP.2015.7451541.
9. Juslin PN, Laukka P. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol Bull*. 2003;129:770–814. https://doi.org/10.1037/0033-2909.129.5.770.
10. Laver J. *The Phonetic Description of Voice Quality*. Cambridge University Press; 1980.
11. Manfredi C, Barbagallo D, Baracca G, et al. Automatic assessment of acoustic parameters of the singing voice: application to professional western operatic and jazz singers. *J Voice*. 2015;29:517.e1–517.e9. https://doi.org/10.1016/j.jvoice.2014.09.014.
12. Schloneger MJ, Hunter EJ. Assessments of voice use and voice quality among college/university singing students ages 18−24 through ambulatory monitoring with a full accelerometer signal. *J Voice*. 2017;31:124.e21–124.e30.
13. Barlow C, LoVetri J. Closed quotient and spectral measures of female adolescent singers in different singing styles. *J Voice*. 2010;24:314–318.
14. Sundberg J, Lã FMB, Gill BP. Formant tuning strategies in professional male opera singers. *J Voice*. 2013;27:278–288. https://doi.org/10.1016/j.jvoice.2012.12.002.
15. Borch DZ, Sundberg J, Lindestad PA, et al. Vocal fold vibration and voice source aperiodicity in "dist" tones: a study of a timbral ornament in rock singing. *Logoped Phoniatr Vocol*. 2004;29:147–153. https://doi.org/10.1080/14015430410016073.
16. Siegwart H, Scherer KR. Acoustic concomitants of emotional expression in operatic singing: the case of Lucia in Ardi gli incensi. *J Voice*. 1995;9:249–260. https://doi.org/10.1016/S0892-1997(05)80232-2.
17. Scherer KR. Expression of emotion in voice and music. *J Voice*. 1995;9:235–248. https://doi.org/10.1016/S0892-1997(05)80231-0.
18. Ekman P. Are there basic emotions? *Psychol Rev*. 1992;99:550–553. https://doi.org/10.1037/0033-295X.99.3.550.
19. Izard CE. Basic emotions, relations among emotions, and emotion-cognition relations. *Psychol Rev*. 1992;99:561–565. https://doi.org/10.1037//0033-295X.99.3.561.
20. Waaramaa T, Laukkanen AM, Alku P, et al. Monopitched expression of emotions in different vowels. *Folia Phoniatr Logop*. 2008;60:249–255. https://doi.org/10.1159/000151762.
21. Airas M, Alku P. Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*. 2006;63:26–46. https://doi.org/10.1159/000091405.
22. Boersma P, Weenink D. *Praat*. 2014.
23. Sundberg J. Emotive transforms. *Phonetica*. 2000;57:95–112.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174. https://doi.org/10.2307/2529310.
25. Scherer KR. Vocal communication of emotion: a review of research paradigms. *Speech Commun*. 2003;40:227–256. https://doi.org/10.1016/S0167-6393(02)00084-5.

26. Scherer KR, Banse R, Wallbott HG. Emotion inferences from vocal expression correlate across languages and cultures. *J Cross Cult Psychol*. 2001;32:76–92. https://doi.org/10.1177/0022022101032001009.
27. Iida A, Campbell N, Higuchi F, et al. A corpus-based speech synthesis system with emotion. *Speech Commun*. 2003;40:161–187. https://doi.org/10.1016/S0167-6393(02)00081-X.
28. Scherer KR, Sundberg J, Tamarit L, et al. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Comput Speech Lang*. 2015;29:218–235. https://doi.org/10.1016/j.csl.2013.10.002.
29. Eyben F, Salomão GL, Sundberg J, et al. Emotion in the singing voice—a deeper look at acoustic features in the light of automatic classification. *J Audio Speech Music Proc*. 2015;2015:19. https://doi.org/10.1186/s13636-015-0057-6.
30. Richard M. *The Structure of Singing, System and Art in Vocal Technique*. Belmont, CA: Wadsworth Group; 1996.
31. Sundberg J. Expressivity in singing. A review of some recent investigations. *Logop Phoniatr Vocol*. 1998;23:121–127. https://doi.org/10.1080/140154398434130.
32. Jansens S, Bloothooft G, de Krom G. Perception and acoustics of emotions in singing. *Proc 5th Eur Conf Speech Commun Technol*. 1997; IV:2155–2158. Available at; http://www.isca-speech.org/archive/archive_papers/eurospeech_1997/e97_2155.pdf.
33. Mazo M, Erickson D, Harvey T. Emotion and expression: temporal data on voice quality in Russian lament. *Vocal Fold Physiology Voice Quality Control*. San Diego, CA: Singular Publishing; 1995:173–187.
34. Titze IR, Martin DW. Principles of voice production. *J Acoust Soc Am*. 1998;104:1148. https://doi.org/10.1121/1.424266.
35. Sundberg J. *The Science of the Singing Voice*; 1987.
36. Peterson KL, Verdolini-Marston K, Barkmeier JM, et al. Comparison of aerodynamic and electroglottographic parameters in evaluating clinically relevant voicing patterns. *Ann Otol Rhinol Laryngol*. 1994;103(5 pt 1):335–346. https://doi.org/10.1177/000348949410300501.

PUBLICATION II

Comparing Contemporary Commercial and Classical Styles: Emotion expression in singing

Tua Hakanpää, Teija Waaramaa, and Anne-Maria Laukkanen

Original publication: Journal of Voice 2021, 35(4)570-580
https://doi.org/10.1016/j.jvoice.2019.10.002

Comparing Contemporary Commercial and Classical Styles: Emotion Expression in Singing

Tua Hakanpää, Teija Waaramaa, and Anne-Maria Laukkanen, *Tampere, Finland*

Summary: Objective. This study examines the acoustic correlates of the vocal expression of emotions in contemporary commercial music (CCM) and classical styles of singing. This information may be useful in improving the training of interpretation in singing.

Study Design. This is an experimental comparative study.

Methods. Eleven female singers with a minimum of 3 years of professional-level singing training in CCM, classical, or both styles participated. They sang the vowel [ɑ:] at three pitches (A_3 220Hz, E_4 330Hz, and A_4 440Hz) expressing anger, sadness, joy, tenderness, and a neutral voice. Vowel samples were analyzed for fundamental frequency (*fo*) formant frequencies (F1-F5), sound pressure level (SPL), spectral structure (alpha ratio = SPL 1500-5000 Hz—SPL 50-1500 Hz), harmonics-to-noise ratio (HNR), perturbation (jitter, shimmer), onset and offset duration, sustain time, rate and extent of *fo* variation in vibrato, and rate and extent of amplitude vibrato.

Results. The parameters that were statistically significantly (RM-ANOVA, $P \leq 0.05$) related to emotion expression in both genres were SPL, alpha ratio, F1, and HNR. Additionally, for CCM, significance was found in sustain time, jitter, shimmer, F2, and F4. When *fo* and SPL were set as covariates in the variance analysis, jitter, HNR, and F4 did not show pure dependence on expression. The alpha ratio, F1, F2, shimmer apq5, amplitude vibrato rate, and sustain time of vocalizations had emotion-related variation also independent of *fo* and SPL in the CCM style, while these parameters were related to *fo* and SPL in the classical style.

Conclusions. The results differed somewhat for the CCM and classical styles. The alpha ratio showed less variation in the classical style, most likely reflecting the demand for a more stable voice source quality. The alpha ratio, F1, F2, shimmer, amplitude vibrato rate, and the sustain time of the vocalizations were related to *fo* and SPL control in the classical style. The only common independent sound parameter indicating emotional expression for both styles was SPL. The CCM style offers more freedom for expression-related changes in voice quality.

Key Words: Emotion expression—Singing voice—Voice quality—Song genre—Acoustic analyses.

INTRODUCTION

Emotions and their expression lie at the core of human life and thought. We are fine-tuned to infer strong emotional messages from our surroundings by nature as an evolutionary coping mechanism.[1] The communication of emotions may be realized both verbally and nonverbally through gestures of the voice and/or body.[2] The way we infer emotional meaning from sound is through sound quality and other prosodic sound variables.[3-5] Multimodal similarities in emotion expression back the assumption that all speech and music have developed from the same origin.[3] The voice is a particularly potent instrument for emotional expression because of its evolutionary background. Acoustic similarities in how emotions are expressed in the speaking voice and in music might explain why music is perceived to be emotional. However, the precision with which music can convey different emotions is limited. There is usually a high agreement within listeners about the broad emotional category (such as sad or angry music), but less agreement concerning the nuances within this category (eg, somber, melancholy, disappointed, distraught, devastated).[3] Research has shown that basic emotions are cross-culturally the easiest to express and perceive in music. It is hypothesized that this is because they have precise expressive characteristics in other verbal and nonverbal channels, such as the speaking voice and human gestures.[3]

The concept of emotion is usually organized through categorical or dimensional emotion models. Categorical models see a number of distinct (discrete) emotion systems in humans, each with its own adaptive behavioral function. These systems usually include so-called basic emotions, such as anger, joy, sadness, fear, disgust, and surprise.[6] According to the dimensional model, emotions form a system where differences of degree can be seen in certain basic dimensions, such as activity/arousal or valence.[7] Furthermore, according to social constructionist theories, emotions are products of culture that are formed by the influence of culture on culture. Therefore, emotions are not seen to have a biological basis; rather, human emotions are always social constructs that serve the general purposes of society. In addition, it is thought that the expression of emotions is governed by predetermined roles in society and the status of the individual in the community in which the emotion occurs.[8] Research leaning toward the assumption of the existence of basic emotions in music is continuously being challenged and criticized by the constructionist account for a lack of precision and evidence. The constructionists

Accepted for publication October 2, 2019.

This research was supported by the Eemil Aaltonen Foundation through a grant (170036 N1) to Tua Hakanpää.

Declarations of interest: None.

From the Speech and Voice Research Laboratory, Faculty of Social Sciences, Tampere University, Tampere, Finland.

Address correspondence and reprint requests to Tua Hakanpää, Speech and Voice Research Laboratory, Tampere University, Åkerlundinkatu 5, Tampere 33014, Finland E-mail: Tua.Hakanpaa@tuni.fi

Journal of Voice, Vol. 35, No. 4, pp. 570−580

0892-1997

https://doi.org/10.1016/j.jvoice.2019.10.002

propose that the phenomenon of decoding emotional meaning from music arises from the interplay of music's ability to express core affects (arousal and valence) and the influence of contextual information in the listener's mind.[9] Valence (the intrinsic attractiveness or averseness of an object, situation, or event) and activity/arousal (a state in which our bodies experience heightened physiological activity) are the two most commonly studied dimensions of vocally conveyed emotion expression.[10]

Emotions are encoded into music in various ways, including contour, modality, tempo, rhythm, pitch, and intensity.[11] These are all attributes that make it harder for singers to encode emotional expression with parameters normally used in the speaking voice. Fundamental frequency (*fo*) sound pressure level (SPL) and duration are the most prevalent indicators of emotions in the speaking voice,[12] and their usage is severely hampered by the constraints of written music. A recent study by Hakanpää et al showed that pitch plays a role in emotion recognition in the singing voice. Low pitches are more easily recognized as low intensity emotions (such as sadness and tenderness) and high pitches as high intensity emotions (such as joy or anger), regardless of the emotion the singer intends to express.[13] However, there are other attributes of voice production that can compensate for the reduction of expressivity in phonation frequency, particularly voice quality and the characteristics of vibrato.[14] The way we change the proportion of energy in the higher and lower part of the spectrum, the amount of noise in the sound, and different phonatory perturbations all contribute to different voice qualities that can be interpreted as emotional expression. Perturbation refers to the small period-to-period changes in period length (jitter) and amplitude (shimmer). All natural voice sounds have some irregularity in them. Harmonics-to-noise ratio (HNR) measures in dB the difference between the partials and the noise component left between them. When the HNR is small, the voice sounds breathy or hoarse. The greater the HNR is, the clearer the voice sounds. The amplitude contour—the short-term variability of sound level—has been found to be an important indicator of emotional content in the singing voice in short phrases.[15] Vibrato refers to the more or less symmetric oscillation of *fo* around its center value on a tone. Voice amplitude typically varies together with *fo* in vibrato.[16,17]

Voice quality is a term that refers to the auditory perception of an individual's characteristic way of using his/her own voice.[18] Acoustic voice quality is determined by the individual anatomy of the voice organ[19] on the one hand, and the way it is used on the other hand. Vocal fold adduction, the symmetry of the vocal folds, muscle tension in the larynx, the size and shape of the vocal tract, and the amount of air escaping to the nasal cavities all affect the voice quality. By changing the natural balance of these factors, it is possible to change the habitual voice quality. Voice quality is always determined according to the norm that we call "neutral." An individual's neutral voice is determined by his/her anatomical and physiological features.[18] Changes are made to the voice quality to match the aesthetic demands of the singing style. Thus, the voice quality for opera is different from the voice quality for pop or rock music. Emotions are psychophysiological events that can cause the voice quality to change.[2] Emotion expression in singing can be produced by imitating the voice changes that naturally occur during an emotional episode.

The ability to code emotion into the voice varies substantially from person to person.[20] Scherer et al[1] state that the central particularly expressive acoustic elements concerning the singing voice are loudness, dynamics, perturbation variation, low-frequency energy, formant amplitude, and *fo*. Eyben et al[21] found that automatized (singing voice) emotion detectors recognize pitch, jitter/shimmer, spectral band, and the mel-frequency cepstral coefficient (MFCC). Mayor et al[22] also list pitch, energy, spectral coefficient, and MFCC and its derivatives as expressive resources that uniquely identify sung expressions of emotion. Vibrato and attack, sustain, and release are also important nominators of expression in the singing voice. In studies investigating the classical singing voice, dynamic features in spectral energy, vibrato extent, and voice quality have been recognized as emotion-carrying components.[23,15] Both dynamic variation and micro variations in tempo have been reported to be prominent indicators of emotional coding in the singing voice.[24,25]

One way of differentiating emotion expressions in singing is to sort them into categories according to their activity level. High activity emotions (such as anger and joy) are characterized by a fast tempo, loud voice or quick changes in loudness, great extent of vibrato, and local departures from the pitch contour at tone onsets.[21,24] By contrast, low activity emotions (such as sadness and tenderness) are characterized by a slow tempo, soft voice, and fewer changes in loudness.[21,24]

Sung expressions of joy are described as fast in tempo and perceptually loud with a lot of short-term amplitude variation and a shallow spectral slope. Scherer et al describe joy as having a low formant bandwidth and formant amplitude but a high formant frequency. Expressions of joy have a high perturbation level (mean jitter, shimmer, and HNR) but a low perturbation variation (standard deviation normalized by the arithmetic mean) and more vibrato (extent and rate) than low activity emotions. They are highest in *fo* variation.[1,15,20,26] Joy is acoustically similar to anger. It has a heightened intensity level, a shallow spectral slope that indicates stronger vocal fold closure, and a bright sound quality that can be induced by the smiling activity of the face (retraction of the lips shortens the vocal tract and thus raises the formant frequencies).[27,28]

Tenderness is a low activity emotion like sadness. It has been described as being characterized by a slow tempo, low vocal energy, and low dynamics.[1,29] Overall, tenderness features low perceptual loudness, but, as Sundberg et al[15] report less short-term variability of the sound level, while

Scherer et al[1] report high loudness(ae)[1] variation. Tenderness has a broad formant bandwidth and low formant frequencies with a tendency to have more sound energy at the low frequency range of the spectrum. It is characterized by a low perturbation level and high perturbation variation, and it has been reported to have less vibrato compared to high activity emotions.[1,15,26,30]

Sadness is a low activity emotion described as having a slow tempo and low level of dynamics. Expressions of sadness are portrayed at a low loudness level, but Sundberg et al[15] and Scherer et al[1] report different results for loudness variation (see above). This discrepancy could be due to individual differences in voice usage and/or culture, as Sundberg et al[15] analyzed the voice samples of one baritone singer, while Scherer et al[1] analyzed samples from eight classical singers—both female and male—and the instruction of how to portray those emotions differed in the two studies. Sadness is high on formant bandwidth and low on formant amplitude with very little energy variation at the low frequencies. It has been reported to be low in perturbation level and high in perturbation variation. Jansens et al[20] report the absence of vibrato in the expression of sadness. Sundberg describes sadness as having slow tone onsets, a soft tone, a dominant fundamental and weak overtones near 3 kHz, low subglottal pressure, and a low degree of glottal adduction.[25] Other reported correlates of sadness include a slower articulation rate and unclear articulation, lower intensity, perturbation in the phonation, and larger irregularities in the periodicity of vocal fold closure.[27,31]

Anger is recognized as a high activity emotion with a fast tempo, high mean sound level, and high dynamics.[30,29] Scherer et al[1] report low loudness(ae) variation, while Sundberg et al[15] report more short-term variability of the sound level. Anger is low on formant bandwidth and formant amplitude, and it has quite a lot of energy variation at low frequencies. Scherer et al[30] found a flat, highly balanced spectrum indicating strong energy in the higher partials, and Scherer et al[29] report weak, low frequency energy. Expressions of anger are high on the perturbation level but low in perturbation variation, and many studies report an increase of vibrato (rate and extent).[15,20,21] Anger has been reported to have faster syllable onsets and faster decays of sound pressure. Other correlates of anger include sharp and tightened articulation and an overall rise in the intensity level.[1,4 27,31,32]

Although there is a growing body of research on emotional expression in the singing voice, most studies deal with the Western classical style of singing, and in that category focus on operatic singing.[1,15,20,21,23-25,29,30] There are still few comparative studies examining the differences between singing styles in the expression of emotions. The present study compares the vocal expression of emotions in contemporary commercial music (CCM) and Western classical styles in female singers. The indicators of singing styles (classical and CCM) in this study are used as a generic descriptor of many styles with related origins. We acknowledge that classical singing can encompass several singing styles with different sound ideals—such as Baroque, Renaissance, operetta, modern music, verismo, Sperchgesang, bel canto, and legit—and CCM singing also encompasses several singing styles with different sound ideals—such as pop, rock, R&B, soul, blues, country, folk, and jazz. In order to investigate the role of voice quality in conveying emotions, we focus on the short vowel [ɑ:], because short vowels do not contain semantic or prosodic information. Therefore, they carry voice quality in its purest sense. Earlier research on both the speaking and the singing voice suggests that emotional information can be received from a short sample.[32-34] We examine the acoustic correlates of sung vowel segments expressing anger, joy, sadness, tenderness, and a neutral state in singing. These emotions have been selected because of their opposing placement on the valence-activation scale, and also because they are categories of emotion frequently encountered in the song literature.

[1]Loudness(ae) is a psychoacoustic measure that is designed to give a value to acoustically estimated loudness. According to Scherer et al. (2017), this correlates better with the vocal affect dimensions than with the raw signal energy.

METHODS

Samples

The voice samples were gathered from 11 Finnish female singers: six with classical training and five with a CCM background (mean age 32 years, mean singing experience 10 years, minimum of 3 years of singing lessons at a professional level). All of the singers were portfolio-type singers (singing multiple styles of CCM and/or classical) who performed at regional or local venues.[35] The singers sang an eight-bar excerpt from Gershwin's "Summertime," expressing emotions of joy, tenderness, sadness, anger, and a neutral state. They repeated the task three times. The vowel [ɑ:] was extracted from three different pitches (220 Hz, 330 Hz, and 440 Hz) in each sample for further analysis. Recordings were made in a sound-treated studio and they were calibrated for SPL measurements using a sinewave generator and sound level meter (a more detailed description of the samples and recordings can be found in Hakanpää et al[13]). The singers were instructed to use either classical or CCM singing technique depending on their orientation as a singer (Table 1). Six of the singers worked in both classical and CCM genres, and these subjects gave voice samples in both styles. No further instructions about the music style were given in order to ensure that the participants would direct as much of their attention to the task of emotion expression as possible. The samples were later evaluated by two known Finnish crossover singing teachers in the field, who confirmed them to be adequate expressions of the generic styles of the classical and CCM singing techniques.

Acoustic analyses

Twenty acoustic parameters were automatically extracted from the voice samples (N = 270; classical n = 120, CCM n = 150). All analyses were made using Praat software

(version 6.0.19). All of the samples were analyzed for *fo* using autocorrelation with a bandwidth set to 75-600 Hz. SPL was measured for the calibrated signals. Formant frequencies F1-F5 were measured up to 5500 Hz.

Two measures of jitter were used: relative average perturbation (RAP), defined as the average absolute difference between an interval (glottal period) and the average of it and its two neighbors, divided by the average time between two consecutive points; and the five-point period perturbation quotient (ppq5), the average absolute difference between an interval and the average of it and its four closest neighbors, divided by the average time between two consecutive points.

For shimmer, the three-point amplitude perturbation quotient (shimmer apq3; the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude) and the five-point amplitude perturbation quotient (apq5; the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbors, divided by the average amplitude) were measured. The alpha ratio was measured by subtracting the mean SPL of lower frequencies from that of higher frequencies in the spectrum: (SPL 1500-5000 Hz)—(SPL 50-1500 Hz). HNR was also measured.[16,36]

The vibrato rate (number of *fo* undulations per second) and extent (how far *fo* departs from its average during a vibrato cycle) were measured, as were the rate and extent of the amplitude vibrato. We also looked into how long it takes from the moment the sound initiates until it reaches its 90% peak level (attack time) and the time that the note is sustained at a 25%-75% dB level in comparison to the maximum level (sustain time), and we calculated the time when the SPL drops to zero from the 25% level of the SPL (release time). A timeframe of 0.01 second was used for sampling.

We measured the envelope of the sound samples extracting measurements on a 0.01-second timeframe (time t [s], f0 [Hz], SPL [dB]). For every sample, we searched for the timeframe where t_imax = 9/10*max(SPL). Attack time and level were set by choosing the smallest timeframe, t_min, where 90% of the maximum SPL level is reached. Then we looked for the timeframe where t_imax = 3/4*max(SPL), we chose the largest timeframe, t_max1, where 75% of the maximum SPL is reached. Sustain time is the subtraction result of the last possible moment where SPL 75%*max(SPL), and attack time difference with a level of 75%*max(SPL), sustain = {t_max1-t_min, 0.9*max(intens)}. Lastly, we extrapolated the timeframes when t_max2, when the SPL is zero and chose the release time the largest timeframe t_max2 release = {max(t_max2), 0}.

Statistical analyses

Mean and median values were calculated for *fo*, SPL, formant frequencies F1-F5, alpha ratio, jitter (rap & ppq5), shimmer (apq3 & apq5), HNR, vibrato rate and extent, amplitude vibrato rate and extent, and duration of attack, sustain, and release.

To evaluate whether the parameter values extracted with Praat differed across emotions for each parameter, we computed a repeated measures analysis of variance (RM-ANOVA) of the general linear model (GLM) with SPSS (v.17; SPSS Inc., Chicago, IL). Due to the small number of participants and the fact that some participants sang in both genres, we ran the RM-ANOVA separately for the classical and CCM styles. Each parameter was set as a within-subjects variable at a time and pitch. Pitch (on a scale 1 = low, 2 = middle, 3 = high) was set as a between-subjects factor to help determine if it had a significant effect on the variation of the parameters. Least significant difference and Bonferroni corrections were used.

TABLE 1.
The Participants

Participant	(Original) Orientation	Professional Status	Repertoire	Years of Experience
1	Classical	Student	Chamber music, baroque, contemporary music, opera	3
2	Classical	Professional (cantor)	Religious music, pop, classical	24
3	Classical	Professional (teacher)	Chamber music, baroque, contemporary music, opera, pop, rock, jazz, musical	14
4	Classical	Student	Chamber music, baroque, contemporary music, opera	3
5	Classical	Professional (teacher)	Opera, lied, rock, pop	6
6	Classical	Professional (teacher)	Pop, rock, soul, jazz, opera	14
7	CCM	Student	Pop, rock, R&B, jazz, folk	3
8	CCM	Student	Pop, rock, R&B, jazz, folk	5
9	CCM	Professional (performer)	Jazz, pop, chanson, classical	20
10	CCM	Professional (teacher)	Pop, R&B, jazz, folk	15
11	CCM	Student	Pop, rock, R&B, jazz, folk	4

Sphericity could not be assumed for all the data (Mauchly's test of sphericity), and for those parameters where the test was significant ($P < 0.05$), the Greenhouse-Geisser correction was used to obtain the results.

It is known that SPL and *fo* have an effect on different sound parameters such as perturbations, HNR, and alpha ratio. To further validate the results obtained from the RM-ANOVA, we used the univariate analysis of the GLM to determine if these parameters varied individually according to emotion and not only in relation to F0 and SPL. In the univariate analysis, each parameter was in turn set as a dependent variable, emotion was set as a fixed factor, and F0 and SPL were set as co-variates.

RESULTS

F0

No clear indication of emotion-related differences in the fine-tuning of *fo* were found. Classical singers sang slightly closer to the target pitches. Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for the CCM singers ($\chi^2(2) = 53, 79$, $P = 0.001$), but not for the classical singers ($\chi^2(2) = 11{,}032$, $P = 0.275$). The emotive singing did not lead to any statistically significant changes in *fo* over different emotion portrayals (CCM: F (1.845, 49.812) = 1.155, $P = 0.32$; classical: F (4, 84) = 1.983, $P = 0.105$).

SPL

SPL increased on the continuum of tenderness → sadness/neutral → joy → anger. If we look at the results from a dimensional perspective, it can be said that emotions of lower activity were sung with a lower volume compared to emotions of higher activity. Pitch had a strong effect on the overall SPL (SPL increased with *fo*), but according to our data, this did not interfere with using SPL adjustment as a means for communicating emotion. The classical style samples were sang slightly louder and with more variation in the loudness when compared to the CCM samples (Table 3). SPL differed significantly across the emotional expressions for both the CCM and classical singers (Table 4).

Formants

The overall formant structure of the samples revealed a few distinctive patterns in regard to emotion expression (Figure 2). For sadness, formants F1 and F2 were low, F3 low/relatively low, and F4-F5 relatively high in comparison to the other emotions and the neutral voice. For anger, the opposite was true: F1-F3 were high and F4-F5 packed lower in both genres. A similar but less pronounced formant structure could be found for joy. For tenderness, the first formant was positioned slightly higher than in sadness, but still relatively low; other formants were positioned fairly high. In the neutral voice, the first formant was neither high nor low, while the second and the third formants were relatively low. The formant structure was most compact for anger and then it scattered in the sequence of anger → joy → neutral → tenderness → sadness (Table 2). F1 was found to be statistically significant in differentiating emotion in both the CCM and classical styles. The same applied to F2 in the CCM style of singing, but F2 correlated with the pitch in both genres. In the CCM style, F4 was found to be statistically significant in differentiating emotional expressions (Table 4).

F1-F2 were positioned higher in the CCM style compared to the classical style (Figure 1). Formants also scattered differently in the CCM style compared to the classical style. In the CCM samples, the structure from the most compact to the most sparse was anger → joy → neutral → tenderness → sadness. In the classical samples, the structure from the most compact to the most sparse was neutral → tenderness → joy → anger → sadness (Table 2).

TABLE 2.
Mean Formant Frequencies for Joy, Tenderness, Sadness, Anger, and Neutral Voice

		F1	F2	F3	F4	F5
		Joy				
CCM	Low pitch	787	1453	2970	3754	4266
	Middle pitch	839	1525	2894	3717	4343
	High pitch	916	1482	2756	3674	4210
Classical	Low pitch	632	1283	2946	3722	4659
	Middle pitch	693	1253	2872	3544	4310
	high pitch	774	1358	2887	3575	4532
		Tenderness				
CCM	Low pitch	773	1433	2991	3887	4350
	Middle pitch	753	1450	2918	3854	4314
	High pitch	779	1385	2732	3683	4309
Classical	Low pitch	599	1342	2938	3754	4651
	Middle pitch	644	1166	2947	3602	4303
	High pitch	744	1273	2851	3684	4395
		Neutral				
CCM	Low pitch	707	1391	2944	3809	4397
	Middle pitch	757	1431	2940	3791	4207
	High pitch	825	1419	2683	3669	4227
Classical	Low pitch	641	1236	2870	3755	4681
	Middle pitch	659	1182	2926	3595	4320
	High pitch	765	1311	2881	3609	4284
		Sadness				
CCM	Low pitch	714	1409	3007	3894	4428
	Middle pitch	719	1412	2929	3709	4192
	High pitch	773	1362	2582	3646	4291
Classical	Low pitch	584	1212	2888	3749	4714
	Middle pitch	663	1200	2911	3578	4472
	High pitch	716	1302	2951	3743	4410
		Anger				
CCM	Low pitch	809	1351	2890	3694	4195
	Middle pitch	873	1454	2936	3617	4251
	High pitch	1003	1533	2851	3650	4290
Classical	Low pitch	654	1177	2919	3744	4642
	Middle pitch	750	1280	2998	3644	4418
	High pitch	791	1391	2917	3648	4597

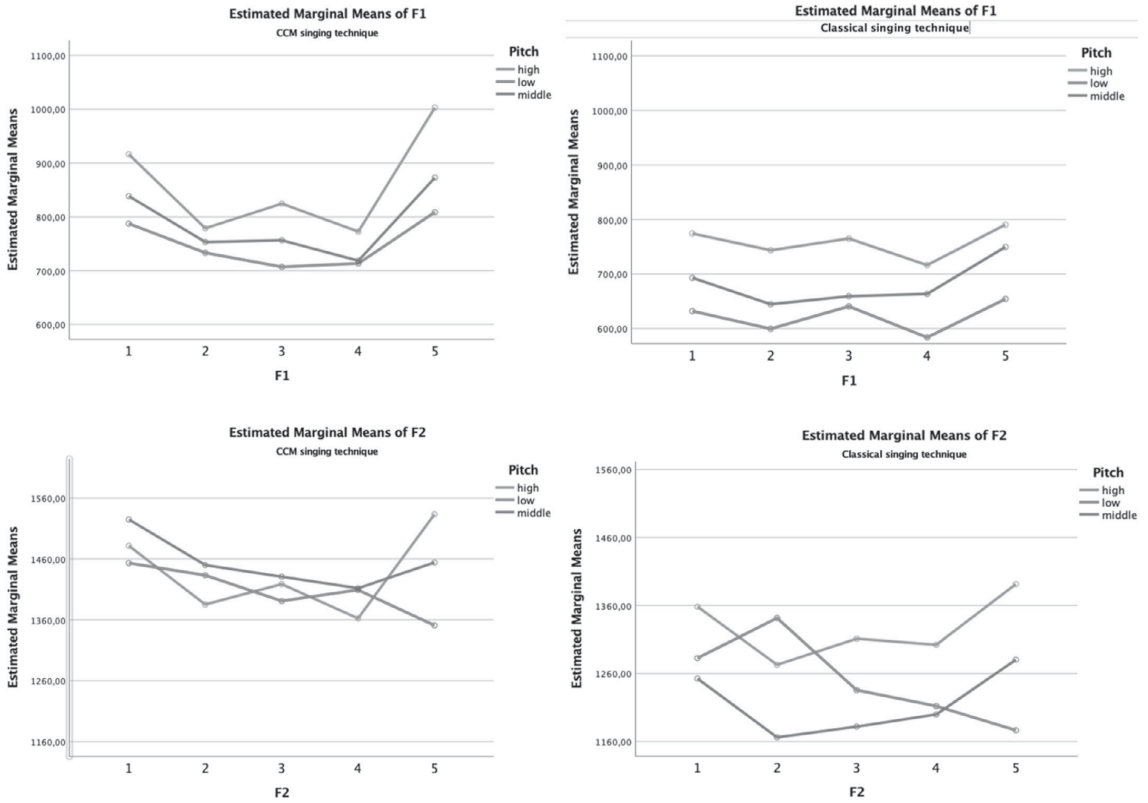

FIGURE 1. Formant positioning (F1 and F2) in the CCM and classical singing styles (1 = joy, 2 = tenderness, 3 = neutral, 4 = sadness, 5 = anger).

Alpha ratio

In almost all cases, the alpha ratio increases (smaller negative absolute value) when singing a higher pitch. All of the samples of the high activity emotions (joy and anger) were characterized by a larger alpha ratio when compared to the low activity emotions (sadness and tenderness). The alpha ratio was larger in the CCM samples than in the classical ones (Table 3).

According to the RM-ANOVA results, the alpha ratio was a significant differentiator of emotions for both the CCM and classical singers (Table 4). This result was confirmed in the univariate analysis for the CCM singers, but not for classical singers (Table 5), which suggests that changes in the alpha ratio in classical style were related to changes in *fo* and SPL.

HNR

In this study, HNR did show a statistically significant difference between the emotions. HNR increased with pitch. The classical samples had a slightly higher HNR (suggesting a clearer sounding voice quality) in all emotions except for anger, in which the CCM samples had a larger HNR (Tables 3 and 4). However, the univariate analysis did not yield statistically significant results for HNR in emotion expression; instead, it showed a positive relation to changes in both F0 and SPL (Table 5).

Perturbation and vibrato

Tenderness and sadness contained more aperiodic fluctuations of *fo* compared to the other emotions (Table 3). *Fo* vibrato was slightly faster in the CCM samples than in the classical ones. The difference in jitter was statistically significant between emotions for the CCM samples in the RM-ANOVA analysis. No significant differences between emotions were found for vibrato in either of the genres (Tables 3 and 4). Univariate analysis did not show any significance for jitter.

The irregular variation of the period amplitude (shimmer) was larger in the low activity emotions (tenderness, sadness) and smaller in the high activity emotions (joy, anger). Shimmer yielded statistically significant results for the CCM samples. The amplitude vibrato rate differed significantly between emotions for both the CCM and classical samples

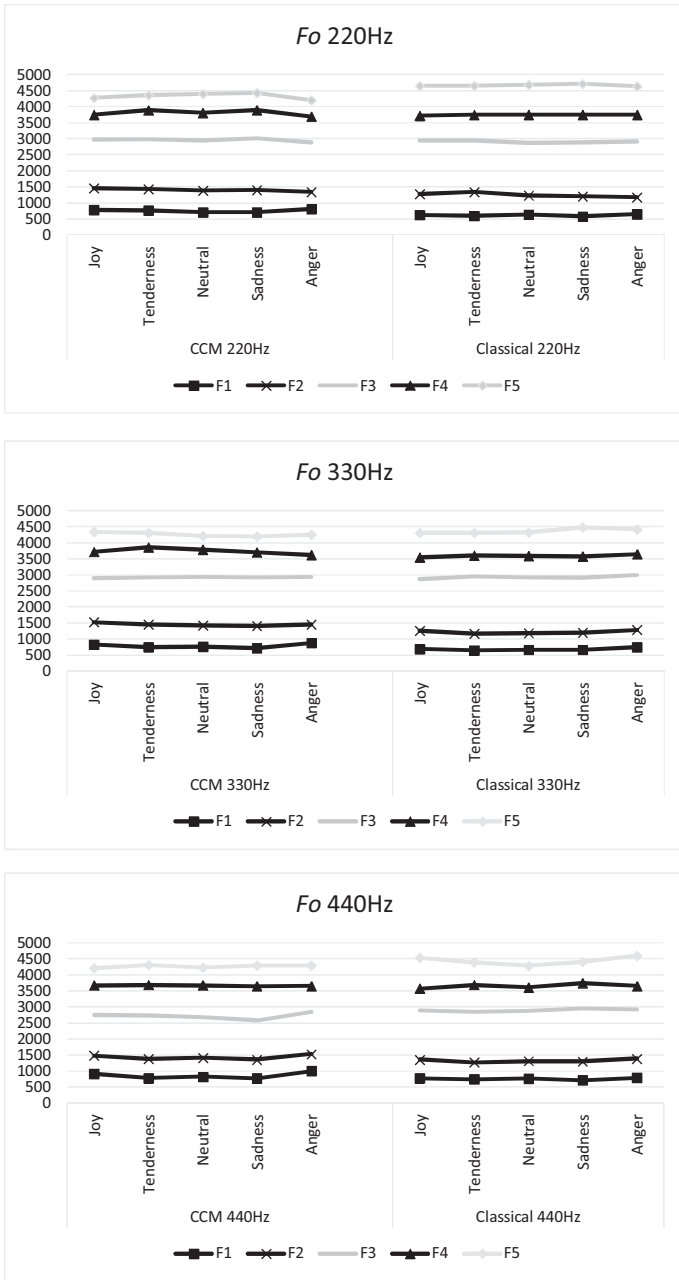

FIGURE 2. Formant positioning in different emotions at different pitches.

(Tables 3 and 4). The univariate analysis showed a significance for shimmer apq5 in the CCM samples, and amplitude vibrato was found to differ significantly between the emotions also in the univariate test for the CCM singers (Table 5).

Onsets and offsets of sound

We wanted to investigate if the amplitude contour revealed some sort of pattern in these short vocal samples as well. The nominal durations for sustained sounds in "Summertime" according to the notation and selected tempo of the

TABLE 3.
Mean Values of Parameters That Differed Significantly Between Emotions in the RM-ANOVA Analysis

		Mean Values																					
		SPL (dB)		HNR (dB)		F1 (Hz)		F2 (Hz)	F4 (Hz)	Alpha Ratio (dB)		Jitter Rap (%)	Jitter ppq5 (%)	Shimmer appq3 (dB)	Shimmer apq5 (dB)	F0 Vibrato rate (Hz)	dB Vibrato rate (dB)		Sustain Time (s)	
		CCM	Classical	CCM	Classical	CCM	Classical	CCM	CCM	CCM	Classical	CCM	CCM	CCM	CCM	CCM	CCM	Classical	CCM	Classical
Joy	220 Hz	62	63	15	16	787	632	1453	3754	−13	−22	0.003	0.003	0.020	0.032	6	8	8	2.4	2.4
	330 Hz	70	69	22	22	839	693	1525	3717	−13	−18	0.002	0.002	0.016	0.018	6	7	7	3.4	3.5
	440 Hz	73	77	22	24	916	774	1482	3674	−12	−17	0.002	0.002	0.014	0.016	6	6	7	1.8	1.9
Tenderness	220 Hz	58	59	14	14	733	599	1433	3887	−16	−22	0.003	0.003	0.028	0.041	7	9	9	2.7	2.4
	330 Hz	65	67	21	22	753	644	1450	3854	−17	−22	0.002	0.002	0.015	0.020	6	7	7	3.6	3.8
	440 Hz	66	73	22	25	779	744	1385	3683	−18	−21	0.002	0.002	0.015	0.014	4	7	6	1.8	2.0
Neutral	220 Hz	59	62	15	17	707	641	1391	3809	−18	−23	0.003	0.003	0.022	0.036	5	8	8	2.4	2.4
	330 Hz	66	68	22	23	757	659	1431	3791	−17	−22	0.002	0.002	0.014	0.017	6	7	6	3.3	3.3
	440 Hz	70	74	24	25	825	765	1419	3669	−17	−20	0.002	0.002	0.018	0.020	6	6	6	1.8	2.0
Sadness	220 Hz	59	60	14	15	714	584	1409	3894	−19	−23	0.004	0.004	0.026	0.045	6	9	8	2.5	2.4
	330 Hz	67	68	22	24	719	663	1412	3709	−18	−23	0.002	0.002	0.014	0.017	6	7	7	3.4	3.8
	440 Hz	69	73	23	25	773	716	1362	3646	−18	−20	0.002	0.002	0.015	0.018	6	7	7	1.7	1.9
Anger	220 Hz	70	64	17	17	809	654	1351	3694	−15	−24	0.002	0.002	0.015	0.023	4	7	7	2.3	2.4
	330 Hz	75	72	23	21	873	750	1454	3617	−14	−18	0.001	0.001	0.011	0.014	5	6	6	2.9	3.2
	440 Hz	77	78	24	24	1003	791	1533	3650	−11	−16	0.002	0.002	0.008	0.010	5	6	6	1.5	1.9

music piece were 2.25 seconds for A_3, 4.5 seconds for E_4, and 2.25 seconds for A_4.

In Table 3, we can see the tendency of the CCM singing style to cut the notes somewhat shorter in comparison to the classical singing style. We can also see that the higher the pitch, the shorter the note value is in comparison to the nominal value. The onset of the vowel sound is very similar in all the samples, but the offset is slightly different depending on the emotion. The sustain time for the CCM samples was found to differ significantly between the emotions (Greenhouse & Geisser correction F (2.430, 65.608) = 3.557, $P = 0.026$). No such significance was found for the classical samples, nor for the attack or release time for either genre in these data (Table 4).

DISCUSSION

The present study investigated acoustic voice quality differences between emotional expressions in short singing excerpts of the vowel [ɑ:] from classical and CCM singers. In addition to the parameters related to voice quality, we also included prosodic elements (onset, offset, and sustain time) that have been identified as potential carriers of emotional content in the singing voice.[24] While caution is needed when extrapolating from data produced by a small number of people making specific deliberate sounds while expressing requested emotions, we found some interesting coherence in this sample set. Our results show that voice quality has a role in emotional expression in the singing voice. They also indicate several differences between the classical and CCM styles in the use of voice quality as a carrier of emotional messages in singing.

Repeated measures ANOVA found a significant effect of emotion on 11 of the 20 parameters measured for the CCM samples and 4 of the parameters measured for the classical samples. The common statistically significant parameters found in both genres were SPL, HNR, alpha ratio, and F1. When the effects of *fo* and SPL were taken into account in the univariate analysis of variance, significance remained for CCM in the alpha ratio, sustain time, shimmer, F1, F2, and amplitude vibrato rate. For the classical samples, none of the parameters remained significant differentiators between emotions after *fo* and SPL had been set as co-variates.

Pitch is one of the most prominent cues of emotional content in the speaking voice.[32,33,37] Its expressive power is diminished in singing due to the nature of music. Rather, it may be an intervening factor in the conveyance of emotions. For instance, according to the results of Hakanpää et al,[13] listeners tend to perceive vowel samples sung at a high pitch as expressions of joy or anger, while samples sung at a low pitch are perceived to convey sadness. In the present study, *fo* was measured, as it may show emotion-related differences in fine tuning. Livingstone et al[26] found that the increased genuineness of emotion expression (according to listener ratings) was associated with decreased pitch accuracy in singing (of male actors). One might expect that

TABLE 4.
Results of the RM-ANOVA Tests of Within-Subject Effects (Gray: Sphericity Violated—Greenhouse-Geisser Corrected; White: Sphericity Assumed)

	CCM			Classical		
	df	F	sig.	df	F	sig.
SPL	2.533, 68.386	54.9	<0.001	2.732, 57.367	18.278	<0.001
HNR	3.098, 83.638	0.844	0.001	4, 84	2.944	0.025
F1	2.455, 66.273	21.382	<0.001	4, 84	6.242	<0.001
F2	3.047, 82.268	7.06	<0.001	2.321, 48.746	1.891	0.156
F2*pitch	6.094, 82.268	4.319	0.001	4.643, 48.746	3.226	0.015
F4	2.898, 78.255	3.602	0.018	2.081, 43,711	1.057	0.359
Alpha ratio	4, 108	23.105	<0.001	1.934, 40.612	5.10	0.011
Jitter rap	2.531, 68.343	5.208	0.004	2.024, 42.503	1.54	0.226
Jitter ppq	2.589, 69.898	4.781	0.006	2.122, 44.568	1.658	0.201
Shimmer apq3	4, 108	10.056	<0.001	2.465, 51.766	1.994	0.137
Shimmer apq3*pitch	8, 108	2.125	0.039	4.930, 51.766	4.194	0.003
Shimmer apq5	2.949, 79.626	10.153	<0.001	2.281, 47.909	2.716	0.069
Shimmer apq5*pitch	5.898, 79.626	2.735	0.019	4.563, 47.909	4.606	0.002
F0 vibratorate*pitch	8,108	2.854	0.006	8, 84	1.09	0.378
dB vibrato rate	4, 108	5.335	0.001	4, 84	4.827	0.001
dB vibrato extent*pitch	5.408, 73.014	3.153	0.011	4.777, 50.155	1.771	0.139
Sustain time	2.430, 65.608	3.557	0.026	4, 84	2.305	0.065

microvariations in pitch might be a way of expressing emotion in the singing voice. However, all participants in the present study sang very close to the target pitch. The accuracy was somewhat better in the classical style. Livingstone et al also reported that the amount of received singing education factored in to singing in tune.[26] It seems that for singers, staying in tune is prioritized over emotional expression via decreased pitch accuracy. The pitch also sets some restrictions on the voice quality manipulation of the singer, as it takes more effort to sustain phonation at the extremities of the vocal range. To add extra features to the sound at very high or low frequencies requires advanced techniques. In the present study, the pitch range used was quite low for the classical style.

SPL is an important parameter in emotional coding and decoding in the singing voice.[1,21,30] It can be recognized as a very clear differentiator between emotions. High activity emotions (joy, anger) were characterized by a larger SPL compared to low activity emotions (sadness, tenderness; Table 2). SPL is known to rise with pitch.[38] According to our data, SPL rose more with the pitch for the participants using the classical singing technique. In the CCM samples, SPL increased an average of 7-11 dB with pitch, depending on the emotions expressed, while in the classical style, the

TABLE 5.
Statistically Significant Parameters of Emotion Expression in the Univariate Analysis of Variance

	CCM			Classical		
Univariate Analysis of Variance	df	F	sig.	df	F	sig.
Alpha ratio	4, 134	7.325	<0.001	4, 104	1.236	0.300
Sustain time	4, 134	2.416	0.052			0.152
Shimmer apq5	4, 134	2.709	0.033	4, 104	0.943	0.442
Shimmer apq5*pitch			0.000			0.000
F1	4, 134	4.124	0.003	4, 104	0.316	0.867
F1*pitch		3.61	0.030			0.615
F1*SPL		8.823	0.004			0.993
F2		2.869	0.026			0.106
F2*pitch			0.094			0.000
F2*SPL			0.342			0.002
dB vibrato rate by emotion pitch with SPL	4, 134	1.756	0.141	4, 104	2.021	0.097
dB vibrato rate by emotion pitch genre with SPL	4, 239	3.546	0.008	4, 239	0.574	0.682

increase was 12-14 dB (Table 3). This may indicate that energy conversion in the classical style becomes more effective at higher pitches when compared to the two lowest pitches used in this study (Table 3).

Formant manipulation is an important tool for singers in both the classical and CCM styles.[39-42] In speech, F1 has been reported to be higher in expressions of joy and anger and lower in sadness.[33] Similar trends could be observed in the present study as well. Participants lifted F1 for the high activity emotions and lowered it for the low activity emotions (Table 3, Figure 1). A statistically significant effect was also found for F2 for the CCM singing style (Table 4, Figure 1). High activity emotions have been reported to have a low formant bandwidth in comparison to low activity emotions.[1] Our findings support this claim, as the formants were scattered further apart from each other in sadness and tenderness, but packed closer together in expressions of anger and joy.

The alpha ratio varied across the emotion portrayals on the activity dimension, much like F1 and F2: it increased in high activity emotions and decreased in low activity emotions (Table 3). While formants are physically linked to the size and shape of the vocal tract, the alpha ratio is physically linked to the voice source level, reflecting the rate of glottal closing. The alpha ratio values in our data were consistently larger for the CCM samples. When testing the data with univariate analyses only, the CCM samples achieved significance in different emotions (Table 5). This may be interpreted as showing that the CCM style uses more expression-related voice source variation than the classical style does. The alpha ratio showed less variation in the classical samples, most likely reflecting the demand for a more stable voice source quality.

HNR varied mostly according to pitch. Low pitches were portrayed with a smaller HNR; conversely, high pitches were portrayed with a larger HNR. Anger was portrayed with a larger HNR compared to the other emotions (Table 3). HNR showed a statistically significant difference between the emotions. However, the univariate analysis of variance did not yield statistically significant values for HNR independently, as F0 and SPL both influenced its effect (Tables 4 and 5).

Vibrato is not used as a code for emotion expression in music, but it can be interpreted as one—namely fast vibrato as anger and slow vibrato as sadness.[3] In vocal music, the vibrato qualities of the voice might affect the perception of emotion. There are two kinds of vibrato: (1) frequency vibrato, in which the pitch undulates and (2) amplitude vibrato, in which the loudness undulates.[43] Only amplitude vibrato rate was found to be a significant factor for emotion expression for both the CCM and classical samples in this study (when tested with the RM-ANOVA). It was still a significant contributor in the CCM samples when tested against the effect of pitch and SPL with the univariate analysis of variance (Tables 4 and 5).

Jitter and shimmer showed statistically significant effects in the RM-ANOVA only for the CCM singers. Shimmer apq5 for the CCM samples was the only perturbation parameter that yielded significant results after the univariate analysis (Tables 4 and 5). Sundberg proposed that the onset and offsets of sound could be especially potent codes in the musical expression of emotion. In anger, the onsets would be fast and in sadness slow.[20,25] In our data, we did not find an effect for onset or offset. There was a statistically significant effect for the sustain time of the CCM samples, where anger was portrayed with a shorter sustain time, while tenderness was portrayed with a longer sustain time (Table 3). We did find significant effects for the dB values at the onset and offset, but the onset and offset time *per se* did not have an effect.

As singers and singing styles are many and various, it is important to keep in mind that the results obtained from this study are valid only in the context of the data gathered for this study. Further study is needed to validate the hypotheses. More research is also needed on the emotion expression of specific song genres. In classical music, for example, there are strict requirements in terms of pronunciation in different languages, pitch accuracy and duration, resonance configuration, sustained legato, evenness of vibrato, and control over volume and phrase length for different types of classical singing. More research is needed on how these genre-specific requirements affect emotional expression. The CCM singing styles usually have a wider degree of freedom when following the notation, but CCM styles also have very specific sound ideals for different genres. Factors that should be addressed in future studies include the use of lyrics vs. scatting or improvising when singing; being emotional vs. expressing emotion when singing; larger musical sections, such as the phrasing, key, tempo, and rhythm of the song; the most comfortable pitch range of each vocalist; and the comparable skill level of the vocalists (while not discriminating against "regular" singers).

CONCLUSIONS

Previous research has shown that multiple sound parameters are used to encode emotion, and our findings support this assessment. Understanding the components of the "code" could help singers and vocal teachers to be more precise when conveying emotions in the singing voice.

The results differed somewhat between the CCM and classical styles. The alpha ratio, F1, F2, shimmer, amplitude vibrato rate, and the sustain time of vocalization were related to F0 and SPL control in the classical style. The only common independent sound parameter indicating emotional expression for both styles was SPL. Our results suggest that compared to the Western classical style, the CCM style offers more freedom for expression by using changes in voice quality.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.jvoice.2019.10.002.

REFERENCES

1. Scherer KR, Sundberg J, Fantini B, et al. The expression of emotion in the singing voice: acoustic patterns in vocal performance. *J Acoust Soc Am*. 2017;142:1805–1815. https://doi.org/10.1121/1.5002886.
2. Darwin C, Ekman P. The Expression of the Emotions in Man and Animals (3rd ed.). *Expr Emot man Anim 3rd ed*.1872. https://doi.org/10.1037/10001-000.
3. Juslin PN, Laukka P. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol Bull*. 2003;129:770–814. https://doi.org/10.1037/0033-2909.129.5.770.
4. Scherer KR. Vocal markers of emotion: comparing induction and acting elicitation. *Comput Speech Lang*. 2013;27:40–58. https://doi.org/10.1016/j.csl.2011.11.003.
5. Juslin PN, Sloboda JA. *Handbook of Music and Emotion: Theory, Research, Applications*.; 2010. https://doi.org/10.1093/acprof.
6. Pervin LA. *The Science of Personality*.; 2003.
7. Leppänen JM. Emootiokategoriat ja niiden tutkiminen. In: Hämäläinen H, Laine M, Aaltonen O, Revonsuo A, eds. *Mieli Ja Aivot Kognitiivisen Neurotieteen Oppikirja*. Turku: Kognitiivisen neurotieteen tutkimuskeskus, Turun yliopisto; 2006:311–317.
8. Niedenthal PM, Ric F. *Psychology of Emotion*.; 2017. https://doi.org/10.4324/9781315276229.
9. Cespedes-Guevara J, Eerola T. Music communicates affects, not basic emotions—a constructionist account of attribution of emotional meanings to music. *Front Psychol*. 2018. https://doi.org/10.3389/fpsyg.2018.00215.
10. Waaramaa T, Laukkanen AM, Airas M, et al. Perception of emotional valences and activity levels from vowel segments of continuous speech. *J Voice*. 2010;24:30–38. https://doi.org/10.1016/j.jvoice.2008.04.004.
11. Quinto LR, Thompson WF, Kroos C, et al. Singing emotionally: a study of pre-production, production, and post-production facial expressions. *Front Psychol*. 2014. https://doi.org/10.3389/fpsyg.2014.00262.
12. Juslin PN, Scherer KR. Vocal expression of affect. In: *The New Handbook of Methods in Nonverbal Behavior Research*.; 2008. https://doi.org/10.1093/acprof:oso/9780198529620.003.0003.
13. Hakanpää T, Waaramaa T, Laukkanen A-M. Emotion recognition from singing voices using contemporary commercial music and classical styles. *J Voice*. 2018. https://doi.org/10.1016/j.jvoice.2018.01.012.
14. Patel S, Scherer KR, Björkner E, et al. Mapping emotions into acoustic space: the role of voice production. *Biol Psychol*. 2011. https://doi.org/10.1016/j.biopsycho.2011.02.010.
15. Sundberg J, Iwarsson J, Hagegard H. A singer's expression of emotions in sung performance. In: Fujimura O, Hirano M, eds. *Vocal Fold Physiology: Voice Quality Control*. Singular Pub Group; 1995:357.
16. Boersma P, Weenink D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.19, retrieved 16 October 2016 from http://www.praat.org/.
17. Guzman MA, Dowdall J, Rubin AD, et al. Influence of emotional expression, loudness, and gender on the acoustic parameters of vibrato in classical singers. *J Voice*. 2012. https://doi.org/10.1016/j.jvoice.2012.02.006.
18. Laver J. *The Phonetic Description of Voice Quality*. Cambridge University Press; 1980.
19. Seikel J, King D, Drumright D. *Anatomy and Physiology for Speech, Language, and Hearing*.; 2009. http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract%5Cnhttp://books.google.com/books?hl=en&lr=&id=LFBOhaD1JHwC&oi=fnd&pg=PR7&dq=Anatomy+&+Physiology+for+Speech,+Language,+and+Hearing&ots=9Gv5WCUMRl&sig=gPuraE2s72r2cJoHaeObmWkLGAs.
20. Jansens S, Bloothooft G, de Krom G. Perception and acoustics of emotions in singing. *Proc Fifth Eur Conf Speech Commun Technol*. 1997;0:0–3. http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=9747D0A838F2790BD0161DCF94739C2E?doi=10.1.1.56.8871.
21. Eyben F, Salomão GL, Sundberg J, et al. Emotion in the singing voice—a deeper look at acoustic features in the light of automatic classification. *EURASIP J Audio Speech Music Process*. 2015;2015:19. https://doi.org/10.1186/s13636-015-0057-6.
22. Mayor O, Bonada J, Loscos A. The singing tutor: expression categorization and segmentation of the singing voice. In: *Proc AES 121st Conv*. 2006.
23. Siegwart H, Scherer KR. Acoustic concomitants of emotional expression in operatic singing: the case of lucia in Ardi gli incensi. *J Voice*. 1995;9:249–260. https://doi.org/10.1016/S0892-1997(05)80232-2.
24. Sundberg J. Emotive transforms: acoustic patterning of speech its linguistic and physiological bases. *Phonetica*. 2000;57:95–112.
25. Sundberg J. Expressivity in singing. A review of some recent investigations. *Logop Phoniatr Vocol*. 1998;23:121–127. https://doi.org/10.1080/140154398434130.
26. Livingstone SR, Choi DH, Russo FA. The influence of vocal training and acting experience on measures of voice quality and emotional genuineness. *Front Psychol*. 2014;5. https://doi.org/10.3389/fpsyg.2014.00156.
27. Scherer KR. Expression of emotion in voice and music. *J Voice*. 1995;9:235–248. https://doi.org/10.1016/S0892-1997(05)80231-0.
28. Tartter VC. Happy talk: perceptual and acoustic effects of smiling on speech. *Percept Psychophys*. 1980;27:24–27. https://doi.org/10.3758/BF03199901.
29. Scherer KR, Trznadel S, Fantini B, et al. Recognizing emotions in the singing voice. *Psychomusicol Music Mind Brain*. 2017. https://doi.org/10.1037/pmu0000193.
30. Scherer KR, Sundberg J, Tamarit L, et al. Comparing the acoustic expression of emotion in the speaking and the singing voice. Comput Speech Lang. 2015;29:218-235. https://doi.org/10.1016/j.csl.2013.10.002.
31. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J Acoust Soc Am*. 1993. https://doi.org/10.1121/1.405558.
32. Laukkanen A-M, Vilkman E, Alku P, et al. On the perception of emotions in speech: the role of voice quality. *Logop Phonatr Vocol*. 1997;22:157–168. https://doi.org/10.3109/14015439709075330.
33. Waaramaa T, Laukkanen AM, Alku P, et al. Monopitched expression of emotions in different vowels. *Folia Phoniatr Logop*. 2008;60:249–255. https://doi.org/10.1159/000151762.
34. Airas M, Alku P. Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*. 2006;63:26–46. https://doi.org/10.1159/000091405.
35. Bunch M, Chapman J. Taxonomy of singers used as subjects in scientific research. *J Voice*. 2000. https://doi.org/10.1016/S0892-1997(00)80081-8.
36. Teixeira JP, Oliveira C, Lopes C. Vocal acoustic analysis—jitter, shimmer and HNR parameters. *Procedia Technol*. 2014. https://doi.org/10.1016/j.protcy.2013.12.124.
37. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*. 1996;70:614–636. https://doi.org/10.1037/0022-3514.70.3.614.
38. Titze IR. *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice-Hall; 1994.
39. Sundberg J, Lã FMB, Gill BP. Formant tuning strategies in professional male opera singers. *J Voice*. 2013;27:278–288. https://doi.org/10.1016/j.jvoice.2012.12.002.
40. Sundberg J, Bitelli M, Holmberg A, et al. The "Overdrive" mode in the "Complete Vocal Technique": a preliminary study. *J Voice*. 2017. https://doi.org/10.1016/j.jvoice.2017.02.009.
41. Titze IR, Worley AS, Story BH. Source-vocal tract interaction in female operatic singing and theater belting. *J Sing*. 2011;67:561–572.
42. Sundberg J. Articulatory configuration and pitch in a classically trained soprano singer. *J Voice*. 2009. https://doi.org/10.1016/j.jvoice.2008.02.003.
43. Sundberg J. Perception of singing. In: *The Psychology of Music*.; 2013:69-105. https://doi.org/10.1016/B978-0-12-381460-9.00003-1.

PUBLICATION III

Training the Vocal Expression of Emotions in Singing: Effects of Including Acoustic Research-Based Elements in the Regular Singing Training of Acting Students

Tua Hakanpää, Teija Waaramaa, and Anne-Maria Laukkanen

Original publication: Journal of Voice, accepted for publication December 2020
https://doi.org/10.1016/j.jvoice.2020.12.032

Training the Vocal Expression of Emotions in Singing: Effects of Including Acoustic Research-Based Elements in the Regular Singing Training of Acting Students

***Tua Hakanpää, *,†Teija Waaramaa, and *Anne-Maria Laukkanen,** **Tampere, and †Vaasa, Finland*

Summary: Objectives. This study examines the effects of including acoustic research-based elements of the vocal expression of emotions in the singing lessons of acting students during a seven-week teaching period. This information may be useful in improving the training of interpretation in singing.

Study design. Experimental comparative study.

Methods. Six acting students participated in seven weeks of extra training concerning voice quality in the expression of emotions in singing. Song samples were recorded before and after the training. A control group of six acting students were recorded twice within a seven-week period, during which they participated in ordinary training. All participants sang on the vowel [a:] and on a longer phrase expressing anger, sadness, joy, tenderness, and neutral states. The vowel and phrase samples were evaluated by 34 listeners for the perceived emotion. Additionally, the vowel samples were analyzed for formant frequencies (F1−F4), sound pressure level (SPL), spectral structure (Alpha ratio = SPL 1500−5000 Hz − SPL 50−1500 Hz), harmonic-to-noise ratio (HNR), and perturbation (jitter, shimmer).

Results. The number of correctly perceived expressions improved in the test group's vowel samples, while no significant change was observed in the control group. The overall recognition was higher for the phrases than for the vowel samples. Of the acoustic parameters, F1 and SPL significantly differentiated emotions in both groups, and HNR specifically differentiated emotions in the test group. The Alpha ratio was found to statistically significantly differentiate emotion expression after training.

Conclusions. The expression of emotion in the singing voice improved after seven weeks of voice quality training. The F1, SPL, Alpha ratio, and HNR differentiated emotional expression. The variation in acoustic parameters became wider after training. Similar changes were not observed after seven weeks of ordinary voice training.

Key Words: Voice quality−Perceived emotion−Acoustic analyses.

INTRODUCTION

An important value of music lies in its capacity to express emotions. Many of the acoustic attributes that musicians use to express emotion are also important in vocal expression. It has been suggested that musical and vocal expressions share a common expressive code.[1,2] This seems to offer an advantage for vocalists in the conveyance of emotions through their music. However, there are still many questions concerning emotional expression in the singing voice, especially regarding its training.

Plenty of techniques have been developed to train interpretation and emotional expression: some are based on the reflective system of information processing, consisting of rigorous self-observation (such as the Stanislavski method[3] or Psychodrama[4]), and others try to engage the more subconscious associative route (such as TRE® or NLP). These methods and their variations are used in voice studios all over the world. The age-old master-apprentice tradition allows for an experimental approach toward these methods, which leads many teachers to use the "good parts" of different methods and rework them into exercises that suit their individual teaching styles. This is both good and bad: it is good in a sense that voice pedagogy keeps reinventing itself, but bad in the sense that the original idea sometimes gets lost in the metamorphoses and can result in a pseudotherapeutic experiment laboratory where things can go awry fast.

The singing voice is an instrument where the self and the voice have a complex inter-relation. A singer's vocal identity is composed of musical and self-identity, but it is also regulated by public and interpersonal transactions that influence perceptions of (or reactions to) the physical instrument and emotional self.[5] The training of emotional expression in the singing studio may be difficult due to (a) a lack of cognitive resources, such as episodic memory (the person has not experienced these emotions), or due to identity-related beliefs about emotions, or (b) motivational issues, such as social status and how one wants to present oneself to others or issues concerning the repression of emotions.[6] This leaves the singing instructor in a difficult position: how should one address the expression of emotion in the singing voice without getting overly involved in the student's emotional life? Furthermore, there are genre-typical esthetic demands that require genre-specific vocal techniques also when expressing emotions. Music pieces themselves regulate many acoustic variables normally used in the conveyance of emotions, like

Accepted for publication December 22, 2020.
Declarations of interest: none.
From the *Speech and Voice Research Laboratory, Faculty of Social Sciences, Tampere University, Tampere, Finland; and the †Communication Sciences, University of Vaasa, Vaasa, Finland.
Address correspondence and reprint requests to Tua Hakanpää, Speech and Voice Research Laboratory, Faculty of Social Sciences, Tampere University, Åkerlundinkatu 5, 33014 Tampere, Finland. E-mail: Tua.Hakanpaa@tuni.fi
Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2020.12.032

pitch, tempo, and also loudness to some extent. All of this complicates the emotional expression in singing and naturally also its training.

One of the most influential changes in teaching singing in the 21st century has been the increasing emphasis on voice science. This has led many teachers to re-examine how they teach vocal technique.[7] Voice science has helped teachers use anatomy, physiology, and the principles of skill acquisition to improve vocal training.[8] There are also a vast amount of studies concerning the acoustic characteristics used in the expression of emotions in singing and the differences between song genres in this respect.[9–18] This leads to the question of whether this information can be exploited in training vocal expression in singing.

The field of emotional expression is naturally very complex. In addition to the basic emotions (happiness, sadness, fear, anger, disgust, and surprise) whose expressions have been found to be relatively universal,[19–22] there are plenty of more subtle emotions whose expressions are culturally shaped.[23,24] Emotions also have degrees (strong–weak) and nuance (eg, cold or hot anger, depressive, submissive sadness *vs*. grief), etc. One way to simplify the topic for research or practical purposes is to classify emotions according to the activity (arousal) level involved and the valence (negative–neutral–positive). Negative valence relates to something unpleasant, potentially threatening, while positive valence relates to something that is interpreted as pleasant and is potentially good for survival. According to this kind of classification, joy and anger have a higher activity level than sadness and tenderness, for example, and joy and tenderness represent a positive valence, while anger and sadness carry a negative valence. We concentrate here on the expression of joy, anger, sadness, and tenderness because of their opposite placement on the valence-activation scale and also because they are categories of emotion that are frequently encountered in the song literature. Table 1 summarizes some results from previous studies.

We know from previous research that the vocal characteristics change when expressing different emotions in the singing voice. Depending on the style of singing, we can deliberately enhance various characteristics to a greater or lesser degree to aid emotion recognition from the singing voice.[9,10,25,26,11–18] Vocal expressions of high-activity emotions compared to expressions of low-activity emotions are typically characterized by a faster tempo, greater loudness, and quick changes in amplitude, a lower level difference between the lowest partials H1 and H2, a flatter spectrum, a large extent of vibrato, local departures from pitch contour at tone onsets, and a higher degree of perturbation and noise (more jitter, lower harmonic-to-noise ratio [HNR]).[10,13,14,17,27] Valence is coded in a more subtle way, through low or intermediate parameter values and differences in formant frequencies. According to Scherer et al,[16] the highest formant frequency mean (ie, mean of the formant frequencies measured) was found in joy and lowest in tenderness. In turn, Waaramaa et al found on average somewhat lower formant frequencies to characterize negative valence.[28] The samples were monopitched vowels produced by acting students in a speaking or speech-like singing voice. Furthermore, according to Waaramaa et al,[29] samples with a synthetically raised F3 were more often perceived as positive in valence than those with the original F3 lowered or totally removed. These differences may be related to differences between singing and speaking or differences in the strength or nuance of emotion expressed. For instance, sadness expressed in a whining voice may be expected to have higher formant frequencies than expressions of depressive sadness.

The supreme parameter used in the coding of emotions is loudness—or its principal acoustic correlate sound pressure level (SPL). Many other parameters (fundamental frequency and also spectral slope) accompany variation in SPL, which is regulated by subglottic pressure (Psub) and vocal fold adduction, and also by vocal tract acoustics. Varying vocal fold adduction along the axis from breathy to pressed induces a change from a steep to a gentle spectral slope (flattening of the spectrum). Loose adduction causes a large amplitude difference between the lowest partials (H1–H2, ie, H1 dominates the spectrum), while the opposite is seen when the adduction is tight (a pressed, strained voice). Likewise, strong adduction results in stronger relative spectral energy above 500 Hz (or 1 kHz). Vocal tract resonances affect SPL. For example, raising F1 closer to F2 (by dropping the lower jaw and/or opening the mouth wider) increases it.[30–35]

Formant positioning also affects the timbral qualities of the voice. Raising the formant frequencies (eg, by smiling or using a more frontal articulation) makes the voice timbre brighter. This increases perceived loudness, as it increases the sound energy in the higher frequency range (between 2 and 4 kHz), where the human ear is more sensitive. Lowering the formant frequencies (eg, by yawning or protruding the lips or vocalizing with a retracted tongue or small mouth opening) darkens the timbre. Adduction and vocal tract acoustics interact. For example, when the adduction is loose, it lowers the amplitude of the formants and broadens their bandwidths.[8,32]

Perturbation (jitter and shimmer) may be introduced by an imbalance between Psub and adduction, which causes irregular vocal fold vibration. The perceptual correlate is a rough, more or less hoarse quality. Turbulence noise may be increased by leaving a gap in the glottis and using a sufficient Psub. The perceived voice contains a hissing component. Both perturbation and turbulence noise may contribute in a decrease in HNR.

This study aims to investigate whether an acoustic research-based parameter modulation technique could be helpful in training the vocal expression of emotions in singing. A 7 × 45-minute training routine was constructed, based on research results on the acoustic characteristics of emotional expressions and their perceptual and physiological correlates. This training was tested on acting students in addition to their ordinary training. The research questions were: 1. Does the specific training improve the recognition

TABLE 1.
Acoustic Parameters Found to Code Emotional Expressions for Joy, Tenderness, Sadness, and Anger in Singing

	Acoustic Characteristic			
Perceptual Characteristic	Joy	Tenderness	Sadness	Anger
Pitch	Higher F0 floor, mean, and ceiling.[10] High on F0 variation, low on F0 rise and fall slopes.[16]	Low F0 floor, mean, and ceiling.[10] Low on F0 variation and F0 rise and fall slopes.[16]	Higher F0 floor, mean, and ceiling.[10] High on F0 variation, low on F0 rise and fall slopes.[16]	Higher F0 floor, mean, and ceiling.[10] Low on F0 variation, high on steepness of F0.[16]
Loudness	High loudness (AE*), low loudness (AE) variation, and moderate on loudness (AE) rise and fall slopes.[16] High equivalent sound level, low Hammarberg index, and low level difference between partials 1 and 2 (H1/H2).[18] Higher mean sound level and more short-term variability of sound level.[12] High SPL.[11]	Low loudness (AE), high loudness (AE) variation and rise and fall slopes, low dynamics.[16] Low equivalent sound level, high Hammarberg index, and high level difference between partials 1 and 2 (H1/H2).[18] Low mean sound level and less short term variability of sound level.[12,17] Low SPL.[11] Low vocal energy.[15]	Low loudness (AE), high loudness (AE) variation, and moderate rise and fall slopes.[16] Low equivalent sound level, high Hammarberg index, and high level difference between partials 1 and 2 (H1/H2).[18] Low mean sound level and less short-term variability of sound level.[12,17] Low SPL.[11] Low level of dynamics.[15] Low intensity.[9]	High loudness (AE), low loudness (AE) variation and rise and fall slopes.[16] High equivalent sound level, low Hammarberg index, and low level difference between partials 1 and 2 (H1/H2).[18] High mean sound level and more short-term variability of sound level.[12,17] High SPL.[11] High vocal energy[25] and high dynamics (rate, F0 contour, loudness variation).[15]
Timbre	Low formant bandwidth, low formant amplitude, high formant frequency, and moderate low-energy frequency variation.[16] Low proportion energy <0.5 kHz, low proportion energy <1 kHz, High spectral flatness, and high spectral centroid.[18] High Alpha ratio.[11,18] Shallow spectral slope[9] and narrow bandwidth.[11]	High formant bandwidth, moderate formant amplitude, small formant frequency, tendency for high low-frequency energy, and small low-energy frequency variation.[16] High proportion energy <0.5 kHz, high proportion energy <1 kHz, low spectral flatness, low spectral slope, and low spectral centroid.[18] Low Alpha ratio.[11,18] Broad bandwidth.[11]	High formant bandwidth, low formant amplitude, small formant frequency, small low-energy frequency variation.[16] High proportion energy <0.5 kHz, high proportion energy <1 kHz, low spectral flatness, low spectral slope, and low spectral centroid.[18] Low Alpha ratio.[11,18] Broad bandwidth.[11]	Low formant bandwidth, low formant amplitude, moderate formant frequency, high low-energy frequency variation.[16] Low proportion energy <0.5 kHz, low proportion energy <1 kHz, high spectral flatness, high spectral slope, and high spectral centroid.[18] High Alpha ratio.[18] Narrow bandwidth.[11] Weak low frequency energy.[15] Flat highly balanced spectrum, indicating strong energy in the higher partials.[17]
Tempo	Fast on tempo.[9,12,15–17,25]	Slowest on tempo.[9,12,15–17,25]	Slow on tempo.[9,12,15–17,25]	Fastest on tempo.[9,12,15–17,25]
Irregularity of sound	Low on perturbation variation, high on perturbation level.[16] Less aperiodic fluctuation of F0, more irregular variation of the period amplitude.[11] More jitter, less HNR.[10]	High on perturbation variation, low on perturbation level, and little waveform irregularity.[16,17] More aperiodic fluctuation of F0, more irregular variation of the period amplitude.[11] Less jitter, more HNR.[10]	High on perturbation variation, low on perturbation level.[16,17] More aperiodic fluctuation of F0, more irregular variation of the period amplitude.[11] More jitter, less HNR.[10]	Low on perturbation variation, high on perturbation level.[16,17] Less aperiodic fluctuation of F0, more irregular variation of the period amplitude.[11] More jitter, less HNR.[10]

* Loudness(ae) is a psychoacoustic measure that is designed to give a value to acoustically estimated loudness. According to Scherer et al (2017), this correlates better with the vocal affect dimensions than with the raw signal energy.

References[9,12,15–18] focus on the operatic voice, while references[10,11,25] concern the nonclassical voice, nonspecific voice technique, or both classical and nonclassical voices.

of emotions from the singing voice? 2. Do the acoustic differences between emotional expressions increase after the particular training? The hypotheses were: 1. The recognition of emotions will increase in the test group and will not change in the control group. 2. Acoustic differences between emotional expressions will increase in the test group. As markers of the latter, we hypothesize that (a) the number of significantly differentiating parameters will increase and (b) the range of parameters will increase.

METHODS

Participants and recording

The participants of this study were six Finnish acting students (three males, three females, mean age 25 years, SD 4 years) with a minimum of 2 years of singing lessons and on average 6 years of singing experience (median 2.25 years). The control group consisted of six gender- and age-matched Finnish acting students also with a minimum of 2 years of singing lessons. On average, they had 2 years of singing experience (median 1 year). All test subjects were native speakers of Finnish.

All subjects were instructed to perform an eight-bar musical excerpt composed especially for the test situation expressing the emotions of joy, tenderness, sadness, and anger plus a neutral state. They all sang using the syllables pa[pa:], da[da:], and fa[fa:], which in themselves in Finnish carry no meaning or emotional content as such. The excerpt was composed using the pentatonic scale in order to avoid sounding too major or too minor (Figure 1). The same test was issued before and after the teaching intervention.

The modality of the song was f-pentatonic, and the pulse was 115 bpm (beats per minute) for all test subjects and every emotion portrayal. The emotion samples and neutrals were performed in a randomized order and repeated three times. The participants were asked to identify the take they liked the best and that take was selected for further analyses.

Recordings of the test group were made in the well-damped recording studio of Tampere University Speech and Voice Research Laboratory using a Brüel & Kjær Mediator 2238 sound level meter and 4188 microphone. The distance between the microphone and the test subjects' lips was 40 cm. Samples were recorded with an external soundcard (Focusrite Scarlett 2-i-4) and Sound Forge Pro 11.0 digital audio editing software using a 44.1 kHz sampling rate and a 16-Bit amplitude quantization. The sound recordings were calibrated for SPL measurements using a sine wave generator with a known SPL.

All control samples were recorded at recording studio 365 of the University of the Arts, Helsinki, using the same microphone and recording distance as for the test subjects. The RME Babyface Pro external sound card and Cubase 10 digital recording and audio editing software were used (44.1 kHz, 16-Bit). The sound recordings were calibrated for SPL measurements in the same way as for the test group.

In order to make the experiment as lifelike as possible, the subjects used a backing track with a neutral accompaniment that was played to them via a SONY MDR V-700 headset through Zoom H-4 in the test condition and a Sennheiser HD 25-SP II 60 Ohm headset in the control condition.

All samples were saved as .wav files for further analyses with *Praat*.[36]

Listening test

A listening test was conducted in which 246 voice samples were replayed to 32 listeners. The participants completed a multiple-choice questionnaire on which emotion they perceived as being expressed.

The listening task was a web-based test with the randomized [ɑ:] vowel and phrase samples and six control samples (120 + 120 + 6). (The control samples were repeated samples of emotion portrayals selected at random.) The test was accessible through a browser by logging in with one's own password. Participants completed the test using their own equipment. The test was accessible from desktop, laptop, tablet computers, and cellphones. The participants were instructed to use headphones to ensure the best possible sound quality. The voice samples were played in a randomized order and it was possible to play the samples as many times as needed. The questionnaire was in Finnish, and the listeners were Finnish speakers. The listening test took approximately 40 minutes to complete.

The number of listeners who completed the test was 32 (27 females, five males, no reported hearing defects). The total number of answers in the listening test was 8160. There were 1632 answers from each emotion category and 1632 answers for the neutral portrayals.

Voice samples in acoustic analyses

The vowel [ɑ:] was extracted from the last bar in each sample for further analyses. The pitch was f1 (F4, 349.23 Hz) for females, and f (F3, 174.61 Hz) for males. The nominal duration of the extracted vowel (including the preceding consonant) was 3.13 seconds according to the notation and tempo of the song. A phrase was extracted consisting of two bars from the beginning of the melody. The [ɑ:] vowels and the phrases were extracted from the sung excerpts using Reaper audio editing software. The vowel samples were cut right after the preceding consonant (the Finnish language

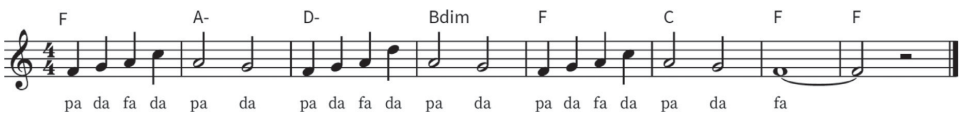

FIGURE 1. The musical excerpt sung.

does not involve aspiration after voiceless plosives). The duration of the sample vowels varied between 1.2 seconds and 4.04 seconds depending on how the test subject had interpreted the time value of the notation. The tail end of the vowel was left as the singer interpreted it (nominal note duration 3.13 seconds), as previous studies have indicated that micromanaging the durations of written notes is one way of expressing emotions in the singing voice.[13]

Acoustic parameters under investigation

Twelve acoustic parameters were automatically extracted from the voice samples. All analyses were made using *Praat* software (Version 6.0.19).[36] The vowel samples (N = 120) were analyzed for the lowest formant frequencies from F1 to F4. SPL was measured with reference to the calibration signal recorded. The Alpha ratio, which reflects the mean strength of the higher spectrum partials as compared to the lower ones, was measured using the formula SPL 1500−5000 Hz − SPL 50−1500 Hz.[37,38] The cut-off frequency was set to 1500 Hz instead of the more traditional 1000 Hz in order to better suit the analysis of the singing voice. The HNR was also measured. Two measures of jitter and shimmer were used. For jitter, we measured the relative average perturbation and five-point period perturbation quotient, and for shimmer the three-point and the five-point amplitude perturbation quotient. The relative average perturbation is the average absolute difference between an interval [glottal period] and the average of it and its two neighbors, divided by the average time between two consecutive points, and the five-point period perturbation quotient (ppq5) is the average absolute difference between an interval and the average of it and its four closest neighbors, divided by the average time between two consecutive points. The three-point amplitude perturbation quotient (Shimmer apq3) is the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude, and the five-point amplitude perturbation quotient (apq5) is the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbors, divided by the average amplitude.

Statistical analyses

The results of the listening test were coded numerically for statistical analyses. Both intended and perceived emotions were given numbers (1 = joy, 2 = tenderness, 3 = neutral, 4 = sadness, 5 = anger). Samples sung within the "test" condition were marked with 1, and those sung in the control group were marked with 2. Furthermore, the before condition was marked as 1 and the after condition as 2. The number of the correct (intended = perceived) answers for emotion are given as percentages and frequencies.

Results of the listening test were analyzed using four different statistical tests.

The first statistical test used was a binomial test (one proportion z test) to evaluate the probability that the observed percentage of the correctly recognized emotions could have resulted from random guessing. The listening test contained five different emotional states, which meant that the expected percentage of correctly recognized emotions in case of random guessing would be 20%. The statistically significant difference between guessing and correct recognition can be shown if the P value of the test is <0.05.

The second statistical test was the unrelated samples t test, which was used to compare the number of correct answers given for the test group samples and the number of correct answers given for the control group samples. The null hypothesis was that the two populations from which the two samples have been drawn have equal means. Separate t tests were run for the first recording (before) and the second recording (after). Recognition between the test group samples and control samples was interpreted to differ statistically significantly if the P value of the test was <0.05.

The third statistical test used was Pearson's chi-squared test of homogeneity to evaluate the probability that two groups of results have the same percentage of correctly recognized emotions. The percentage of correctly recognized emotions is statistically significant in two groups of results if the P value of the test is <0.05. Pearson's chi-squared test was used to compare the correct recognition within the same population under different conditions, ie, to determine if there was any difference in the recognition in the test group (and separately for the control group) between the two recordings.

The fourth statistical test used was Cronbach's alpha to evaluate the internal consistency of listener evaluations. Values >0.7 indicate acceptable internal consistency.

To evaluate whether the parameter values extracted with *Praat*[36] differed across emotions for each parameter, we computed the Friedman test (a nonparametric alternative to the one-way repeated measures ANOVA) with *SPSS* (v.17; SPSS Inc., Chicago, IL). We ran the Friedman test separately for the test group and control group and the before and after conditions. Bonferroni corrections were used for multiple comparisons.

Training procedure

General structure

Test subjects participated in a workshop consisting of seven individual lessons, each lasting 45 minutes in duration. The aim was to introduce the basic acoustic characteristics typically observed in the expressions of four particular emotional states and the perceptual correlates of these characteristics.[11,39] After introducing the characteristics, we wanted to rehearse the voluntary variation of these voice characteristics so that they resulted in clearly recognizable emotional expression.

At the beginning of the teaching intervention, the rich tradition of emotion expression coaching in actor voice training through mind imagery, self-reflection, and interaction exercises was acknowledged and discussed. This was done because we felt it was important to point out that the method we were using was somewhat rigid and it was not our intention to imply that it would be a method the acting students

should use exclusively when expressing emotions with the singing voice. After this, a more mechanical approach was agreed upon for the duration of the workshop.

Exercises used for emotion expression included basic tension and release techniques, movement with singing, and work with different breathing patterns and standard drills for varying loudness, articulation, and timbre. Such drills can be found in many books and YouTube tutorials related to the art of singing.[40–43]

The parameter modulation technique

The participants were first offered the polar opposite scale of valence and activity of emotions and placed the target emotions there. We then introduced a system of acoustic parameter manipulation for expressing different emotions. As the parameter manipulation as such is quite a mechanical way of expressing emotion, we emphasized at all times that this exercise regime was just a tool for exploring the possibilities of voice quality in emotion expression, not a definitive way of arriving at stellar expression. For the purposes of this study, we asked the students to try out the following voice quality manipulations:

- Anger: loud volume with pressed phonation, very clear articulation and no vibrato.
- Sadness: soft voice with a few volume outbursts, more breathy phonation, unclear articulation, and a lot of vocal perturbation and noise.
- Tenderness: moderate loudness and projection, slightly breathy phonation, but clear articulation, no perturbation.
- Joy: loud and well projecting voice, phonation balance (neither breathy nor pressed), inclusion of vibrato acceptable (Table 2).

The starting point of the voice modulations was the participants' habitual neutral voice. In order to use the parameter modulation technique safely, students should be aware of what their individual optimal (well-balanced, effortless) voice use is like. The extent to which the parameter manipulation can be executed (ie, how wide deviations from the optimum can be introduced) needs to be scaled individually and also for the esthetics of the singing style in use. In this study, the acting students were singing in nonclassical styles, which allowed for the maximum amount of parameter manipulation.

As each exercise needs to be fitted individually to the students' conceptual understanding of the voice and to their individual way of using it, a specific description of the exercises used cannot be given in the scope of this article. Instead, we will give a general description of how the parameter modulation was taught.

For volume control, we used exercises exploring the loudness range of each individual student from the softest possible to the loudest. We discussed each participant's habitual loudness use, comfort loudness, air flow regulation, vocal fold adduction, and the influence of the oral cavity and mouth opening on the perceived loudness.

For phonation, we used phonation balance exercises fitted to the individual need of the student (soft attack and general "hypofunction" for the "hyperfunctional" student and *vice versa*). Vocal fold movement between the barely abducted and barely adducted is said to produce a resonant voice,[44] and the goal of these exercises was to establish this zone for the students so that they can safely depart from it. We also drilled polar opposite exercises ranging from very breathy voice through optimal sound balance to pressed phonation. The goal of these exercises was to clearly demonstrate the perceptual (both acoustic and tactile/sensory) differences between the different modes of phonation.

For resonance and articulation, we used exercises that shape the vocal tract in various ways.

The articulatory exercises addressed different possibilities of the physiological positioning of the tongue, jaw, velum, and lips. The tongue position has a role in shaping the vowels and the general sound quality. The advancement of the root of the tongue results in the "fronting" of the sound, while retracting it makes the sound darker.[32,45] We used exercises for moving the tongue forward and backward (genioglossus), flattening the tongue (hyoglossus, chondroglossus), pulling the tongue back and up while depressing the soft palate (palatoglossus) and working with the intrinsic muscles of the tongue. The point of these exercises was to

TABLE 2.
Acoustic Parameters That Were Modulated During Exercises Using Either More (+) or Less (−) of That Parameter

	Perceived Acoustic Element			
Emotion Expression	Volume—Loudness	Phonation/Sound Balance	Resonance/Articulation	Perturbation/Noise*
Anger	++	+++	++	+
Sadness	−− (+−)	− (−+)	−−	++
Tenderness	−	−	−	−
Joy	+	++	+	−−
Measurement	SPL (& Alpha ratio)	Alpha ratio, HNR	F1−F4	Jitter, shimmer, HNR
	Acoustic correlates			

* Noise: increase in jitter/shimmer, decrease in HNR.

acknowledge the major role that the tongue has in shaping the oral cavity and the resulting sound. The students were encouraged to look for sounds that would (in their opinion) fit the acoustic descriptions given to the target emotions.

Exercises of the jaw movement were presented on an open-closed continuum. The students were given different exercises addressing the relaxed, open and closed jaw positions and they were instructed to experiment with different jaw openings as well as the fixed jaw position (with a bite block). The aim of these exercises was to demonstrate the full range of jaw movement as well as the possibility to hold the jaw in place and still sound intelligible. Again, the students were encouraged to explore and pick different sounds to be used in emotion expression.

The velum exercises we used were either velum up (levator veli palatini, musculus uvulae) or velum down (palatoglossus, palatopharyngeus), and the way we approached them was through mind imagery instructions such as "smelling the flower" or "like you are just about to cry."

The lips have a role in lengthening and shortening the vocal tract, which effects the formant frequencies and makes the voice color sound darker or shriller.[32,46] This effect can be achieved using exercises extending the lips outward (pouting) and retracting them sideways as in a smile. The lips (orbicularis oris) are continuous with the buccinator muscle, and ultimately with the superior pharyngeal constrictor,[47] and in this way any movement of the lips will have an effect on the shape of the oral cavity and its possibility of reinforcing formants. For this study, we used phonation with protruded and retracted lips, as well as with different lip openings and with restricted lip movement.

Another method we used was facial expressions. Facial expressions have been found to alter the voice quality, and finding a preferred sound through a facial expression is a known technique in singing instruction that lends itself readily to acoustic emotion expression.[48] By simply asking the student to make a sad, happy, angry, or tender face while phonating, we were able to generate different voice qualities through muscle movement.

Jitter and shimmer measures indirectly assess laryngeal function by quantifying acoustic correlates of irregular vocal fold vibration. Jitter measures *fo* perturbation and shimmer measures SPL perturbation, caused by vibratory variations from one vocal fold cycle to the next. Jitter is affected mainly because of the lack of control of vocal fold vibration and shimmer because of the reduction in SPL-related tension in the vocal folds. [49] Although sound perturbations refer to abnormal stability of the period length, amplitude and waveform of the vocal fold cycle, the sound is somewhat tolerant of the wave's asymmetry, and a little bit of perturbation occurs in all natural sounds. Small irregularities in the acoustic wave are considered as normal variation associated with physiological body function and voice production.[50–52] The occurrence of jitter or shimmer in voiced sound can be perceptually described as a hoarse, husky, or rough voice. As it is not exactly a desirable effect in a professional voice, we used extreme vibratos, both frequency (undulating between several semitones) and amplitude (crying-like volume changes) modulation, and even breaks in the voice to simulate perturbation in a voice-friendly way. For this exercise set, we did not use distorted sounds such as growls or screams. The extreme vibratos and tremolos were chosen for this exercise regime as they offer a perceptually concrete and singer-friendly way of practicing the unwanted undulation of sound. Adding a vocal tract-induced noise component to a steady vocal fold vibration cycle—such as we do in dist-sounds[53]—often takes a considerable amount of practice to be done safely, and as we had time restrictions, we felt that this was the better option.

RESULTS

Recognition of emotions

The emotion appraisals (answers) given in the listening test indicate that it is possible to recognize emotion from the singing voice. The overall recognition was 47.7% (χ^2 *z* value 62.57, $P = 0.000$). The recognition of phrases at 52.7% (χ^2 *z* value 52.28, $P = 0.000$) was slightly better than the recognition of short vowel samples at 42.6% (χ^2 *z* value 36.24, $P = 0.000$).

Recognition from vowel samples

For the purposes of this study, we are mainly interested in the short vowel samples, as they are seen as a carrier of information about the voice quality and as such are reflective of the usefulness of the practice regime used in the teaching intervention. Our hypothesis was that by teaching specific voice use (different voice qualities), we could improve the recognition of emotion from the singing voice.

We first ran unrelated *t* tests to determine if the recognition differed between the test group samples and the control samples. We established that there were no outliers in the data, as assessed by an inspection of the boxplot. The distribution of correct answers given in the listening test was normally distributed in all other conditions except for the test group before condition, as assessed by the Shapiro-Wilk test ($P > 0.05$). In the test group before condition, the data distribution was not normally distributed (Statistic 0.927, DF 30, $P = 0.041$). There was homogeneity of variance for the correct answers in the listening test for the test group and the control group, as assessed by Levene's test for the equality of variances ($P = 0.166$ in the before condition; $P = 0.201$ in the after condition). There were 30 voice samples evaluated by 34 listeners in the test group and 30 voice samples evaluated by 34 listeners in the control group. There were more correct recognitions of intended emotions in the test group in both the before (M = 15, SD = 10) and the after (M = 17, SD = 10) conditions. For the control group, the correct recognition of intended emotions was M = 14 (SD = 8) in the before condition and M = 11 (SD 8) in the after condition (Table 3). The mean difference in the correct answers given in response to the heard samples was 1.50 (95% confidence

TABLE 3.
Number of Correctly Recognized Vowel Samples

Expressed Feeling	Correctly Recognized Samples BEFORE	Correctly Recognized Samples AFTER	Change in Recognition
Test group			
Joy	33	46	13
Tenderness	63	77	14
Neutral	191	122	−69
Sadness	112	123	11
Anger	86	141	55
Correctly recognized samples all together	485	509	
Control group			
Joy	32	38	6
Tenderness	93	43	−50
Neutral	78	89	11
Sadness	99	89	−10
Anger	110	78	−32
Correctly recognized samples all together	412	337	

interval [CI], −3.28 to 6.28) higher in the test group in the before condition in comparison to the control group and 5.97 (95% CI, 1.31−10.62) higher in the test group in the after condition in comparison to the control group. In the before condition, there was no statistically significant difference in the mean recognition of emotion between the test group and control group ($t(58) = 0.628$, $P = 0.532$). There was a statistically significant difference in mean correct recognition of emotion between the test group samples and the control group samples after the teaching intervention ($t(58) = 2.565$, $P = 0.013$).

The results indicate that for the test group samples, the recognition of emotion increased in all emotion portrayals in the after condition. The recognition of neutral samples decreased in the after condition. For the control group samples, the recognition of emotion decreased for the after condition in all other emotion portrayals except joy. The recognition of neutral also increased (Table 3).

The internal consistency of the answers was tested with Cronbach's alpha, and the results showed a mean consistency of 0.93. Anger yielded the most consistent answers, while neutral yielded the least consistent answers (Table 4).

We ran Pearson's chi-squared test to see if there was a statistically significant difference between the answers given for samples recorded before and after the 7-week training period. For the sake of comparison, we also ran the test for the control group samples recorded at the 7-week interval. Pearson's chi-squared test showed a significant difference in the answers given for the neutral and anger portrayals in the test group and tenderness and anger portrayals in the control group. The recognition of anger increased by 28.4% in the test group from before to after training. The recognition of neutral decreased by 20.6%. In the control group, the situation was reversed: the recognition of emotion decreased from before to after in tenderness (24.4%) and anger (15.7%) (Table 4).

Recognition from phrases

Recognition from phrases seemed to be easier than recognition from the vowel samples in this study.

There were 30 voice samples evaluated by 34 listeners in the test group and 30 voice samples evaluated by 34 listeners in the control group. There were no outliers in the data, as assessed by an inspection of the boxplot. The distribution of correct answers given in the listening test was normally distributed, as assessed by the Shapiro-Wilk test ($P > 0.05$), except in the test group before condition, where the data were not normally distributed (Statistic 0.917, DF 30, $P = 0.023$). The homogeneity of variances was observed, as assessed by Levene's test for equality of variances ($P = 0.689$ in the before condition and $P = 0.218$ in the after condition). In the before condition, the phrases were better recognized from the control group samples (M = 18, SD = 10) than from the test group samples (M = 17, SD = 9). In the after condition, the situation was reversed. Phrases were better recognized from the test group samples (M = 20, SD = 8) than from the control group samples (M = 17, SD = 9; Table 5). There were no statistically significant differences in recognition. The mean difference in the correct answers given in response to the heard samples was −0.67 (95% CI, −5.52 to 4.18) in the test group in the before condition in comparison to the control group, and 3.23 (95% CI, −1.25 to 7.71) in the test group in the after condition in comparison to the control group. In the before condition, there was no statistically significant difference in the mean recognition of emotion between the test group and control group ($t(58) = -0.275$, $P = 0.784$). There was no statistically significant difference in

TABLE 4.
Correctly Recognized Short Vowel Samples in the Listening Test, the Internal Consistency of Answers, and the Statistical Significance of the Perceptual Difference in Answers Given from the Samples Recorded Before and After the Exercise Regime

		Test Group			Control Group		
		Before	After	Pearson's Chi-Squared	Before	After	Pearson's Chi-Squared
Joy	% of recognition	16.2%	22.5%	H0:%1 = %2	15.7%	18.6%	H0 = %1 = %2
	z value	1.37	0.91	2.7	-1.54	-0.49	0.6
	P value	0.17	0.363	0.103	0.123	0.624	0.431
	Cronbach's alpha	0.76	0.86		0.86	0.87	
Tenderness	% of recognition	30.9%	38.2%	H0:%1 = %2	45.6%	21.2%	H1 = %1<>%2
	z value	3.89	6.51	2.4	9.14	0.42	27.2
	P value	0.00	0	0.118	0	0.0674	0
	Cronbach's alpha	0.87	0.92		0.57	0.54	
Neutral	% of recognition	81.4%	60.8%	H1:%1<>%2	38.2%	43.6%	H0 = %1 = %2
	z value	23.64	14.56	23.1	6.51	8.44	1.2
	P value	12.46	0	0	0.000	0	0.268
	Cronbach's alpha	0.10	0.9		0.52	0.75	
Sadness	% of recognition	54.9%	60.8%	H0:%1 = %2	48.5%	43.6%	H0 = %1 = %2
	z value	12.46	14.56	1.4	10.19	8.44	1
	P value	0.00	0	0.229	0	0	0.321
	Cronbach's alpha	0.74	0.88		0.93	0.96	
Anger	% of recognition	42.2%	70.6%	H1:%1<>%2	53.9%	38.2%	H1 = %1<>%2
	z value	7.91	18.06	33.5	12.11	6.51	10.1
	P value	0.00	0	0	0	0	0.001
	Cronbach's alpha	0.97	0.96		0.97	0.96	

TABLE 5.
Correctly Recognized Phrases in the Listening Test

Expressed Feeling	Correctly Recognized Samples Before	Correctly Recognized Samples After	Change in Recognition
Test group			
Joy	71	92	21
Tenderness	114	115	1
Neutral	100	97	−3
Sadness	112	157	45
Anger	112	130	18
Correctly recognized samples all together	509	591	
Control group			
Joy	87	80	−7
Tenderness	121	125	4
Neutral	73	90	17
Sadness	142	101	−41
Anger	109	109	0
Correctly recognized samples all together	532	505	

the mean correct recognition of emotion between the test group samples and the control group samples after the teaching intervention ($t(58) = 1.445$, $P = 0.154$).

The results indicate that for the test group samples, the recognition of emotion increased in all emotion portrayals in the after condition. The recognition of neutral samples decreased in the after condition. For the control group samples, the recognition of emotion increased for the after condition in tenderness and neutral and decreased in joy and sadness. The recognition of anger was similar in both conditions (Table 5).

The internal consistency of the answers was tested with Cronbach's alpha, and it showed a mean consistency of 0.87. Anger yielded the most consistent answers, while neutral yielded the least consistent answers (Table 6).

We ran Pearson's chi-squared test to see if there was a statistically significant difference between the answers given for samples recorded before and after the 7-week training period. For the sake of comparison, we also ran the test for the control group samples recorded at the 7-week interval. Pearson's chi-squared test showed a significant difference for answers given for the neutral and anger portrayals in the test group and tenderness and anger portrayals in the control group. The recognition of anger increased by 10.1% from the test group samples from before to after training. The recognition of joy increased by 11.8% and the recognition of sadness increased 22.5%. In the control group, the recognition of tenderness increased by 2% but decreased in

TABLE 6.
Correctly Recognized Phrase Samples in the Listening Test, the Internal Consistency of Answers, and the Statistical Significance of the Perceptual Difference in Answers Given From the Samples Recorded Before and After the Exercise Regime

Recognition From Phrase		Test Group Before	After	Pearson's Chi-Squared	Control Group Before	After	Pearson's Chi-Squared
Joy	% of recognition	34.8%	46.6%	H1:%1<>%2	42.6%	39.2%	H0:%1 = %2
	z value	5.29	9.49	5.9	8.09	6.86	0.5
	P value	0.00	0	0.016	0	0	0.481
	Cronbach's alpha	0.83	0.91		0.91	0.87	
Tenderness	% of recognition	56.4%	57.1%	H0:%1 = %2	59.3%	61.3%	H0:%1 = %2
	z value	12.99	13.23	0	14.04	14.74	0.2
	P value	0.00	0	0.875	0	0	0.686
	Cronbach's alpha	0.91	0.67		0.24	0.94	
Neutral	% of recognition	49.5%	49%	H0:%1 = %2	35.8%	44.1%	H0:%1 = %2
	z value	10.54	10.36	0	5.64	8.61	3
	P value	0.00	0	0.921	0	0	0.086
	Cronbach's alpha	0.93	0.85		0.9	0.81	
Sadness	% of recognition	55.4%	77.9%	H1:%1<>%2	69.6%	49.5%	H1:%1<>%2
	z value	12.64	20.69	23.3	17.71	10.54	17.1
	P value	0.00	0	0	0	0	0
	Cronbach's alpha	0.96	0.93		0.98	0.95	
Anger	% of recognition	54.9%	65 %	H1:%1<>%2	53.4%	53.4%	H0:%1 = %2
	z value	12.46	16.04	4.3	11.94	11.94	0
	P value	0.00	0	0.037	0	0	1
	Cronbach's alpha	0.94	0.92		0.98	0.98	

joy by 3.4% and in sadness by 20.1%, while the recognition of anger remained the same (Table 6).

Acoustic results

A Friedman test was run to determine if there were differences in the usage of different sound parameters in different emotion expressions. Pairwise comparisons were performed (*SPSS*, 2019) with a Bonferroni correction for multiple comparisons. The acoustic parameters F1 and SPL differed significantly between the expressions of emotions in both the before and after conditions (Tables 7 and 8). In addition, there was a statistically significant difference of the HNR between emotions in the test group samples. In the samples recorded before the teaching intervention, F3, jitter, and shimmer distinguished emotions in the test group and F4 in the control group, but the effect was not repeated in the samples recorded after the teaching intervention/waiting period. Instead, the Alpha ratio was found to differ statistically significantly between the emotions in the after condition for the test group.

Post hoc analysis of the before condition revealed statistically significant differences in F1 between sadness (Mdn = 620 Hz) and joy (Mdn = 801 Hz; P= 0.010) and between sadness and anger (Mdn = 807 Hz; P= 0.001) in the test group. In the control group, differences were found between sadness (Mdn = 684 Hz) and joy (Mdn = 811 Hz; P= 0.019) and between sadness and anger (Mdn = 887 Hz; P= 0.010). In the after condition, post hoc analysis showed statistically significant differences between tenderness (Mdn = 658 Hz) and anger (Mdn = 852 Hz; P= 0.019), between sadness (Mdn = 658 Hz) and anger (P= 0.019), and between neutral (Mdn = 658 Hz) and anger (P= 0.019) in the test group samples.

For SPL, post hoc analyses for the before condition revealed statistically significant differences between tenderness (Mdn = 76 dB) and anger (Mdn = 86 dB; P= 0.035), between tenderness and joy (Mdn = 86 dB; P= 0.019), and between sadness (Mdn = 75 dB) and joy (P= 0.035) in the test group samples. In the control group samples, differences were found between tenderness (Mdn = 71 dB) and anger (91 dB; P= 0.019), between tenderness and joy (Mdn = 84 dB; P= 0.000), and between sadness (Mdn = 73 dB) and anger (P= 0.035).

In the after condition, statistically significant differences in SPL were found between sadness (Mdn = 73 dB) and joy (Mdn = 82 dB; P= 0.019) and between sadness and anger (Mdn = 87 dB; P= 0.001) in the test group. Statistically significant differences for the control group were found between tenderness (Mdn = 67 dB) and anger (Mdn = 80 dB; P= 0.019), between sadness (Mdn = 66 dB) and anger

TABLE 7.
The Friedman Test for Correlated Samples Recorded Before the Teaching Intervention

Before						
The Friedman Test for Correlated Samples						
	Test Group			Control Group		
	DF	χ^2	Sig.	DF	χ^2	Sig.
F1	4	21.067	0.000	4	13.733	0.008
F2	4	8.400	0.078	4	8	0.092
F3	4	9.600	0.048	4	7.600	.
F4	4	7.600	0.107	4	11.467	0.022
Alpha ratio	4	6.267	0.108	4	6.800	0.147
SPL	4	17.467	0.002	4	21.733	0.000
HNR	4	10.400	0.034	4	7.092	0.131
Jitter rap	4	10.475	0.033	4	3.322	0.505
Jitter ppq	4	11.898	0.018	4	1.544	0.819
Shimmer apq3	4	11.067	0.026	4	3.467	0.483
Shimmer apq5	4	11.200	0.024	4	5.517	0.238

(P= 0.035), and between neutral (Mdn = 67 dB) and anger (P= 0.019).

In addition to these, in the test group after condition, post hoc tests revealed statistically significant differences in the Alpha ratio between neutral (Mdn = −27 dB) and joy (Mdn = −20 dB; P= 0.035) and between tenderness (Mdn = −27 dB) and joy (P= 0.035), and in HNR between sadness (Mdn = −17 dB) and joy (Mdn = −21 dB; P= 0.019).

SPL

For the test group, SPL increased on the continuum of tenderness → sadness → neutral → joy → anger. Low intensity emotions (sadness, tenderness) were sung with a lower volume than the emotions with higher activity. The effect of the exercise routine can be seen in the wider variety in SPL control in the after condition as opposed to the before condition in the test group. In comparison to the control group, the test group showed a more consistent SPL between the recordings. In the control group, singers sang slightly louder the first time and softer the last time(Figure 2). SPL differed significantly between emotional expressions for both groups.

Alpha ratio

All of the samples for high-activity emotions (joy and anger) were characterized by a larger Alpha ratio than for the low-activity emotions (sadness and tenderness). The effects of the training routine can be seen again in the wider variety of Alpha ratio usage in the test group from before to after samples. The difference in the test group after samples in comparison to the before samples indicates that the test group started to vary their sound balance more after the teaching intervention (Figure 3).

HNR

In this study, HNR did show a statistically significant difference between the emotions in the test group in both the before and after conditions. The HNR decreased in the second recording for the test group, suggesting the use of more noise components in the signal. The most harmonic content was found in joy and the least in sadness. For the control group, HNR increased for the second recording, suggesting a more sonorous sound. The effects of the training routine were again observed in the increased variety of the use of the HNR component in the test group (Figure 4).

Formants

The overall formant structure of the samples revealed a few distinctive patterns in regard to emotion expression. In sadness, F1 was lower, and F2, F3, and F4 were higher in comparison to other emotions and neutral, suggesting a more

TABLE 8.
The Friedman Test for Correlated Samples Recorded After the Teaching Intervention

After						
The Friedman Test for Correlated Samples						
	Test Group			Control Group		
	DF	χ^2	Sig.	DF	χ^2	Sig.
F1	4	16	0.003	4	15.018	0.005
F2	4	3.930	0.416	4	5.754	0.218
F3	4	1.600	0.809	4	3.733	0.443
F4	4	4.772	0.312	4	1.263	0.868
Alpha ratio	4	17.263	0.002	4	4.772	0.312
SPL	4	21.193	0.000	4	17.825	0.001
HNR	4	10.386	0.034	4	.421	0.981
Jitter rap	4	5.895	0.207	4	.587	0.964
Jitter ppq	4	7.604	0.107	4	2.400	0.663
Shimmer apq3	4	3.228	0.520	4	2.847	0.584
Shimmer apq5	4	8.140	0.087	4	3.856	0.426

diffuse formant pattern. In anger, the opposite was true: F1 was higher and F2, F3, and F4 were lower in both groups. A similar but less pronounced formant pattern could be found in joy. In tenderness, the first formant was positioned slightly higher than in sadness, but it was still relatively low; other formants were positioned fairly high. In neutral expression, the first formant was neither high nor low, while the second and the third formants were in a relatively low position. The formant structure was most compact in anger and then it scattered in a sequence of anger → joy → neutral → tenderness → sadness. F1 was found to be statistically significant in differentiating the emotional expressions in both groups (Figure 5).

DISCUSSION

This study investigated the effects of a specific training strategy, the "parameter manipulation technique," for emotional expression in the singing voice. The technique is based on accumulated research findings on the acoustic characteristics of vocal emotion expression. The aim was to see whether the specific training improves the recognition of emotions from the singing voice and whether the acoustic differences between emotional expressions increase after the particular training. We hypothesized that the recognition of emotions would increase in the test group and not change in the control group, and that the number of significantly differentiating parameters and the range of the parameters would increase after training. While one needs to take caution in extrapolating data that have been produced by a small number of people repeating specific deliberate tasks while expressing requested emotions, the results seem to support the hypotheses at least partially. Our results suggest that training with the parameter modulation technique increased the correct recognition of emotional expression from the short vowel and phrase samples.

The number of significantly distinguishing parameters did not increase in our investigation, but the range of how various parameters were used became broader.

For the vowels, we found a statistically significant difference in mean correct recognition between the test group samples and the control group samples after the teaching intervention (P= 0.013). The recognition was 4.5% units better from the test group samples after the intervention. Our results show that for the test group samples, recognition of emotion increased in all emotion portrayals in the after condition. The recognition of neutral samples decreased in the after condition. It is fairly common to get a lot of "neutral" answers with the type of forced choice questionnaire that we used for the listening test. If the listeners are not

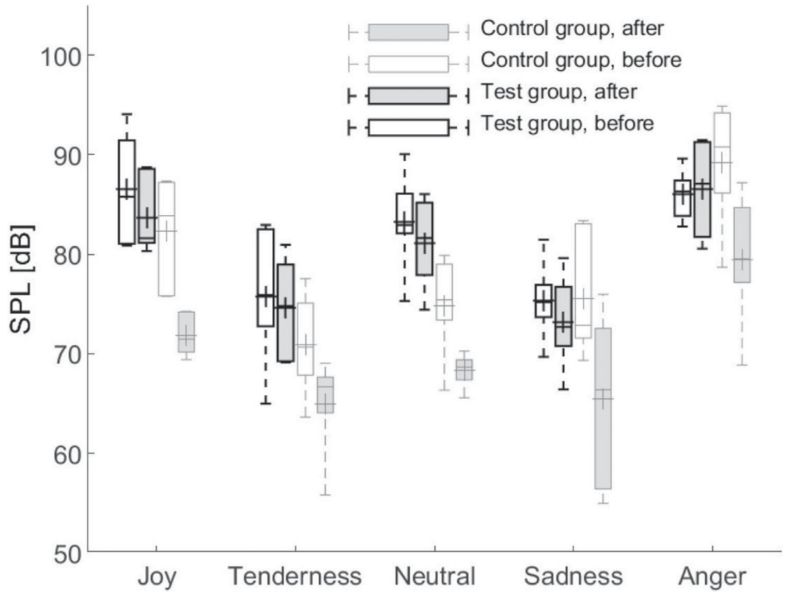

FIGURE 2. Test group SPL control before and after the training intervention and control group volume control in two separate recordings.

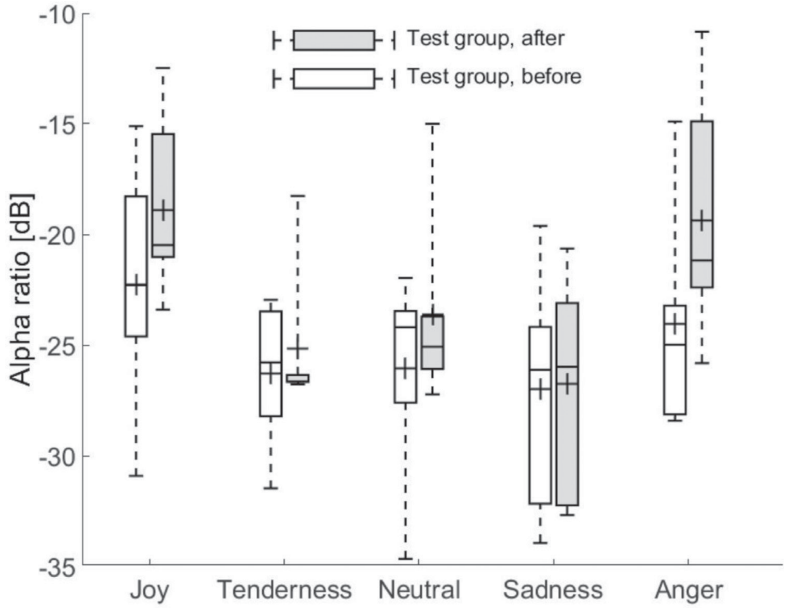

FIGURE 3. Measured Alpha ratio in the test group samples before and after the training intervention.

sure what they heard, they are more likely to select "neutral."[54] The decrease in neutral answers in the test group after condition samples can be therefore interpreted as an increase in the expressivity of the singing voice. For the control group, vowel sample recognition of emotion decreased for the after condition in all other emotion portrayals except joy. The recognition of neutral on the other hand increased. This can be interpreted as a difficulty in

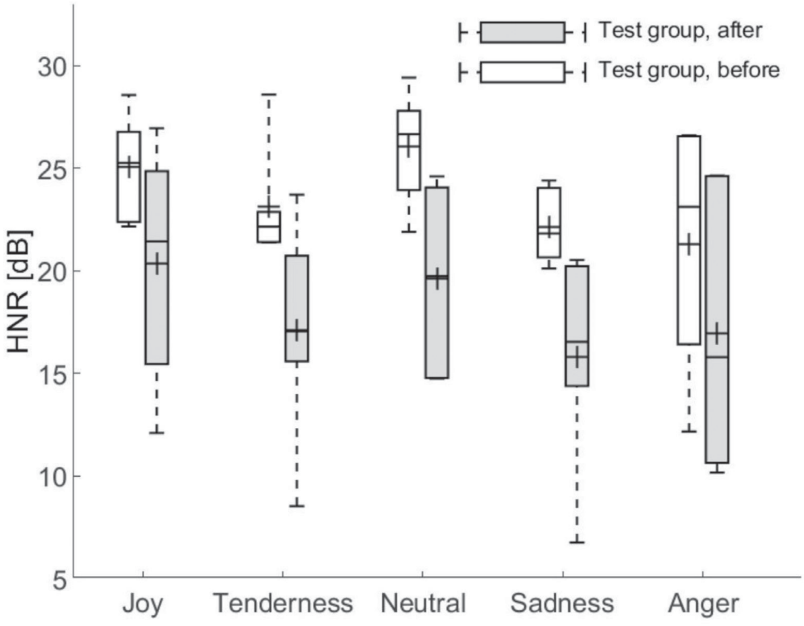

FIGURE 4. Measured HNR in the test group samples before and after training.

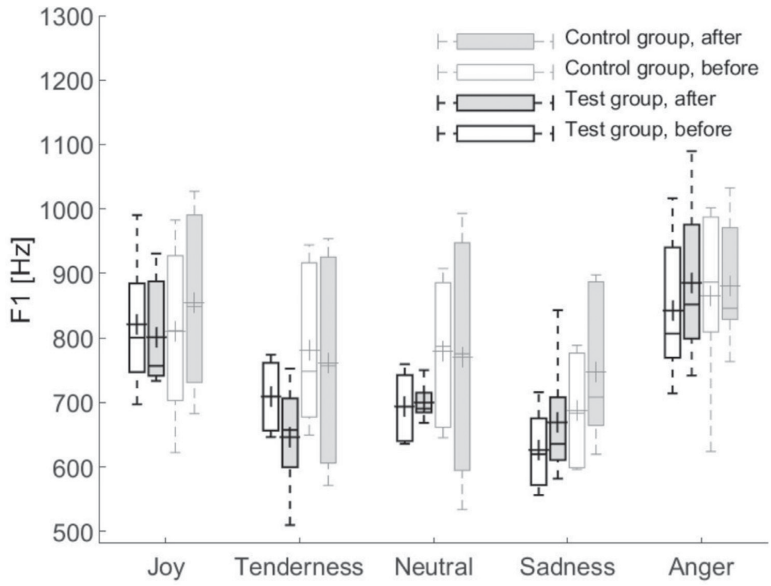

FIGURE 5. Range of F1 positioning in recorded samples.

arriving at a specific emotion appraisal on the listener side. This lends support to our hypothesis that the systemic parameter manipulation of the singing voice can help in building expressivity in contrast to a situation where the students receive just regular voice tuition without rehearsing the use of voice quality in emotional expression.

For phrases, there was no statistically significant difference found in the recognition of emotions between the test

and control groups (even though for the test group, the recognition was significantly better in the after condition as compared to the before condition). Emotions were already recognized better from phrases than vowels at the beginning. This can be interpreted as resulting from the fact that phrases contain more acoustic variables and thus more redundancy, which makes it easier to convey emotions through them. Furthermore, the training technique used targeted skills to vary loudness and voice quality, not various other aspects relevant at the level of the phrase. As such, the result that the recognition of emotions improved for vowels and not for phrases supports the suggestion that the result is due to the training intervention and not other intervening factors (like various elements of the ordinary actors' training obtained during the 7-week period).

For the test group vowel samples in the before condition, the Friedman test found statistical significance in differentiating emotion expression for F1, SPL, HNR, F3, jitter(s), and shimmer(s). In the samples recorded after the teaching intervention only, F1, SPL, HNR, and Alpha ratio were found to be statistically significant in distinguishing the emotion expressions. For the control group, F1, SPL, and F4 showed statistical significance in differentiating the emotions in the samples collected in the former recording, but only F1 and SPL remained statistically significant in the samples collected in the latter recording. This means that the results did not give support to our hypothesis that the number of differentiating acoustic parameters would increase. The result could be interpreted to show that rehearsing stereotypic use of voice parameters narrows down the choice of parameters. On the other hand, it may reflect a more focused and conscious use of certain key parameters and the reduction of parameters that may not be related to expression as such or that may not have any relevance to emotion recognition.

The exercise routine we used to train the parameters of emotional expression in the sound signal included exercises for volume control, phonation balance, articulation, and extreme vibrato, but excluded resonant voice training and distortions of sound signal. Resonant voice with a singer's ring at around 3000 Hz has been identified as a typical marker of a quality voice that can carry over a large orchestra.[32,55–57] The resonance features can be explored in a context of emotional expression by altering the carrying power of the voice from a very resonant, sonorous voice to a nonresonant, muffled sound. Articulation and vocal tract settings affect the resonance frequencies of the vocal tract.[15] Clear articulation, which can be viewed as the voluntary maximation of the intelligibility of one's singing for the benefit of the listener, is another mark of a good singer. There are, however, several different configurations of the vocal tract that can lead to perceptually similar sounding vowels and consonants, depending on the pitch used and individual singing style.[58] As the resonance and articulation have plenty of overlapping anatomical and physiological elements and as time was scarce with our 7-week training regime, we felt that it was not efficient in a teaching situation to try to compartmentalize these two functions in the context of emotional expression through sound qualities. This is especially the case since jaw and lip movements have been widely studied and linked to both an increase or decrease in clarity of articulation and the resonance of sound, which in turn have been linked to the perception of loudness.[44,59–68] The five bodily systems needed for singing are the nervous, breathing, phonation, resonance, and articulatory systems.[47] Already for the perceptually seemingly easy task of volume control, the students need to control these systems simultaneously. It was our pedagogical challenge to adjust the exercises of the "parameter modulation" model to correspond with the students' own way of using and coordinating these systems. Our solution to this problem was to use the most simple instruction of moving clearly definable anatomical organs on the vertical, horizontal, and anterior/posterior axis, and, when this was not possible, to use vocal sounds that could be easily segregated, demonstrated, and mimicked using the auditory feedback loop.[47,69] In this way, we felt that the possibilities of misunderstandings due to a lack of conceptual knowledge or terminology could be avoided and the idea of the "parameter modulation" technique could be conveyed impartially to all participants.

There is so much individual variation in singing voices and the way that we use them that there will never be a ready-made mold that will fit all singers. However, we can say that based on previous research and the results of this investigation, SPL/loudness plays a large role in emotion expression, phonation balance is one of the defining factors, and formant frequencies is another.[15–18] When we move on to longer units of sound, such as phrases, verses, or songs, there will be many more factors that will contribute to the expressivity of the singing voice.

This study is limited by the small number of participants, and therefore the potential effect of individual factors cannot be excluded. On the other hand, the recognition rate was very similar for both the test and the control group samples before training, which indicates that the groups expressed emotions in a similarly recognizable way. The fact that the (singing) participants were acting students also may affect the results, since they can be expected to have a better baseline ability to convey emotion through their voices. However, the acting students were chosen as subjects as this kind of training fits their curriculum, and they are supposed to be more motivated and also more capable of exploring their vocal resources. The samples were recorded in two different locations, but with similar equipment and similar (studio) surroundings. The listener-participants used their own devices in the listening test, which creates possible fluctuations in audio quality from one participant to another. However, the flexible use of listening equipment allowed us to have significantly more listener-participants than the use of fixed listening equipment.

CONCLUSIONS

The parameter modulation exercise routine helped in expressing emotions vocally, as indicated by the increase in the correct recognition of emotions in the listening tests.

The 7-week training program helped the acting students to broaden their voice use and tactically implement the usage of F1 positioning, SPL, HNR, and Alpha ratio in the expression of emotion in comparison to the control group, who did not receive instructions in parameter modulation.

Training different voice qualities may help in expressing emotions in the singing voice.

REFERENCES

1. Juslin PN, Laukka P. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol Bull*. 2003;129:770–814. https://doi.org/10.1037/0033-2909.129.5.770.
2. Scherer KR. Expression of emotion in voice and music. *J Voice*. 1995;9:235–248. https://doi.org/10.1016/S0892-1997(05)80231-0.
3. Stanislavski K, Männistö M. *Näyttelijän Työ*. Helsinki: Kustannusosakeyhtiä Tammi; 2015.
4. Karp M, Holmes P, Tauvon KB, et al. *The Handbook of Psychodrama*. London and New York: Routledge; 2005. https://doi.org/10.4324/9780203977767.
5. Sweet B, Parker EC. Female vocal identity development: a phenomenology. *J Res Music Educ*. 2019;67:62–82.
6. Quirin M, Kazén M, Kuhl J. When nonsense sounds happy or helpless: the implicit positive and negative affect test (IPANAT). *J Pers Soc Psychol*. 2009;97:500–516. https://doi.org/10.1037/a0016063.
7. Mesiä S. *Developing Expertise of Popular Music and Jazz Vocal Pedagogy Through Professional Conversations: A Collaborative Project among Teachers in Higher Music Education in the Nordic Countries*. vol. 77. Helsinki: The Sibelius Academy of the University of the Arts Helsinki Studia Musica; 2019.
8. Titze IR, Verdolini Abbot K. *Vocology the Science and Practice of Voice Habilitation*. Salt Lake City, Utah: NCVS; 2012.
9. Jansens S, Bloothooft G, de Krom G. Perception and acoustics of emotions in singing. *Proc Fifth Eur Conf Speech Commun Technol*. 1997;0:0–3. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=9747D0A838F2790BD0161DCF94739C2E?doi=10.1.1.56.8871. Accessed January 15, 2021 .
10. Livingstone SR, Choi DH, Russo FA. The influence of vocal training and acting experience on measures of voice quality and emotional genuineness. *Front Psychol*. 2014;5:1–13. https://doi.org/10.3389/fpsyg.2014.00156.
11. Hakanpää T, Waaramaa T, Laukkanen A-M. Comparing contemporary commercial and classical styles — emotion expression in singing. *J Voice*. 2019. https://doi.org/10.1016/j.jvoice.2019.10.002.
12. Sundberg J, Iwarsson J, Hagegård H. A singer's expression of emotions in sung performance. In: Fujimura O, Hirano M, eds. *Vocal Fold Physiology: Voice Quality Control*. Stockholm Sweden: Singular Pub Group; 1994;35:81–92. http://www.speech.kth.se/prod/publications/files/qpsr/1994/1994_35_2-3_081-092.pdf.
13. Sundberg J. Emotive transforms: acoustic patterning of speech its linguistic and physiological bases. *Phonetica*. 2000;57:95–112.
14. Eyben F, Salomão GL, Sundberg J, et al. Emotion in the singing voice —a deeperlook at acoustic features in the light ofautomatic classification. *EURASIP J Audio, Speech Music Process*. 2015;2015:19. https://doi.org/10.1186/s13636-015-0057-6.
15. Scherer KR, Trznadel S, Fantini B, et al. Recognizing emotions in the singing voice. *Psychomusicol Music Mind Brain*. 2017;27:244–255. https://doi.org/10.1037/pmu0000193.
16. Scherer KR, Sundberg J, Fantini B, et al. The expression of emotion in the singing voice: acoustic patterns in vocal performance. *J Acoust Soc Am*. 2017;142:1805–1815. https://doi.org/10.1121/1.5002886.
17. Scherer KR, Sundberg J, Tamarit L, et al. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Comput Speech Lang*. 2015;29:218–235. https://doi.org/10.1016/j.csl.2013.10.002.
18. Sundberg J, Salomão GL, Scherer KR. Analyzing emotion expression in singing via flow glottograms, long-term-average spectra, and expert listener evaluation. *J Voice*. 2019. https://doi.org/10.1016/j.jvoice.2019.08.007.
19. Darwin C, Ekman P. *The Expression of the Emotions in Man and Animals*. 3rd ed. 1872. https://doi.org/10.1037/10001-000.
20. Izard CE. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspect Psychol Sci*. 2007;2:260–280. https://doi.org/10.1111/j.1745-6916.2007.00044.x.
21. Izard CE. Basic emotions, relations among emotions, and emotion-cognition relations. *Psychol Rev*. 1992;99:561–565. doi:10.1037//0033-295X.99.3.561.
22. Purves D, Cabeza R, Huettel S, et al. *Principles of Cognitive Neuroscience*. 2nd ed. Sundeerland, MA U.S.A.: Sinauer Associates, Inc. Publishers; 2013.
23. Bericat E. The sociology of emotions: four decades of progress. *Curr Sociol*. 2016;23:1307–1351. https://doi.org/10.1177/0011392115588355.
24. Scherer KR. The dynamic architecture of emotion: evidence for the component process model. *Cogn Emot*. 2009. https://doi.org/10.1080/02699930902928969.
25. Kotlyar G, Morozov V. Acoustical correlates of emotional content on vocalized speech. *Sov Phys Acoust*. 1976;22:208–2011.
26. Siegwart H, Scherer KR. Acoustic concomitants of emotional expression in operatic singing: the case of lucia in Ardi gli incensi. *J Voice*. 1995;9:249–260. https://doi.org/10.1016/S0892-1997(05)80232-2.
27. Sundberg J. Expressivity in singing. A review of some recent investigations. *Logop Phoniatr Vocology*. 1998;23:121–127. https://doi.org/10.1080/140154398434130.
28. Waaramaa T, Laukkanen AM, Alku P, et al. Monopitched expression of emotions in different vowels. *Folia Phoniatr Logop*. 2008;60:249–255. https://doi.org/10.1159/000151762.
29. Waaramaa T, Alku P, Laukkanen A-M. The role of F3 in the vocal expression of emotions. *Logop Phoniatr Vocol*. 2006;31:153–156. https://doi.org/10.1080/14015430500456739.
30. Fant G. *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*. 2nd ed. Berlin/Boston: De Gruyter, Inc.; 1971.
31. Pulkki V, Karjalainen M. *Communication Acoustics an Introduction to Speech, Audio and Psychoacoustics*. West Sussex, UK: Wiley; 2015.
32. Sundberg J. *The Science of the Singing Voice*. Illinois, U.S.A.: Dekalb North Ill University Press; 1987.
33. Titze IR. *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice-Hall; 1994.
34. Titze IR, Baken RJ, Bozeman KW, et al. Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *J Acoust Soc Am*. 2015;137:3005–3007. https://doi.org/10.1121/1.4919349.
35. Welch G, Thurman L, Theimer A, et al. How your vocal tract contributes to basic voice qualities. *Bodymind and Voice*. 2000:449–469.
36. Boersma P, Weenink D. Praat. 2014.
37. Frøkjaer-Jensen B, Prytz S. Registration of voice quality. *Bruel-Kjaer Technol Rev*. 1976;3:3–17.
38. Lã FMB, Sundberg J. Pregnancy and the singing voice: reports from a case study. *J Voice*. 2012;26:431–439. https://doi.org/10.1016/j.jvoice.2010.10.010.
39. Hakanpää T, Waaramaa T, Laukkanen A-M. Emotion recognition from singing voices using contemporary commercial music and classical styles. *J Voice*. 2018;33:501–509. https://doi.org/10.1016/j.jvoice.2018.01.012.
40. Behrman A, Haskell J. *Exercises for Voice Therapy*. 2nd ed. San Diego, CA: Plural Publishing; 2013.
41. Linklater K. *Freeing the Natural Voice*. London, UK: Nick Hern Books; 2006.
42. Brown OL. *Discover Your Voice*. New York: Delmar Cenage Learning; 2007.
43. Richard M. *The Structure of Singing System and Art in Vocal Technique*. Belmon CA, U.S.A.: Wadsworth Group/Thomson Learning; 1986.

44. Myers BR, Finnegan EM. The effects of articulation on the perceived loudness of the projected voice. *J Voice*. 2015;29:390.e9–390.e15. https://doi.org/10.1016/j.jvoice.2014.07.022.
45. Edmondson JA, Esling JH. The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies. *Phonology*. 2006;23:157–191. https://doi.org/10.1017/S095267570600087X.
46. Tartter VC. Happy talk: perceptual and acoustic effects of smiling on speech. *Percept Psychophys*. 1980;27:24–27. https://doi.org/10.3758/BF03199901.
47. Seikel J, King D, Drumright D. *Anatomy and Physiology for Speech, Language, and Hearing*. 5th edition. New York, U.S.A.: Cenage Learning ; 2014.
48. Aura M, Geneid A, Bjørkøy K, et al. The nasal musculature as a control panel for singing—why classical singers use a special facial expression? *J Voice*. 2019;33:510–515. https://doi.org/10.1016/j.jvoice.2017.12.016.
49. Farrús M, Hernando J. Using jitter and shimmer in speaker verification. *IET Signal Process*. 2009;3:247. https://doi.org/10.1049/iet-spr.2008.0147.
50. Brockmann M, Drinnan MJ, Storck C, et al. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *J Voice*. 2011;25:44–53. https://doi.org/10.1016/j.jvoice.2009.07.002.
51. Orlikoff RF, Baken RJ. The effect of the heartbeat on vocal fundamental frequency perturbation. *J Speech Hear Res*. 1989;32:576–582. https://doi.org/10.1044/jshr.3203.576.
52. Titze IR. A model for neurologic sources of aperiodicity in vocal fold vibration. *J Speech Hear Res*. 1991;34:460–472. https://doi.org/10.1044/jshr.3403.460.
53. Borch DZ, Sundberg J, Lindestad P, et al. Vocal fold vibration and voice source aperiodicity in "dist" tones: a study of a timbral ornament in rock singing. *Logoped Phoniatr Vocol*. 2004;29:147–153. https://doi.org/10.1080/14015430410016073.
54. Marcel Z, Eerola T. Self-report measures and models. In: Juslin PN, Sloboda JA, eds. *Handbook of Music and Emotion Theory, Research, Applications*. Oxford, UK: Oxford University Press; 2012:187–221.
55. Sundberg J. Level and center frequency of the singer's formant. *J Voice*. 2001;15:176–186. https://doi.org/10.1016/S0892-1997(01)00019-4.
56. Lee S-H, Kwon H-J, Choi H-J, et al. The singer's formant and speaker's ring resonance: a long-term average spectrum analysis. *Clin Exp Otorhinolaryngol*. 2008;1:92–96. https://doi.org/10.3342/ceo.2008.1.2.92.
57. Barrichelo VMO, Heuer RJ, Dean CM, et al. Comparison of singer's formant, speaker's ring, and LTA spectrum among classical singers and untrained normal speakers. *J Voice*. 2001;15:344–350. https://doi.org/10.1016/S0892-1997(01)00036-4.
58. Rua Ventura SM, Freitas DRS, Ramos IMAP, et al. Morphologic differences in the vocal tract resonance cavities of voice professionals: an MRI-based study. *J Voice*. 2013;27:132–140. https://doi.org/10.1016/j.jvoice.2012.11.010.
59. Sundberg J. Articulatory interpretation of the "singing formant". *J Acoust Soc Am*. 1974;55:838–844. https://doi.org/10.1121/1.1914609.
60. Vurma A. Amplitude effects of vocal tract resonance adjustments when singing louder. *J Voice*. 2020. https://doi.org/10.1016/j.jvoice.2020.05.020.
61. Titze IR, Story BH. Acoustic interactions of the voice source with the lower vocal tract. *J Acoust Soc Am*. 1997;101:2234–2243. https://doi.org/10.1121/1.418246.
62. Acker BF. Vocal tract adjustments for the projected voice. *J Voice*. 1987;1:77–82. https://doi.org/10.1016/S0892-1997(87)80028-0.
63. Sundberg J. What's so special about singers? *J Voice*. 1990;4:107–119. https://doi.org/10.1016/S0892-1997(05)80135-3.
64. Garnier M, Henrich N, Smith J, et al. Vocal tract adjustments in the high soprano range. *J Acoust Soc Am*. 2010;127:3771–3780. https://doi.org/10.1121/1.3419907.
65. Fant G. *Acoustic Theory of Speech Production*. The Hague: Mouton; 1960.
66. Schulman R. Articulatory dynamics of loud and normal speech. *J Acoust Soc Am*. 1989;85:295–312. https://doi.org/10.1121/1.397737.
67. Dromey C, Ramig LO. Intentional changes in sound pressure level and rate: their impact on measures of respiration, phonation, and articulation. *J Speech, Lang Hear Res*. 1998;41:1003–1018. https://doi.org/10.1044/jslhr.4105.1003.
68. Wohlert AB, Hammen VL. Lip muscle activity related to speech rate and loudness. *J Speech, Lang Hear Res*. 2000;43:1229–1239. https://doi.org/10.1044/jslhr.4305.1229.
69. Tourville JA, Guenther FH. The DIVA model: a neural theory of speech acquisition and production. *Lang Cogn Process*. 2011;26:952–981. https://doi.org/10.1080/01690960903498424.