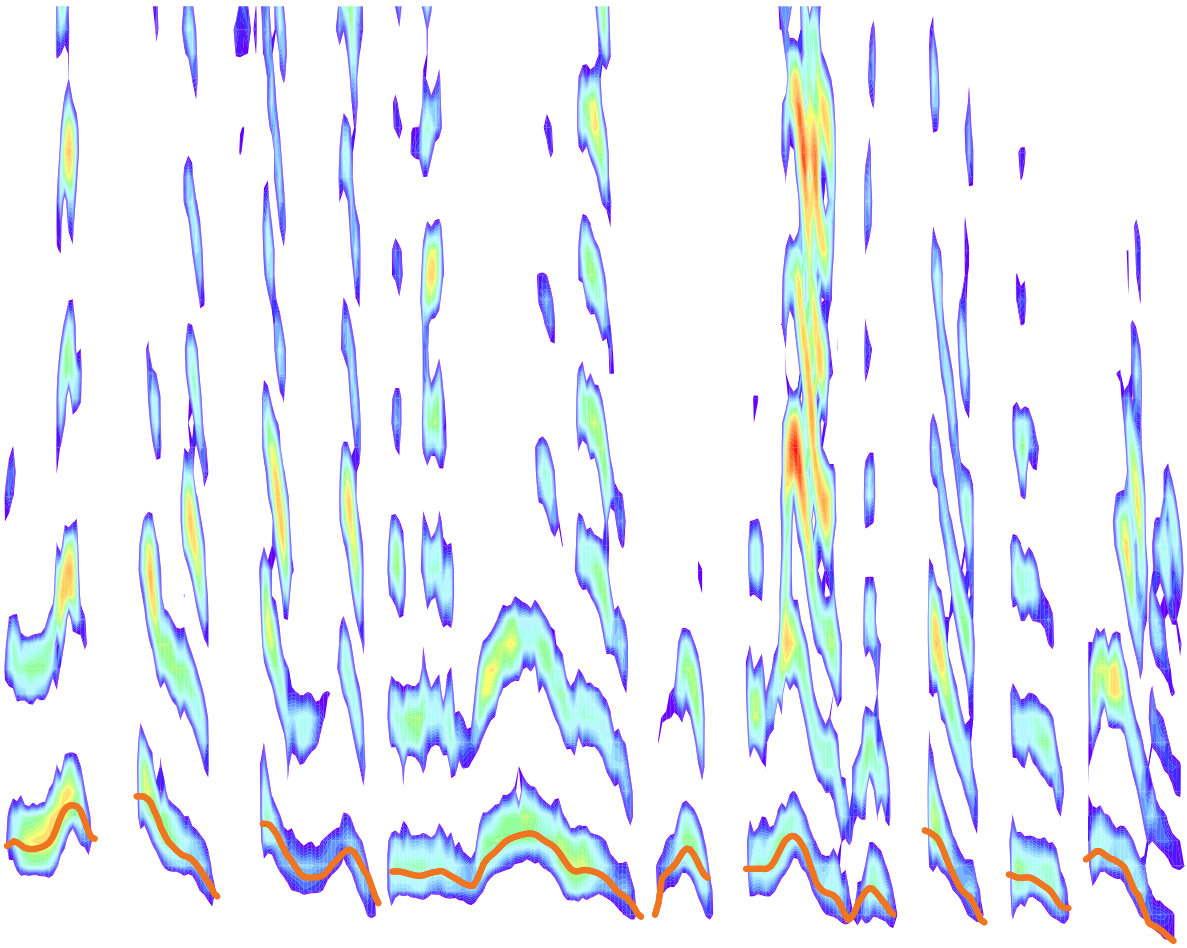
Stockholm University

Emotional Communication in the Human Voice

Henrik Nordström



Doctoral Thesis in Psychology at Stockholm University, Sweden 2019

Emotional Communication in the Human Voice

Henrik Nordström

Academic dissertation for the Degree of Doctor of Philosophy in Psychology at Stockholm University to be publicly defended on Tuesday 4 June 2019 at 13.00 in David Magnussonsalen (U31), Frescati Hagväg 8.

Abstract

Emotional communication is an important part of social interaction because it gives individuals valuable information about the state of others, allowing them to adjust their behaviors and responses appropriately. When people use the voice to communicate, listeners do not only interpret the words that are said, the *verbal* content, but also the information contained in how the words are said, the *nonverbal* content. A large portion of the nonverbal content of the voice is thought to convey information about the emotional state of the speaker. The aim of this thesis was to study how humans communicate and interpret emotions via nonverbal aspects of the voice, and to describe these aspects in terms of acoustic parameters that allow listeners to interpret the emotional message.

The thesis presents data from four studies investigating nonverbal communication of emotions from slightly different perspectives. In a yet unpublished study, the acoustic parameters suggested to communicate discrete emotions – based on theoretical predictions of how the voice may be influenced by emotional episodes – were compared with empirical data derived from listeners' judgments of actors portraying a wide variety of emotions. Results largely corroborated the theoretical predictions suggesting that previous research has come far in explaining the mechanisms allowing listeners to infer emotions from the nonverbal aspects of speech. However, potentially important deviations were also observed. These deviations may be crucial to our understanding of how emotions are communicated in speech, and highlight the need to refine theoretical predictions to better describe the acoustic features that listeners use to understand emotional voices.

In the first of the three published studies, Study 1, the common sense notion that we are quick to hear the emotional state of a speaker was investigated and compared with the recognition of emotional expressivity in music. Results showed that listeners needed very little acoustic information to recognize emotions in both modes of communication. These findings suggest that low-level acoustic features that are available to listeners in the first tenths of a second carry much of the emotional message and that these features may be used in both speech and music.

By investigating listeners recognition of vocal bursts – the kind of sounds people make when they are not speaking – results from Study 2 showed that listeners can recognize several emotional expressions across cultures, including emotions that are often difficult to recognize from speech. The study thus suggests that the voice is an even more versatile means for emotional communication than previously thought.

Study 3 also investigated emotional communication in a cross-cultural setting. However, instead of studying emotion recognition in terms of discrete categories, this study investigated whether nonverbal aspects of the voice can carry information about how the speaker evaluated the situation that elicited the emotion. Results showed that listeners were able to infer several aspects about the situation, which suggests that nonverbal expressions may have a symbolic meaning comprising several dimensions other than valence and arousal that can be understood across cultures.

Taken together, the results of this thesis suggest that humans use nonverbal manipulations of the voice to communicate emotions and that these manipulations can be understood quickly and accurately by listeners both within and across cultures. Although decades of research has investigated how this communication occurs, the acoustic parameters allowing listeners to interpret emotions are still elusive. The data from the four studies in this thesis, the methods used, and the acoustic analyses performed shed new light on this process. Future research in the field may benefit from a more standardized approach across studies, both when it comes to acoustic analysis and experimental design. This would facilitate comparisons of findings between different studies and allow for a more cumulative science within the field of emotional communication in the human voice.

Keywords: *emotion recognition, vocal expression, speech, acoustic parameters.*

Stockholm 2019
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-167973

ISBN 978-91-7797-735-3
ISBN 978-91-7797-736-0

Department of Psychology

Stockholm University, 106 91 Stockholm

EMOTIONAL COMMUNICATION IN THE HUMAN VOICE

Henrik Nordström

Stockholm
University

Emotional Communication in the Human Voice

Henrik Nordström

ISBN print 978-91-7797-735-3
ISBN PDF 978-91-7797-736-0

Printed in Sweden by Universitetsservice US-AB, Stockholm 2019

Till snart bebis

Abstract

Emotional communication is an important part of social interaction because it gives individuals valuable information about the state of others, allowing them to adjust their behaviors and responses appropriately. When people use the voice to communicate, listeners do not only interpret the words that are said, the *verbal* content, but also the information contained in how the words are said, the *nonverbal* content. A large portion of the nonverbal content of the voice is thought to convey information about the emotional state of the speaker. The aim of this thesis was to study how humans communicate and interpret emotions via nonverbal aspects of the voice, and to describe these aspects in terms of acoustic parameters that allow listeners to interpret the emotional message.

The thesis presents data from four studies investigating nonverbal communication of emotions from slightly different perspectives. In a yet unpublished study, the acoustic parameters suggested to communicate discrete emotions – based on theoretical predictions of how the voice may be influenced by emotional episodes – were compared with empirical data derived from listeners' judgments of actors portraying a wide variety of emotions. Results largely corroborated the theoretical predictions suggesting that previous research has come far in explaining the mechanisms allowing listeners to infer emotions from the nonverbal aspects of speech. However, potentially important deviations were also observed. These deviations may be crucial to our understanding of how emotions are communicated in speech, and highlight the need to refine theoretical predictions to better describe the acoustic features that listeners use to understand emotional voices.

In the first of the three published studies, Study 1, the common sense notion that we are quick to hear the emotional state of a speaker was investigated and compared with the recognition of emotional expressivity in music. Results showed that listeners needed very little acoustic information to recognize emotions in both modes of communication. These findings suggest that low-level acoustic features that are available to listeners in the first tenths of a second carry much of the emotional message and that these features may be used in both speech and music.

By investigating listeners recognition of vocal bursts – the kind of sounds people make when they are not speaking – results from Study 2 showed that listeners can recognize several emotional expressions across cultures, includ-

ing emotions that are often difficult to recognize from speech. The study thus suggests that the voice is an even more versatile means for emotional communication than previously thought.

Study 3 also investigated emotional communication in a cross-cultural setting. However, instead of studying emotion recognition in terms of discrete categories, this study investigated whether nonverbal aspects of the voice can carry information about how the speaker evaluated the situation that elicited the emotion. Results showed that listeners were able to infer several aspects about the situation, which suggests that nonverbal expressions may have a symbolic meaning comprising several dimensions other than valence and arousal that can be understood across cultures.

Taken together, the results of this thesis suggest that humans use nonverbal manipulations of the voice to communicate emotions and that these manipulations can be understood quickly and accurately by listeners both within and across cultures. Although decades of research has investigated how this communication occurs, the acoustic parameters allowing listeners to interpret emotions are still elusive. The data from the four studies in this thesis, the methods used, and the acoustic analyses performed shed new light on this process. Future research in the field may benefit from a more standardized approach across studies, both when it comes to acoustic analysis and experimental design. This would facilitate comparisons of findings between different studies and allow for a more cumulative science within the field of emotional communication in the human voice.

Sammanfattning på svenska

Förmågan att kunna kommunicera och förstå känslor är en viktig del av vår sociala tillvaro. Denna kommunikation ger oss information om hur människor i vår omgivning kan tänkas bete sig vilket gör att vi bättre kan anpassa vårt eget beteende på ett lämpligt sätt. När lyssnare tolkar innehållet i talad kommunikation så tolkar de inte bara de ord som sägs, det verbala innehållet, utan även all annan information som finns i rösten, dvs. det icke-verbala innehållet. En stor del av det icke-verbala innehållet i rösten tros förmedla information om talarens känslotillstånd. Syftet med denna avhandling var att studera hur människor kommunicerar och tolkar känslouttryck med icke-verbala förändringar i rösten samt att beskriva de akustiska egenskaper som hänger ihop med dessa förändringar. Avhandlingen presenterar resultat från fyra studier som alla undersökte olika aspekter av hur denna typ av kommunikation går till. I alla fyra studier blev skådespelare instruerade att uttrycka olika känslor varpå lyssnare fick göra olika bedömningar av dessa uttryck.

Baserat på antaganden om hur rösten påverkas av talarens känslotillstånd har tidigare forskning försökt göra förutsägelser om hur akustiska egenskaper i rösten hänger ihop med olika känslouttryck. I en studie som ännu inte är publicerad jämfördes dessa akustiska egenskaper med dem som lyssnarna i denna studie använde för att kategorisera känslouttryck. För många känslor och akustiska mått stämde resultaten väl överens med de förutsägelser som tidigare forskning har gjort men det fanns också många avvikelser. Dessa avvikelser kan tyda på att de teoretiska antagandena från tidigare studier inte räcker till för att beskriva hur människor gör för att förstå icke-verbala känslouttryck.

I den första av de tre publicerade studierna i avhandlingen, Studie 1, undersöktes den allmänna föreställningen om att människor är bra på att snabbt höra om en röst eller musik uttrycker en viss känsla. Denna studie visade att lyssnare kan känna igen flera olika känslouttryck även om de bara får höra en väldigt kort del av talet eller musiken. Detta resultat tyder på att den akustiska information som är tillgänglig redan i de första tiondelarna i ett ljud kan användas för att känna igen känslor i både tal och musik.

Studie 2 undersökte vilka känslouttryck som lyssnare kan känna igen då känslor kommuniceras med olika typer av läten, det vill säga sådana ljud som vi gör när vi inte talar. Studien undersökte denna kommunikation tvärkulturellt vilket innebär att skådespelarna och lyssnarna var från olika län-

der. Resultaten visade att lyssnarna kunde känna igen många olika känslouttryck – fler än vad som vanligtvis kan kännas igen i tal – vilket tyder på att rösten kan förmedla fler känslor än vad man tidigare trott, även då uttryckaren och lyssnaren är från olika länder.

Även Studie 3 undersökte hur känslor kommuniceras tvärkulturellt. Men istället för att undersöka hur lyssnare tolkar olika kategorier av känslor, som de tidigare tre studierna, undersökte denna studie om lyssnare kan förstå hur talaren upplevde den situation som orsakade känslan. Här fick skådespelarna föreställa sig att de befann sig i olika situationer som var tänkta att väcka känslor. Sedan fick lyssnarna bedöma hur de trodde att talaren upplevde situationen baserat endast på deras röst. Resultaten visade att lyssnarna kunde bedöma flera aspekter av hur situationen upplevdes av talaren. Detta visar att icke-verbala aspekter av rösten kan ha en symbolisk mening eftersom de kan förmedla information utöver de typer av känslouttryck som tidigare forskning har studerat.

Sammanfattningsvis så tyder dessa studier på att människor förändrar icke-verbala egenskaper i rösten för att förmedla känslor och att dessa förändringar kan tolkas snabbt och korrekt av lyssnare, även tvärkulturellt. Trots att decennier av forskning har försökt hitta akustiska mått som beskriver hur denna kommunikation går till så är de akustiska sambanden som kan förklara hur känslor kommuniceras fortfarande otydliga. De resultat som omfattas i denna avhandling, de metoder som har använts, och de akustiska samband som presenteras, kastar nytt ljus på hur denna kommunikation går till. Framtida forskning kan troligtvis ha nytta av att använda mer standardiserade tillvägagångssätt, både när det gäller design av experiment såväl som för akustiska analyser. På så sätt skulle resultat från olika studier lättare kunna jämföras med varandra vilket skulle göra så att nya studier kan bygga vidare på tidigare resultat. Detta skulle kunna leda till att vi får mer kunskap om hur människor förmedlar och förstår känslouttryck i rösten.

Acknowledgements

Petri Laukka
Mats E. Nilsson
Stefan Wienes
(Anders Sand)

The friends from before and after science

Family

Maia. I love you.

List of studies

This doctoral thesis is based on the following studies:

I. Nordström, H., & Laukka, P. (in press). The time course of emotion recognition in speech and music. *Journal of the Acoustical Society of America*.

II. Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Chui, W., … & Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology, 4*, 353.

III. Nordström, H., Laukka, P., Thingujam, N. S., Schubert, E., & Elfenbein, H. A. (2017). Emotion appraisal dimensions inferred from vocal expressions are consistent across cultures: A comparison between Australia and India. *Royal Society Open Science, 4*(11), 170912.

In Chapter 3, data from a yet unpublished study on the acoustic correlates of emotional communication in speech is presented.

Contents

1. General introduction

Whenever we hear a person talk, we make inferences about the speaker from how the voice sounds. The voice may reveal physical characteristics such as age, sex, health status, intoxication, and tiredness. It may also reveal social characteristics such as education, status, whether the language is the speaker's mother tongue, or in which part of a country they were brought up. The voice may reveal traits such as intelligence and personality, and also psychological states such as, stress, truthfulness, and last but not least; emotions (Kreiman & Sidtis, 2011). The aim of this thesis was to investigate how humans communicate and interpret emotions via nonverbal aspects in the voice.

When we talk to each other, we do not only interpret the words that are said, the *verbal* content, but we also interpret the information in how the words are said, the *nonverbal* content. Among laymen, there is a general belief that the nonverbal aspects, or the "tone of voice", may reveal the speaker's true intentions, their inner state, and how they really feel about something. Often when I talk to people about the topic of my research they say that they think the tone of voice says "more" than what the person says. This observation has been a hot topic in politics and public speaking ever since the ancient schools of rhetoric (Aristoteles, Cicero, Quintilian, cited in Scherer, 2018) and in psychology, biology and linguistics ever since these faculties of science branched off from philosophy. The interest in the voice and the nonverbal information it conveys seem to raise interest in laymen and researchers alike. Perhaps this interest reflects the common sense notion that the voice is conveying something else than the verbal content, something more truthful that cannot be completely hidden. For this reason, the voice has been called 'the mirror to our soul'' (Sundberg, 1998).

Another thing that people often bring up when I talk to them about my research are personal anecdotes of communication with animals. It seems, they say, that for example their pet can sense what they are feeling, perhaps because the animal interprets changes in the voice. Common sense seems to imply that humans and other animals have the same language when it comes to nonverbal communication. In fact, Darwin made this observation in his books (e.g. The Expression of the Emotions in Man and Animals, Darwin, 1872) and recent studies have shown that humans can recognize emotional arousal in vocalizations from all land-living vertebrates (Filippi et al., 2017). Some researchers have even suggested that non-human animals may serve as

a better model than humans of how emotions affect the voice because animal expressions, contrary to human speech, are assumed to be free of control and therefore represent more direct expressions of emotions (Briefer, 2012). In support of this notion, research suggests that the acoustic patterns of animal vocalizations are mainly determined by phylogenetic rather than ontogenetic factors (Owren, Amoss, & Rendall, 2011; Wheeler & Fischer, 2012). In other words, all members of a non-human species use more or less the same sounds to communicate, and the message is mainly emotional. Studies show that even though human language is extremely flexible and obviously culturally dependent, many of the sounds that humans produce are nonverbal, and these sounds are much more similar across cultures and even species than other parts of our communication (Jürgens, 2009).

It is well established that the neural networks that humans use to master the flexibility of language are not completely separated from the much older networks responsible for the nonverbal aspects of communication (Ackermann, Hage, & Ziegler, 2014; Jürgens, 2009). As language and speech depends on a pathway connecting the motor cortex to neurons controlling the muscles of the voice apparatus, another separate neural circuit located in the brain stem has much more influence on nonverbal vocalisations or "affect bursts" (Jürgens, 2009). These separate pathways may explain the fact that saying a word or a sentence is obviously under voluntary control while affective bursts are not; it is much more difficult to cry or laugh at will than saying a sentence. Actors usually need years of training to produce credible expressions of emotions and this training commonly involves techniques of emotion induction such as the Stanislavski system or "method acting". The purpose of these techniques is to mobilize the actor's thoughts and imagination and that these conscious processes will mobilize other less conscious processes, both physical and psychological. Instead of learning to copy the vocalizations associated with an emotion, it seems that it is easier for actors to first induce the emotional state and then let the unconscious processes do the rest (Provine, 2012).

These and other observations has led scientists to believe that humans and other animals have at least partly shared neural networks for emotional communication. At some point in our evolutionary past though, early hominids may have begun to use these networks to develop a more complex type of communication, a kind of proto-language consisting of hmms and grunts acquiring meaning other than that of the original emotional message (Mithen, Morley, Wray, Tallerman, & Gamble, 2006; Scheiner & Fischer, 2011; Wallin, Merker, & Brown, 2001). Once the communicative function of such expressions has been established, expressing and interpreting them could clearly have a survival value and therefore continue to increase in complexity. Sequences of hmms and grunts may have begun to acquire syntax- and melody-like intonations allowing for more effective communica-

tion, in turn making it possible for hominids to live in larger, safer groups. Therefore, it could be argued that the signals of emotional communication are at the very basis of the development of language and thus of the success of our species.

Motivation and research gaps in research on emotional communication in the voice

In recent years, there has been a rapid increase in research focused on how emotions are communicated in the voice. Many of these studies have aimed to develop automatic emotion classifier systems (Mustafa, Yusoof, Don, & Malekzadeh, 2018). These studies use databases of recordings of acted or non-acted speech with the aim to train machines to become as good as (or better than) humans to understand emotions in the voice. Besides the technical aspects needed for success in this endeavour (which I leave to others to describe), there are at least three aspects where the current thesis may contribute. First, to appreciate what humans are capable of doing when it comes to emotion recognition. Second, to understand how we do what we do. Third, to find ways to separate the kinds of emotional information that humans do understand from those we do not. My hope is that this will allow the empathic machines of the future to base their knowledge on, from a human perspective, more sensible information.

However, the findings and ideas presented in this thesis may not only be of interest to machines, but also to humans. The curiosity driving basic research, the will to understand ourselves, has led many researchers over the years to throw themselves into the field of emotional communication. Also, as we come closer to understanding the processes involved when communication does work this may also tell us something about when it does not. Therefore, these findings may not only appeal to curiosity but may also be useful in clinical settings, in entertainment (acting), and perhaps to improve communication, and thus understanding, between people from different cultures.

Two underlying assumptions of emotional communication (whether done by humans or machines) are that emotions have predictable effects influencing the voice and that the interpreter can use this information to infer the emotional message of the speaker. If these assumptions are justified, there must be a set of vocal qualities that can be used to associate measurable acoustic parameters to specific emotion expressions. During decades of research, much time and effort has been put into discovering the acoustic parameters that represent the vocal qualities listeners use to encode the emotional message of a voice. Descriptions of such parameters are usually made in the form of "acoustic parameter-patterns". This concept stresses the no-

tion that there is no single feature of the voice that listeners can rely on to understand an expression. Rather, listeners have to attend to a combination of features, or "cues", that allow them to recognize specific emotions.

One aim of this thesis was to evaluate how close we have come to discovering the acoustic parameter-patterns allowing listeners to recognize specific emotions in the voice. In Chapter 3, I give an overview of this research. In that chapter, I also attempt to contribute to the field by presenting results from a yet unpublished study investigating the acoustic parameter-patterns that a group of Swedish listeners used to infer several emotions and compare them with those suggested by previous research. The aim was thus to test how well the parameter-patterns suggested by the literature corresponded to the ones used by this group of listeners. Although the literature is vast, it was surprisingly difficult to find easily testable predictions. Predictions based on theoretical work could only give a vague picture of the acoustic parameter-patterns because they usually suggest them in the form of "high/medium/low" for specific emotion- and parameter-combinations. Empirical studies, on the other hand, were difficult to compare because they differ widely in their stimulus materials, experimental design, and in the acoustic analyses performed. Although this is far from a novel observation, another aim of this thesis thus became to suggest ways to analyse and present data so that the findings may be more easily tested and compared by future research.

In Chapter 4, I present an overview of the literature on the time-course of emotion recognition. Common sense tells us that we are quick to hear an emotional voice and that we don't need much acoustic information to determine what state a speaker is in. Ending the overview, I present my own contribution to the field by describing the findings of Study 1 exploring how much acoustic information listeners need to reliably classify vocal and musical expressions. This study contributes to the field by including a wider range of emotions and a more fine-grained resolution in the time domain than previous studies. Also, the comparison with music highlights that the voice is not the only tool people use to communicate emotions. The acoustical signals listeners use to infer emotions from very brief utterances might thus be generalized to other means of auditory emotion communication.

Next, Chapter 5 goes on to present a brief overview of the literature on emotion recognition of non-linguistic expressions. Non-linguistic expressions comprise all the variety of sounds that humans produce when we are not speaking, such as laughing, crying or screaming. These types of sounds are not restricted by the rules of language and are therefore thought to be especially salient carriers of the emotional message. In that chapter, I present the results of Study 2 investigating the limits of what listeners are able to perform in terms of vocal emotion recognition. To this end, I used non-linguistic expressions of 18 emotions uttered by people from four different

cultures. This study gives new insights into the capabilities of the voice as a carrier of emotional information. It also highlights that the generally low recognition rates of some emotions in the literature may be remedied by the use of non-linguistic expressions rather than speech.

Similar to Chapter 5, Chapter 6 also investigates the boundaries of what type of emotional information listeners can perceive in a voice. In that chapter I use the assumptions of the "appraisal theory of emotion" to speculate that the emotional message in speech may not be limited to emotion categories such as "happy" or "sad", or to ratings of how aroused the speaker was, but may also include inferences about the situation that led to these emotional states. Here, I present data from Study 3 investigating listeners' ability to infer several aspects of the speakers' evaluations of the events that may have triggered the emotion. Similar to Study 2, this study also used a cross-cultural design but this time listeners heard expressions of speakers from another culture as well as from their own. The contribution of this study may be to encourage research to look for other ways to think about emotion recognition, especially in cross-cultural studies in which emotion words may have slightly different meanings in different cultures and/or languages.

Research questions

Each chapter of this thesis will investigate one of the broad research questions listed below. The chapters begin with an overview of the literature and then go on to present the contribution of one of the four empirical studies relevant for each question.

- How are emotions communicated in speech? (Chapter 3)
- How much acoustic information do listeners need to infer emotions from speech and music? (Chapter 4 and Study 1)
- What emotions can listeners infer from non-linguistic vocalizations? (Chapter 5 and Study 2)
- Can listeners infer appraisal dimensions across cultures from speech? (Chapter 6 and Study 3)

2. Defining emotions

Although this thesis is not investigating emotions per se, but rather *communication* of what we commonly refer to as emotions, it seems necessary to at least briefly mention the many controversies surrounding the term.

Among the mechanisms that psychology tries to describe – such as perception, behaviour, thinking, learning, decision-making, and memory etc. – emotion may be the most intriguing yet most elusive. The paradox of emotions is that they seem so obvious and self-evident at one hand; everyone knows what they are when examined introspectively, yet they have been extremely difficult to define scientifically. Emotions are not only what makes us feel good or bad. They change how we perceive the world, how we think, learn, remember, decide and behave. Emotions are deeply personal and volatile, whenever we try to describe them they seem to lose their original meaning and lessen the experience, and they are closely related to our personal goals and values.

Although (or perhaps because) every human being can agree on the impact and importance of emotions, there is no consensus in the scientific literature of how they should be defined and this debate will continue (Adolphs, 2016; Barrett, 2014; Moors, 2014, 2017; Tracy, 2014; Tracy & Randles, 2011).

A scientific definition needs to specify the necessary and sufficient conditions for an exemplar to be categorised as belonging to that phenomenon (Moors, 2017). In a general sense, the process of defining a phenomenon scientifically comprises four steps. In the first step, researchers try to demarcate the phenomenon in a descriptive sense using common words that people use in everyday life. This way, a list of words describing the phenomenon's superficial features is drawn up. In the second step, an explanation of the phenomenon is starting to develop. These explanations may be structural ones that define what elements the phenomenon is comprised of, and/or a causal-mechanistic one that defines what causes the phenomenon and how. In the third step, these explanations are tested empirically in the hope of finding an explanation that is common to all instances of the phenomenon. If an explanation holds for rigorous testing, then there can be a fourth step in which the superficial list of features from the first step is replaced by the empirically supported explanations that demarcate the phenomenon. Borrowing an example from Moors (2017); in the first step water may be de-

fined as a liquid that falls from the sky and runs through rivers. In the second step a structural definition may be suggested; that water consists of one hydrogen- and two oxygen atoms. In the third step this definition is tested by collecting samples of water (from rain, from rivers) to determine if all instances of what is called water have this composition. If the observations are consistent with the structural definition, H_2O can replace the list of common sense words as the scientific definition of water.

In the case of emotion however, this process has proven difficult. Following the steps described above, early definitions used common language to describe emotions as one or several aspects (called components by later theorists) of the emotional process as observed in others or experienced by oneself.

Charles Darwin (1872) was the first to define emotions as basic biological functions shared with other animals. He defined emotions from the observation that some expressions seemed to be apparent in both humans and other animals. Although he mainly used these observations to support his claim that humans and other animals were evolved from the same ancestor, his thoughts have had a huge impact on the theories of emotion. He proposed that emotional expressions are universally recognized not only in humans, but also in closely related mammals such as great apes as well as domestic animals such as cats and dogs. Darwin did not, however, define what he meant by emotions further than to describe the final outcome, the expressions, and this description was solely based on his own intuition and a high degree of anthropomorphizing stereotypical expressions.

William James (1884) also tried to define emotions from a common sense description but instead of focusing on the expression, he focused on the subjective feeling that accompanies an emotional episode. James argued that the physical reaction to a situation, for example the arousal caused by a potential threat, preceded the feeling. Thus he defined emotions as the subjective experience, or perception, of the physiological changes that typically occurs in specific situations.

Today, the definition may be said to have reached the second and third steps to a scientific definition. In other words, structural and mechanistic definitions are being empirically tested but a consensus that would fulfil the fourth step, in which these explanations can completely replace the list of commons sense descriptors is nowhere to be seen, at least in the near future.

Most theorists, however, agree that a structural definition should describe an internal response that is coupled with several “components”; (1) a cognitive evaluation of the situation, a physiological change in the (2a) central and (2b) peripheral (autonomous) nervous systems (CNS, ANS) that is seen as a (3) preparation for action or adaptation to the situation, (4) the resulting action tendencies (approach/avoidance behaviour) and (5) a subjective feeling (Lench, 2018; Moors, 2017; Sander & Scherer, 2009).

By defining emotion as an *internal response*, theorists intend to stress that emotions should be seen as a reaction that occurs within the individual and that this reaction is not random but rather directed toward a specific object or event. The emotion is thus directed toward the object or event that triggered it. This means that the emotional response is depending on the interaction between the individual and its environment. The event might be something that happens to the individual or someone else, the individual's own actions or the actions of others. It might be that the individual is struck by new information or acts of nature. Events triggering emotions might also be internal processes such as memories, ideas, imagination or even voluntary decisions to experience certain emotions. Illness, drugs, or hormonal changes are also internal events that can lead to emotional episodes. Objects that trigger emotions might thus be seen, heard, felt, tasted, smelled, and involuntarily or voluntarily imagined.

In this definition, the emotional response is said to be coupled with an evaluation, or appraisal of the situation. This implies that the type of emotion elicited is depending both on the emotional trigger and the mental setting of the individual. Thus, the situation and its social contexts together with the individual's goals, expectations, values, locus of control, and experiences will influence the valence, strength, complexity, nuance, and controllability of the emotional response.

The definition also states that a change should occur in the individual's physiological condition. This change may help the individual to cope with the emotion eliciting object or event by preparing for action. The physiological response to an emotional event in the ANS can be measured for example as heart rate, blood pressure and skin conductance, and the response in the CNS are most commonly measured with electroencephalography (EEG) or functional magnetic resonance imaging (fMRI).

Action tendencies, sometimes called the "motivational" component, can be observed and measured as approach or avoidance behaviour of the individual subjected to the emotion in relation to the emotion-eliciting object. Some theorists suggest that action tendencies should be seen as a part of the physiological component while others suggest that it should be seen as a separate component. The reason that action tendencies are sometimes seen as part of the physiological component is that the physiological response is a preceding step to take action and sometimes these tendencies are not visible in the behaviour but rather measurable changes in the physiology. The startle response, an unconscious reflex to sudden or threatening events, is a good borderline example. If the startle response is strong enough it is visible as avoidance behaviour but in the typical experimental setting, where approach/avoidance behaviour is usually not required, the startle response can only be observed as electrical activity via electromyography (EMG) in the muscles involved.

The component of subjective feeling, or emotional experience, is obviously not directly measurable but is still at the philosophical core of most theories of emotion. Some researchers have claimed that the emotional experience *is* emotion, and that all other components are just correlates of this experience. For example, Lieberman (2018) argues that an episode in which all the components described above are displayed except the subjective feeling (i.e. acting) would not be categorized as an emotional episode. A counter argument to this claim is that the activation of the other components *causes* the conscious feeling of emotion (James, 1890). However we choose to interpret the causal relationship between the observable and non-observable aspects of emotion, the subjective feeling component serves the purpose to reflect the everyday experience people have of an emotional episode.

Some theories, inspired by Darwin's early definition, also include a sixth component describing a specific sort of motor tendency that serves the purpose of signalling the emotional state to other individuals via emotion-specific and cultural-independent expressions. The theoretical assumption is that an individual's emotional state affects physiology and action tendencies in such a way that it is visible/audible via several cues that observers use to infer the internal state of the expresser. Following Darwin, the theories that include expressive behaviours as part of the definition of emotion usually give this component a central role in their investigations.

Historically, some or only one of these components has served to define emotions. Emotions have been defined as only the subjective feelings component (James, 1890), the peripheral (ANS) part of the physiological response (Watson, 1919), the central (CNS) part of the physiological response component (i.e., evolutionary brain modules as part of an affect program; Tomkins, 1962), the action tendencies and the subjective feeling component (Frijda, 1986), all components except the cognitive component (Izard, 1972; Lang, 1994), and all components except the signal/expression component (Zeelenberg & Pieters, 2006).

Current definitions usually include all of these components but argue about their relative importance and how specific components should be defined. Most theories also accept the common notion that emotions are expressed and recognized as discrete categories, at least in some sense, but the lingering controversy seems to be if these categories should be seen as the fundamental building blocks of emotion, i.e. as universal, biologically hard-wired categories or not.

Emotion expressions: Basic versus constructed view

Darwin (1872) suggested that a small set of emotion expressions are universal, shared with other animals, and have their origin in adaptations to specif-

ic challenges and opportunities in our evolutionary past. Almost a hundred years after they were presented, these ideas were popularized in psychology by Silvan Tomkins in his theory of evolutionary brain modules (Tomkins, 1962) and inspired Paul Ekman and Carroll Izard (Ekman, Sorenson, & Friesen, 1969; Izard, 1971) to empirically test Darwin's claims. Ekman and Izard collected data from several cultures around the world in experiments where participants were asked to classify pictures of facial expressions by selecting an appropriate label from a list of emotion terms. Results from these studies were the first to show that facial emotion expressions could be recognized with above chance accuracy across cultures. In what is now called the "classical" or "basic" view of emotion (Ekman, 1992), these findings together with hundreds of replications (Elfenbein & Ambady, 2002; Juslin & Laukka, 2003) are still viewed as the strongest support that Darwin was right in his claims about universal emotion categories.

The "basic" emotion theories (BETs) thus put specific emphasis on the expression-component, which is thought to be closely linked with the physiological component. Following Darwin, BET postulates that emotions are caused by specific challenges and opportunities that have been especially important in our evolutionary past. The neural causes of emotions are thus thought to be hardwired responses to specific events and therefore produce stereotypical sets of output in the face, voice, autonomic and peripheral nervous system, and behaviour. The theory predicts that there is a mapping between the types of events that cause emotions in a specific "family" of emotions. Each emotion is thought to have a specific physiological and behavioural output that can be interpreted as signals by an observer, and these outputs can be "read" as emotions by others. BET suggests that a small set of emotions, typically five or six, are the building blocks that cannot be further reduced and that all the nuanced affective states that we experience as emotions are combinations of these building blocks. A strength of BETs is that they have strong predictions, the line between emotions should not be fuzzy, and that they separate emotions from other affective states such as moods, emotional traits and attitudes, and emotional disorders.

Another, closely related theoretical framework, is "appraisal" emotion theories (AET; e.g., (Arnold, 1960; Lazarus, 1968; Lazarus, Coyne, & Folkman, 1984; Roseman, 2013). AETs put more emphasis on the cognitive component and thus put less emphasis on the claim that specific types of challenges or opportunities automatically cause specific emotions. The cognitive appraisal of the situation may thus have a larger effect on the type of emotion produced by an event than the event itself. An event may lead to different emotions depending on how the person evaluates the importance and possible impact of the event. The same event may trigger different emotions in different people or even in the same person from one time to another. However, many responses to specific events are thought to initiate largely

automatic evaluations of the significance of a situation or event, which then drive the responses in the other components. Thus, once an emotion is triggered, the impact on physiology and expression is usually described in a similar way as BET. However, the direct association between emotion and expression postulated by BETs is downplayed in the definition of emotions and thus AETs leave room for a more nuanced view of the expressions associated with specific emotions. Rather than saying that the expression is a consequence of the emotion, they argue that the expression may or may not be a part of the emotion episode depending on the social circumstances and the expectations it puts on the individual subjected to the emotion (Ellsworth & Scherer, 2003; Moors, 2014; Moors, Ellsworth, Scherer, & Frijda, 2013; Shuman, Clark-Polner, Meuleman, Sander, & Scherer, 2017).

Putting more emphasis on the cognitive component in the generation of expressions leads to differing predictions about the continuity of cross-cultural emotion expression and recognition. Whereas BET postulates strongly that emotions are universally expressed and recognized, AET predicts more cross-cultural variation of expressions. Even so, AET usually suggest at least some cross-cultural continuity of emotion expression. Although AETs focus on the evaluation of events and are therefore more flexible in the type of emotions and expressions that are produced in a given situation, both AET and BET usually study expressions of discrete emotions such as happiness, anger, fear, sadness, and disgust. However, whereas BET studies expressions as a means to study emotions, AETs study expressions as a means to study appraisals which in turn elicits emotions. Because the appraisals may lead to more nuanced expressions that can be recognized by observers, AET usually accepts additional emotion labels to be dubbed "emotions" compared to BET, but both theories keep to only a few "families" of emotions.

There is something intuitively right about thinking of emotions as a few basic categories or families because it makes sense with the qualitative experience we have when we feel happy, angry or any of the other basic emotions. It is easy to argue that the experience of anger is qualitatively different from what we feel when we are afraid or happy. However, a third lineage of theoretical work, "constructed" emotion theories (CETs), suggest that the categories we interpret and feel as discrete emotions are constructed by culture and language rather than specific neural mechanisms. CET postulates that emotions are not caused by specific "brain modules" shaped by evolution to respond to important events, but rather that they are caused by more general functions of the brain (Barrett, 2006, 2017; Russell, 2003; Siegel et al., 2018). These functions are thought to be more general in the sense that they may elicit the subjective feeling of a discrete emotion, but also have other cognitive functions such as attribution and pattern recognition in general. In this way, the experience we have when feeling a discrete emotion is

based on an attribution of what we usually associate with that feeling and what we expect will happen next based on previous experience of similar situations. Even though emotions are not seen as specific brain modules, emotions can still be viewed as evolutionarily driven adaptations guiding humans and other animals to cope with their environment and predict the behaviours of others. But these behaviours are caused by the same processes in the brain as any other predictions about our surroundings and not specific for generating emotions.

CETs postulate that the attribution of emotions is driven by what is called "core affect", basic physiological states of valence and arousal that are then interpreted as discrete emotions by processes related to language and culture-specific expectations. Discrete emotions thus "become" emotions when they are categorised as such, and this categorization is learned from language and socialization. Because discrete emotions are seen as culturally dependent, CETs puts very little emphasis on the expression component in their definition of emotions. Rather, CETs usually study emotion expressions in terms of core affect, which will have a much less direct effect on the expression.

The main difference between theories that propose that emotion categories are innate versus constructed by culture is thus that the former argue that there are specific networks or modules in the brain that are activated in specific situations. The latter instead argues that there are more general/nonspecific networks that have evolved to cope with a myriad of situations and that these networks sometimes give rise to what is often interpreted as emotions. Both viewpoints thus have their roots in an evolutionary view, but they differ in their view of domain-specific or domain general mental processes that give rise to emotional episodes. A consequence of this difference is the differing interpretations of emotions and the physiological reaction that is coupled with that emotion as adaptations to specific events (freezing, fleeing in the sight of a predator) and the emotion it elicits (fear). The "innate" viewpoint says that an evaluation of threat produces fear whereas the "constructed" viewpoint says that to experience fear is to experience something as threatening. In the latter, the appraisal process is a general process that evaluates all kinds of situations. Thus the same system that creates perceptions and cognition sometimes also produces emotions. (Adolphs, 2016; Barrett, 2014; Moors, 2014, 2017; Tracy, 2014; Tracy & Randles, 2011).

In summary, BET, AET, and CET have different predictions regarding communication of discrete emotions. BET postulates that basic emotions have unique and distinct features signalling different emotions. Expressions are based on automatic physiological responses and evaluations of antecedent events. This proposal is based on the notion that emotions have evolved to deal with a specific task or situation, and that specific emotions will be

evoked by similar situations for all humans. In consequence, BET predicts that expressions of a few emotion categories are communicated and understood in a similar way across cultures and even species. AET has a more flexible view on how emotions are communicated because the evaluation of the situation can produce different emotions in similar situations. However, appraisals can be more or less automatic. For automatic appraisals, such as those evoked by loud noises, expressions are expected to initially function in a similar way as those described by BETs. Less automatic appraisals however, are more influenced by cultural and contextual demands as well as by individual differences and thus have more differentiated expressions. Even though AET proposes that expressions are influenced by physiological changes coupled with an emotion, these changes may or may not produce a stereotypical expression. The expresser is thus thought to have more voluntary control of the communication process and therefore AET predicts more variability across cultures. CET views discrete emotions as culturally learned concepts based on evaluations of internal states of valence and arousal (core affects). Emotional communication is thus based on what expression a person has learned to associate with a specific emotion concept. Therefore, CET predicts that expressions of discrete emotions are poorly recognized both within and across cultures in the absence of knowledge about emotion concepts.

Definitions of emotion terms used in this thesis

Across the four studies described in this thesis, discrete emotions have been defined mainly in terms of the AET framework. This means that the emotion terms presented below are both defined as dictionary definitions and as scenarios typically associated with each emotion. The scenarios describing the typical situations in which each emotion may be elicited were based on current AET (e.g. Ellsworth & Scherer, 2003; Lazarus, 1991; Ortony, Clore, & Collins, 1988). Some emotion terms are commonly agreed to represent basic emotions (anger, disgust, fear, happiness, sadness, and positive/negative surprise; e.g., Ekman, 1992), and others considered as basic emotions by some but not all BETs (contempt, interest, lust, and relief; see Tracy & Randles, 2011). Three emotions that were included were intended to represent a self-conscious emotion family (pride, shame, and guilt; see Tangney & Tracy, 2012) and the other emotion terms were intended to capture affective states that are rarely studied (and may not be considered as "true emotions" by BETs) but that may have distinct expressions in the voice (affection/tenderness, peacefulness/serenity, amusement, and distress; e.g. Goudbeek & Scherer, 2010)

The definitions and scenarios were also used to describe the emotion terms to actors and/or listeners in studies 1, 2, 3, and the yet unpublished study presented in Chapter 3.

Amusement (study 2)

A moderately arousing state of finding something funny or comical.

Scenario: "Think of a situation where you experienced something that you found very amusing. For example, you saw or heard something that amused you. The situation is pleasant."

Anger (all studies)

A highly arousing state of displeasure caused by real or imagined injury to oneself or someone valued. Anger is usually accompanied by the desire and possibility to retaliate, either immediately or manifested in plans to retaliate later.

Scenario: "Think of a situation where you experienced a demeaning offense against you and yours. For example, somebody behaves rudely toward you and hinders you from achieving a valued goal. The situation is unexpected and unpleasant, but you have the power to retaliate."

Contempt (unpublished study and study 2)

A weakly arousing state of displeasure directed toward someone that is seen as immoral, dishonest, corrupt or inferior.

Scenario: "Think of a situation where you disagree with the actions of a person that you regard as inferior to you. For example, you feel that the actions of a person are against your wishes. As a consequence you find the person repulsive and feel superior to him/her."

Disgust (unpublished study and study 2)

A moderately arousing state of repulsion caused by something considered offensive or infectious, such as a foul smell, rotten food, or contagious disease.

Scenario: "Think of a situation where you were taking in or being too close to an indigestible object. For example, somebody offers you food that smells rotten and that repulses you. The situation is very unpleasant."

Distress (study 2)

A highly arousing state of pain and powerlessness caused by a direct harm to oneself.

Scenario: "Think of a situation where you have been harmed and want to get out of the situation immediately. For example, somebody hits you and you are in pain and want to escape. The situation is very unpleasant, and you do not have any control over what is happening."

Fear (all studies)

A moderately to highly arousing state of agitation and anxiety caused by the presence or imminence threat of danger.

Scenario: "Think of a situation where you faced an immediate, concrete and overwhelming physical danger. For example, something or somebody threatens to harm you and yours. The situation is unexpected and unpleasant, and you are uncertain about your ability to cope."

Guilt (study 2)

A weakly arousing state of anxiety and remorse caused by a realization of some shortcoming or transgression (or belief thereof, accurate or not). The transgression might be a behavior that compromise ones own standards and values, are considered immoral, cause harm or violate an agreement or the rights of others.

Scenario: "Think of a situation where you transgressed a moral imperative. For example, you did something that you knew that you were not supposed to do, and this action caused somebody harm. Though you did achieve your goal, you now feel bad about this."

Happiness/Joy/Elation (all studies)

A moderately arousing state of pleasurable content of mind, which results from success or the attainment of what is considered good and sought after. Often associated with the realization or definite progress towards realization of a valued goal.

Scenario: "Think of a situation where you made reasonable progress toward the realization of a goal. For example, you have succeeded in achieving a valued goal. Your success may be due to your own actions, or somebody else's, but the situation is pleasant and you feel active and in control."

Interest (unpublished study, studies 1, 2)

A moderately arousing state of curiosity, concern, and focused attention toward an event, process or object that the individual wants to learn more about.

Scenario: "Think of a situation where you encounter something that you want to learn more about. For example, you encounter something new that you feel could help you achieve a valued goal. The situation is pleasant and you feel that it is possible for you to learn more."

Negative/positive surprise (study 2)

A moderately arousing state of encountering or discovering something unexpected. If the surprise is negative or positive is determined by if the event was something that the person wanted to happen or not (e.g. helped the person to reach a valued goal or not)

Negative scenario: “Think of a situation where you experienced something that you were not expecting and did not wish for to happen. For example, something or someone unexpectedly hinders you from achieving a valued goal. The situation is very unexpected and you need time to take it in.”

Positive scenario: “Think of a situation where you experienced something that you were not expecting, but that you did wish for to happen. For example, something or someone unexpectedly helps you to achieve a valued goal. The situation is very unexpected and you need time to take it in.”

Pride (unpublished study, studies 2, 3)

A moderately arousing state of a pleasant or sometimes exhilarating self-evaluation, often caused by the achievement of a valued goal.

Scenario: “Think of a situation where you experienced the enhancement of positive feelings about yourself by taking credit for a valued object or achievement. For example, you (or someone you identify with) did achieve a valued goal. The situation is pleasant, and you deservedly receive the credit for the positive outcome.”

Relief (unpublished study, studies 1, 2)

A low-arousing state of relaxation caused by the easing of a burden or distress, such as pain, anxiety, or oppression.

Scenario: “Think of a situation where you experienced that a distressing condition changed for the better or went away. For example, you can finally relax after an ordeal of some sort. The situation is pleasant, and you feel certain about the positive outcome.”

Sadness/Grief (all studies)

A low-arousing state of disappointment, sorrow, despair and helplessness caused by disadvantage or irrevocable loss of something valued.

Scenario: “Think of a situation where you experienced an irrevocable loss. For example, you lose someone or something very valuable to you, and you have no way of getting back that what you want.”

Serenity/peacefulness (all studies)

A low-arousing state of peacefulness, complete fulfillment of a want and freedom from disturbance or agitation.

Scenario: “Think of a situation where you experienced the complete fulfillment of a want. For example, you have achieved a valued goal, and now no longer have to put in any effort. The situation is pleasant, and you feel calm and secure.”

Sexual lust (unpublished study, studies 1, 2)

A highly arousing state of overwhelming desire or craving of a sexual relationship with a person whom you feel affection toward.

Scenario: "Think of a situation where you desired or participated in a sexual relationship. For example, you desire to have a sexual relationship with a person whom you feel affection toward. The situation is pleasant, and you are aroused."

Shame (unpublished study, studies 2, 3)

A weakly arousing state of anxiety and pain caused by the consciousness of something dishonoring in one's own conduct or failure to live up to ideals.

Scenario: "Think of a situation where you failed to live up to your ideals. For example, you failed to perform according to your standards on an important task, and feel bad about not living up to your own ideals. The situation is unpleasant, and it was your own fault."

Tenderness/affection (unpublished study, studies 1, 2)

A low-arousing state of trust, compassion, appreciation, and understanding directed toward a friend, spouse, child or pet.

Scenario: "Think of a situation where you desired or participated in affection. For example, you interact with a very close friend (with whom you have a non-romantic relationship) who you really appreciate and are drawn to. The situation is very pleasant and you also feel appreciated in return."

Neutral scenario: "Think of an ordinary everyday situation where you did not feel any particular affective state. For example, you are engaged in doing the activities that you most commonly do on an everyday basis. The situation is neither positive nor negative and you do not wish to express anything besides the verbal content of the utterance."

3. How are emotions communicated in speech?

Reliable communication of emotion is beneficial for survival in both humans and other animals because it gives individuals valuable information about their surroundings. This information may be used to adjust the individual's behaviour in ways to avoid danger and to promote goals (Wheeler & Fischer, 2012). For animals living in social groups, emotional communication is especially important because it allows individuals of the group to sustain relationships, resolve conflicts, and coordinate other behaviours related to reproduction, foraging, and defence against predators (Juslin & Scherer, 2005; Wheeler & Fischer, 2012). With language and speech, humans have developed a communication system transferring such information with exceeding precision and complexity. A large portion of the information contained in speech however, is still based on the same principles as the phylogenetically older ways of communication (Banse & Scherer, 1996; Owren et al., 2011). As opposed to the verbal content (information transmitted with words), the nonverbal content of speech is referred to as prosody. Emotional prosody refers to the nonverbal information speakers use to communicate emotion (intentionally or unintentionally) by manipulating certain aspects of the voice. To a large extent, this type of communication follows the same principles as those of non-linguistic expressions (i.e. the emotional sounds humans make when we are not speaking) (Sauter, Eisner, Calder, & Scott, 2010), which in turn follow the same principles as those used by other animals (Filippi et al., 2017; Owren et al., 2011). The prosody of a speaker thus conveys other information than the words, much of which has the function of communicating emotion.

An assumption that has to be justified before investigating *how* emotions are communicated in speech is to establish that there *is* nonverbal communication of emotions at all, that is, communication independent of the verbal content. Accordingly, this was the focus of the early studies on emotional communication (Banse & Scherer, 1996; Juslin & Laukka, 2003). Because there is no such thing as speech without words, a first obstacle in the study of emotional communication is to separate the meaning of the words, the verbal content, from the meaning of the prosody, the nonverbal content. Most studies have solved this issue by using some version of the standard content paradigm. In this paradigm, a speaker (i.e. an actor) is asked to read

a standard sentence while portraying specific emotions (e.g., anger, happiness, fear). The standard sentence may consist of a series of numbers, letters of the alphabet, sentences with neutral words, nonsense words, or words in a foreign language (Juslin & Laukka, 2001). The rationale of the standard sentence paradigm is thus to keep the information in the words constant while letting the nonverbal information vary. If listeners are able to use the nonverbal information to infer the intended emotion, emotional communication has occurred.

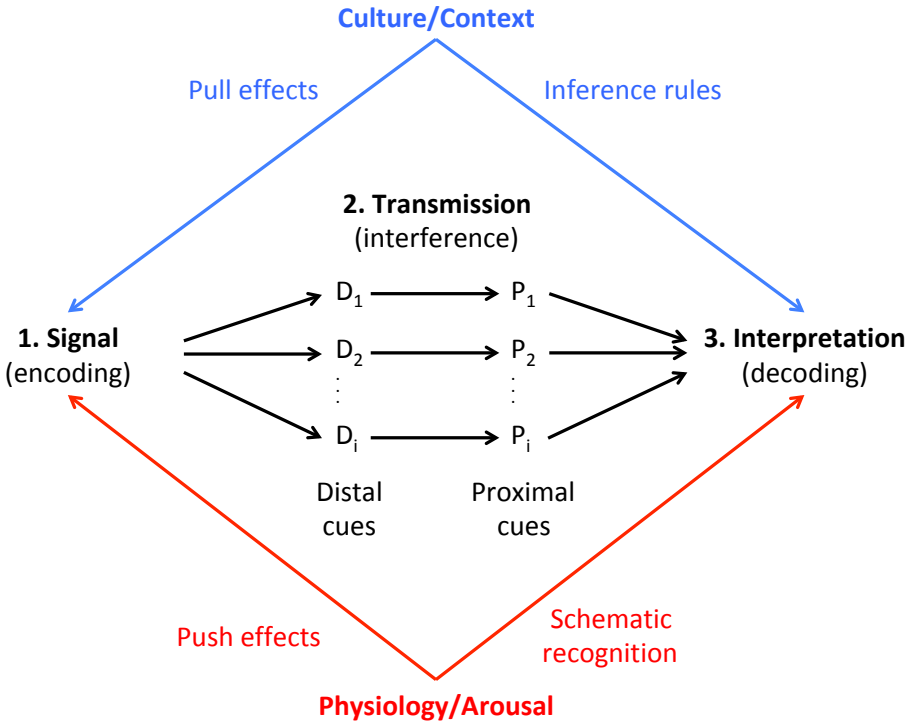
Almost all of the early studies on emotion recognition used a procedure called the forced-choice task in which listeners are presented with recorded emotion expressions and then asked to classify them by selecting an emotion word from a list of alternatives. Although the forced-choice task has been heavily criticized (e.g. Russell, 1994) because (1) it may inflate recognition rates (listeners guessing the correct answer), (2) recognition rate is affected by which emotion-alternatives are included (some emotions are more easily confused) and (3) because of its low ecological validity (the procedure is not even close to how we evaluate emotions in real life), most researchers agree that the results of these studies (and hundreds of replications, e.g. Elfenbein & Ambady, 2002; Juslin & Laukka, 2003) show that emotions can be communicated via nonverbal information in speech. Though criticized, one benefit of the forced-choice task is that listeners immediately understand what they are supposed to do. The fact that listeners intuitively understand the task suggests that it, at least in some way, reflects the mental abilities people would use in real-life situations. Also, it is a relatively easy way for researchers to get estimates of how well different emotions are communicated, and confused, because the percentage of listeners selecting the intended emotion, and each of the non-intended emotions, can easily be computed. In this field of research, recognition of an *expression* is thus defined as some proportion of listeners who selected the intended emotion label, and recognition of an *emotion category* is defined as the mean across all expressions belonging to that category.

Having established that there *is* nonverbal information that listeners can recognize reliably as discrete emotions, researchers went on to investigate how speakers manipulated their voices to express emotions. The early focus on recognition rather than expression was not only driven by the necessity to justify that communication occurred but also because the acoustical analysis required was inaccessible to most researchers in the field (there are exceptions, see e.g., Skinner, 1935). Though there has been a rapid increase in research using acoustic analyses to study emotional expressions in recent years, there are still considerably fewer studies of this kind.

A framework to describe how humans communicate emotion

When researchers became more interested in studying emotion expressions, it was evident that expression and recognition was part of the same communication process and should be studied together. Emotional communication in speech can, in fact, be studied much the same way as any other type of communication; as transformation of information from one person to another. The dynamic tripartite emotion expression and perception (TEEP) framework has been proposed as a way to illustrate how emotions are communicated via vocal, facial and bodily cues (Scherer, 2018). It describes the communication process as "dynamic" because it assumes that the emotional state of the speaker, and therefore also the message, changes continuously, and "tripartite" in the sense that it describes the communication process in three steps (see Figure 1).

Figure 1. The dynamic tripartite emotion expression and perception (TEEP) framework



First, there has to be some kind of signal or set of "cues" that the speaker expresses. The process used to express a set of cues to signal an emotion is sometimes referred to as "encoding". In this case the cues are changes of the character of the voice related to the emotional state of the speaker. The changes of the character of the voice are thought to be influenced by so called "push" and "pull" effects intended to highlight that manipulations are caused both by the physiology of the speaker as well as contextual demands. The push effects are the involuntary manipulations of the voice caused by the arousing emotional episode. The level of arousal causes the muscles involved in voice production to ease or tense, which in turn will have an effect on the character of the voice. On the other hand, the pull effects are the more or less intended manipulations of the voice that the speaker uses to adjust their voice to fit the social context of the communication. These adjustments comprise cultural and contextual agreements that could either enhance or conceal the push effects or even introduce other manipulations of the voice that are commonly agreed to signal an emotion in that specific context. The pull effects are thus intended to influence the interpretation of the expression in a specific way. In other words, the pull effects are the vocal manipulations reflecting what the speaker intends to communicate, consciously or unconsciously to the listener, while the push effects are caused by involuntary physiological changes. Appreciating that the emotional state of a speaker is assumed to change continuously further increases the complexity of the information that is transmitted. The information contained in the voice across an utterance may thus be an ambiguous mix of several emotions with varying levels of arousal. And even if the emotional state of the speaker was fairly stable across the utterance, the information may still be ambiguous because of the interaction of push and pull effects. For example, a speaker may have tried to control the voice to fit the contextual demands, succeeding at first but failing towards the end of the utterance.

The second step of the communication process illustrates the transmission of information from speaker to listener. The many changes of the voice related to the expression are now split up into what is called "distal cues" in the framework, presumably because it takes the perspective of the listener who will perceive them as "proximal cues". This division is intended to highlight the fact that interference is possible during transmission of the information. For example, even though subtle changes in the voice could have been used as information to infer the emotion, the physiology and function of the auditory and nervous systems may limit the listener's ability to actually perceive them. Interference could of course also arise from external factors such as the distance to the speaker and/or due to background noise. The information may also be distorted if the medium of transmission filters the acoustic characteristics of the signal in some other way. For example, the signal is filtered if a conversation is heard through a wall (low pass filter) or

if it is played back from a recording (filtered by the characteristics of the microphone, recording device, and speakers). Transmission is said to be successful if the proximal cues contain the same information as the distal cues.

In the third step of the communication process, the listener is appealed to process the available information (i.e. interpret the proximal cues) to come to a decision of what the emotional message might be. The process used to interpret the cues that signal an emotion is sometimes referred to as "decoding". Because of the complexity and potentially ambiguous nature of the information contained in the voice, the interpretation is said to be "probabilistic", highlighting three aspects of the content of the emotional message. First, the information is potentially ambiguous because of the ever-changing emotional state of the speaker. Second, the influence of the push and pull effects on the voice may be contradictory because the speaker may try to conceal the push effects, or have a cultural expression-schema that the listener is unfamiliar with. Third, specific manipulation of a cue may be used to signal more than one emotional state. Thus, there is no one-to-one mapping between a single cue and a specific emotion that the listener may use in isolation to come to a conclusion. Rather, the listener has to rely on a pattern of many cues, with subtle differences between qualitatively similar emotions, which together may or may not guide the listener in the intended direction. When making the probabilistic inference, the listener also has to integrate the information they receive with previous knowledge of how an emotional state usually influences the voice of a speaker and to figure out what contextual demands the speaker may be influenced by. In this framework, these two cognitive processes are called "schematic recognition" referring to knowledge about physiological effects commonly associated with an emotion, and "inference rules" referring to knowledge about contextual effects. If the listener knows the speaker, the knowledge about how that speaker's voice usually sounds and changes under emotional influence is also an important part of the contextual effects.

Once the listener has come to a decision about the internal emotional state of the speaker this will influence the listener's behaviour. Whether it is to comfort the speaker, say something nice, run away, or to select an emotion label in a list of more or less relevant words. If the information in the proximal cues could be used to guide the listener to behave in a manner expected by the speaker (or researcher), emotional communication has occurred.

At a first glance, this framework may seem rather complex, but this is mainly caused by the type of information that is being transmitted; the subtle changes in the tone of a voice. A way to demystify the framework is to apply it to a more direct type of communication, for example verbal communication. In the first step, a speaker might produce a series of words with potentially ambiguous meaning or even different meanings from one time to an-

other. In the second step, the information is transmitted more or less intact. In the third step, the listener searches their memory and vocabulary in an attempt to decode the meaning of the words and if the interpretation is close enough to the intended meaning, verbal communication occurred.

The purpose of the framework is thus to highlight that any functional communication of information relies on unique patterns that will signal meaning from one agent to another. It highlights the need of agreement between the expresser and the perceiver and that successful communication of emotions relies on emotion-specific patterns of vocal-acoustic cues that are sufficiently specific for, and reliably related to, an emotion category. Thus, if each emotion has a unique signalling expression they can be understood, categorized and differentiated from other emotions in a similar way as words are articulated by a speaker and understood by a listener (Scherer et al., 2011). It also highlights the need to study the complete communication process, in other words that expression and perception cannot be studied separately.

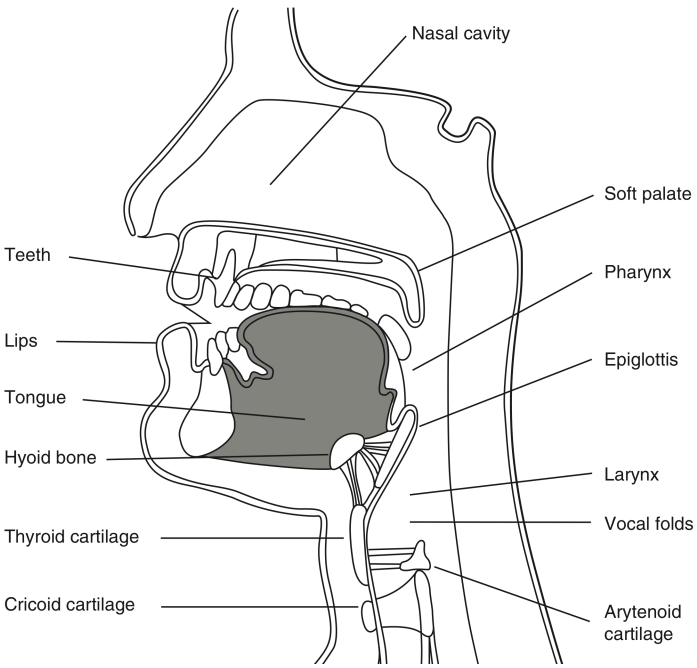
Physiology of the vocal apparatus

The vocal apparatus is illustrated in Figure 2. Humans produce speech and nonverbal expressions by pressing air from the lungs through the vocal tract. The vocal tract consists of the oral cavity, the nasal cavity, the pharynx, and the larynx. Each of these four components consists of several subcomponents. The larynx is what produces the sounds and the other parts are cavities that amplify or resonate the sounds produced and thus have the ability to change the characteristics of the sounds by amplifying some frequencies while attenuating others. The oral cavity is the mouth, lips, teeth, and cheeks and the nasal cavity is the hollow space behind the nose. The pharynx is the upper part of the throat that connects the nasal and oral cavities to the larynx (Kreiman & Sidtis, 2011).

The larynx, commonly called the voice box, consists of a set of muscles, ligaments, and cartilage that can move up and down to prevent food from entering the lungs and to alter the length and position of the vocal folds. When we speak or make other vocalizations, the vocal folds are pulled together forming a thin space between them called the glottis. When air is expelled from the lungs, the air pressure on the glottis increases until it opens to release the pressure and then closes again rapidly until the pressure forces it to open again. This repeated opening and closing of the glottis causes the vocal folds to vibrate and thus produce sound. The frequency of the vibration determines the pitch, and the air pressure determines the loudness of the sounds produced.

Because the emotional episodes together with the speaker's more or less intentional manipulations produce changes in the physiology of the vocal apparatus, including the muscles controlling respiration, phonation, and articulation, the perceived loudness, pitch, sharpness, and tempo of the utterance is affected. The listener may then interpret these effects as an emotional message. The changes in the voice can also be recorded and measured in order to find the patterns of acoustic parameters that speakers and listeners use to communicate emotions.

Figure 2. Physiology of the vocal apparatus



Acoustic parameters

Over decades of research, a large number of acoustic parameters have been suggested as physical measures of the auditory information that listeners perceive as emotion in speech. These acoustic features can be sorted into four major categories related to what listeners would perceive as pitch (frequency parameters), loudness (energy/amplitude parameters), voice quality or timbre (spectral balance parameters), and speech rate (temporal parameters). Over the years, the computation and vocabulary used to compute and describe these acoustic parameters has not been standardized making it diffi-

cult to compare results reported in different studies. In an attempt to solve this issue, a recent initiative by leading researchers in the field suggested a "Minimalistic" and "Extended" acoustic parameter set (GeMAPS and eGeMAPS, Eyben et al., 2016). Researchers investigating the acoustic correlates of vocal expressions are encouraged to use these parameter sets alongside any other relevant acoustic parameters that may be important for the specific study. Throughout this thesis, I have used the extended version of GeMAPS containing 88 parameters and then used Principal Component Analysis (PCA) and/or other feature selection procedures to reduce and select parameters in each study. Below follows a short description of the parameters that were used in this thesis.

Frequency related parameters:

The **fundamental frequency** (F0) is determined by the rate of vibration of the vocal folds and is the main acoustic parameter to measure what listeners would perceive as the pitch of a voice. In GeMAPS, F0 is expressed both in hertz (cycles per second) and as semitones on a frequency scale starting at 27.5 Hz (semitone 0). F0 is computed via sub-harmonic summation in the spectral domain for overlapping time windows that are 60 ms long and 10 ms apart for voiced regions of the recording. From this series of time-windows, or "low-level descriptors", the average, standard deviation, percentiles, range, and contour (falling and rising slope) can be computed to describe different aspects of F0 across time. Such "functional" parameters can be computed for all parameters that are based on a series of time-windows.

Jitter is a perturbation measure computed as the average window-to-window deviations in consecutive F0 time windows. Presence of jitter is associated with perceived "roughness" of the voice (Barsties v. Latoszek, Maryn, Gerrits, & De Bodt, 2018).

The **frequency of the first, second and third formants** are perceived as vowel quality (or modes of articulation of vowels). In the acoustic signal, formants are frequency bands with stronger amplitude that can be visually inspected in a spectrogram such as the one on the cover of this thesis. The formants reflect the natural resonance and speaker modulations of the vocal tract. The third and higher formants may be more influenced by individual physiological differences in the vocal tract rather than intended modulations of the vowel sounds (Juslin & Laukka, 2003; Laver, 1980). In GeMAPS, the formants are computed (over 20 ms time windows, 10 ms apart) as the centre of frequency for each formant using the roots of Linear Predictor-coefficient polynomial.

The **bandwidth of the first, second, and third formants** are also perceived as vowel quality. The computation is based on the same principle as

for the formant frequencies but instead of computing the centre of frequency for each formant the width of the formant is computed.

Amplitude related parameters:

To model humans' non-linear perception of the **amplitude or energy** of the acoustic signal, GeMAPS filters the signal to produce what is referred to as an "auditory spectrum" that mimics human amplitude-perception. The main measure of energy in the voice is therefore called loudness, referring to the perceptual quality of the measure. When the auditory spectrum has been obtained, loudness is computed as the sum over all frequency bands of the amplitude in the spectrum. This parameter reflects the effort the speaker used to produce the utterance and estimates how loud listeners would perceive it. Similar to F0, functional parameters for the average, standard deviation, percentiles, range, and contour (falling and rising slope) are computed. GeMAPS also computes an amplitude measure that is not based on the auditory spectrum. This measure is called **equivalent sound level** and expresses the amplitude on a logarithmic (decibel) scale by computing the average of the logarithmized root mean square of each frame.

Shimmer is, like jitter, a perturbation measure that is associated with perceived "roughness". Unlike jitter though, which measures pitch-instability, shimmer measures amplitude-instability. It is computed in a similar way as jitter but uses the average window-to-window deviations in *peak amplitudes* of consecutive F0 time windows.

Harmonics-to-noise ratio (HNR) is a measure of the proportion between harmonic and nose components in the voice, and is related to excessive airflow, or turbulence, through the glottis during vowel articulation. HNR is computed as the ratio of the F0-amplitude (amplitude of the main harmonic) over the summed amplitude across the other frequency bands. HNR has been associated with perception of voice "breathiness" (Barsties v. Latoszek, et al., 2018).

Spectral balance parameters:

Alpha Ratio, Hammarberg Index, Spectral slope, energy proportion, and Harmonic difference are all measures of what is perceived as "soft"- or "sharpness" of a voice. These measures are sometimes referred to as measures of "high frequency energy" because they generally measure the proportion of energy below versus above a certain cut-off in the frequency spectra. In GeMAPS, **alpha ratio** is computed as the ratio between the summed amplitude in the 50-1000 Hz and 1-5 kHz frequency bands. **Hammarberg index** is computed as the ratio of the strongest peak amplitude in the 0-2 kHz and the 2–5 kHz frequency bands. The two measures of **spectral slope** are computed as the linear regression slope of the amplitudes of two frequency bands; one for the amplitude in frequency bands centered on

0 versus 500 Hz and the other for frequency bands centered on 500 versus 1500 Hz. The two measures of **energy proportion** are computed as the ratio of energy below and above 500 Hz, and 1000 Hz respectively. The two measures of **harmonic difference** are computed as the ratio of energy in the first harmonic of F0 (H1) to the energy in the second harmonic of F0 (H2), and third harmonics of F1 (A3) respectively.

The **relative energy in formant one, two, and three** are, like formant frequency and bandwidth, also related to the perception of vowel quality but these parameters instead measure the amplitude in the formats relative to the amplitude of F0. The three measures are computed as the energy of the peak spectral harmonic at the first, second, and third formant's centre of frequency divided by the energy of the spectral peak at F0.

Spectral flux and the Mel Frequency Cepstral Coefficients (MFCCs) are more difficult to tie to a specific perceptual quality. Usually it is simply said that these measures are related to what is perceived as the "timbre" of a voice, which is just another way of saying that it is not perceived as either pitch or loudness. **Spectral flux** is an average measure of how quickly the energy distribution across frequencies is changing over time. It is computed in two steps. First, the distribution of energy across a set of frequency bands is computed (a power spectrum), and then the squared differences between consecutive time windows in each frequency band is computed and summed. The **MFCCs (1 – 4)** are four measures that are thought to represent the human perception of spectral balance. They are based on the auditory spectrum described above and the mel frequency scale. The mel frequency scale is a scale of frequencies that listeners judge to be equal in distances from one another and thus models humans' non-linear perception of frequency in a similar way as the auditory spectrum models energy perception. The MFCCs are computed by first computing the auditory spectrum across 25 ms windows and then apply a mel-filter to segment the frequency bands according to the mel scale (narrower bands for lower frequencies and wider bands for higher frequencies). The cepstral coefficients 1 – 4 are then obtained via the discrete cosine transform of the (logarithmized) energy in each frequency band of the auditory spectrum.

Temporal related parameters:

The four measures in this category are related to the perceived speed or tempo, and to the rhythm, timing, and pausing of the speech. **Rate of loudness peaks** and the **number of continuous voiced regions per second** are both influenced by how fast the person speaks and can thus be seen as measures of syllable rate per second (though they do not actually measure the number of syllables). Respectively, they are computed by counting the number of loudness peaks in the auditory spectrum divided by the length of the recording, and counting the number of regions in which speech was detected divid-

ed by the length of the recording. Because the sentences uttered in the unpublished study were all 20 syllables long, I added an extra parameter measuring the actual number of syllables per second by dividing 20 with the length of each recording.

The **Length of continuously voiced regions** and the **Length of continuously unvoiced regions** are intended to measure what a listener would perceive as the rhythm or fluency of the speech. They are computed as the mean and standard deviation of the length of the voiced and unvoiced regions of the recording.

Unpublished study: Acoustic Correlates of Emotional Communication in Speech

The main purpose of this study was to find acoustic parameter-patterns suggested by previous research describing how specific emotions are communicated in speech, and to compare these patterns with those used by the listeners in the current study. Because the literature is vast and partly contradictory, I have focused on the influential framework presented by Scherer and colleagues (Banse & Scherer, 1996; Scherer, 2018), but I have also tried to incorporate findings of other studies more or less related to their work. The aim was thus to evaluate the predictability of the currently most influential ideas of how discrete emotions are communicated in speech.

Acoustic analysis of vocal emotion expressions suggests that discrete emotions are expressed with a combination of absolute and relative acoustic-perceptual cues that together form a pattern unique for each emotion (Banse & Scherer, 1996; Hammerschmidt & Jürgens, 2007; Juslin & Laukka, 2001; Sauter, Eisner, Ekman, & Scott, 2010; Scherer et al., 2011). When asked to report what type of nonverbal information they use to infer the emotional state of a speaker in experiments or in their everyday life, people most often report cues related to the intensity of the voice and the speech rate (Planalp, 1996). Accordingly, acoustic features related to loudness and the number of syllables per second are, together with the mean and variation of F0, commonly suggested to be the most important carriers of the emotional information in nonverbal communication (Johnstone & Scherer, 2000; Juslin & Laukka, 2003).

The study by Banse and Scherer (1996) describing the patterns of acoustic features used in emotion communication of discrete emotions, has been immensely influential to the development of the field over the last two decades. With impressive thoroughness, especially with regard to the technical equipment available at the time, they examined how expression and perception of 14 emotion categories were related to 29 acoustic parameters. Their results showed that expressed emotion predicted a large proportion of the variance in several of the acoustic parameters, especially for mean F0 and

loudness. They also showed that a combination of seven acoustic parameters could account for a large portion of the variation in listeners' ratings for different emotions. The seven functional parameters suggested to represent the information listeners use to classify emotions in speech were: mean F0, standard deviation of F0, mean loudness, mean duration of voiced periods, hammarberg index, proportion of energy up to 1000 Hz, and "spectral drop-off" (comparable to spectral slope described above). For the emotions relevant in the context of the current study the following acoustic features (as indicated by significant multiple correlation coefficients) suggested that expressions rated by listeners as:

- Intense (hot) anger was associated with increased F0 mean and F0 sd, and decreased hammarberg index.
- Moderate (cold) anger was not significantly related to any parameter.
- Intense (panic) fear was related to increased F0 mean.
- Moderate fear (anxiety) was related to increased F0 mean, and decreased F0 sd and loudness mean.
- Intense sadness (despair) was related to increased F0 mean, duration of voiced periods and hammarberg index, and decreased F0 sd.
- Moderate sadness was related to decreased loudness.
- Intense happiness (elation) was related to increased F0 mean and proportion of voiced energy up to 1000 Hz.
- Moderate happiness was not significantly related to any parameter.
- Interest was not significantly related to any parameter.
- Shame was related to decreased loudness mean.
- Pride was related to decreased F0 mean.
- Disgust was related to decreased proportion of voiced energy up to 1000 Hz.
- Contempt was related to decreased F0 mean and increased F0 sd.

Over the years, these findings have been extended and slightly modified and incorporated in a more theoretical framework, taking into account how speech-production is assumed to be related to specific emotions and acoustic parameters. However, they are strikingly similar to the latest version (Scherer, 2018) and to other studies related to the same theoretical framework.

The predictions used in the current study were based on both theoretical suggestions and empirical findings. First, theoretical predictions of 17 acoustic parameters were derived from an interpretation of the acoustic parameter patterns presented in Scherer (2018, Tables 4.2, 4.3, which in turn are based on Scherer & Juslin 2005, table 3.2; Juslin & Laukka, 2003, Table 7; Scherer, Johnstone, & Klasmeyer, 2003, table 23.2; and Scherer, 1986, p. 161-162; and empirical findings in Bänziger, Mortillaro, & Scherer, 2012;

and Laukka et al., 2016). The theoretical predictions for each emotion-parameter combination are visualized in Figure 3 on pages 41-53 as grey areas above 1 sd (predicted as "high" compared with a speakers normal tone of voice), +/-1 sd (predicted "medium") and below -1 sd (predicted "low"). Second, because the theoretical predictions are rather vague, the figures also show the results presented in Banse and Scherer (1996, Table 6), Juslin and Laukka, (2001, Figures 3-6), and Goudbeek and Scherer (2010, Figures 1-2) as colored areas. These results are intended as empirical predictions with a more fine-grained resolution than the theoretical predictions because they also show approximations of how much each parameter is expected to deviate from the speakers' normal tone of voice. It should be noted that the predictions based on Goudbeek and Scherer (2010) and Juslin and Laukka (2001) concern expression rather than what was perceived by listeners. However, the selection of expressions in these studies were based on the fact that expressions were well-recognized in validation studies and thus imply that these acoustic patterns are those that allow listeners to recognize the expressions. It should also be noted that the three studies have used other acoustical measures than those used in the current study. Thus, the predictions should be seen as guidelines to be compared, rather than tested, with the acoustic feature patterns presented in the current study.

When trying to describe the acoustic parameter-patterns that listeners use to recognize emotions, a potential problem that arises from the forced choice task is that many judgments may be based on guessing or that the listener perceived an emotion not listed in the alternatives. To minimize the influence of such judgments, studies usually select expressions that are well recognized in validation studies. However, well recognized, as discussed above, is defined as a certain proportion of the listeners classifying them correctly. This means that listeners may still be uncertain about the intended emotion, regardless of whether they selected the intended emotion label or not. This may render the acoustic results misleading, or at best diluted, because the parameter-patterns produced will be based on a mix of (1) expressions that listeners actually perceived as the selected emotion, (2) expressions not perceived as a specific emotion (guessing), and (3) expressions perceived as some other emotion (not listed in the alternatives). In the current study, I took another approach to this problem. The expressions were not validated in terms of recognition of intended emotion, instead, participants were asked to rate their confidence on each judgment and only confident judgments were kept in the statistical analysis.

Methods

Recording of vocal expressions

Fourteen actors (7 females) were recruited via advertisement on a web page for actors looking for parts in Swedish TV/film productions and commercials. Their acting experience ranged from amateurs with no training to professionals with years of training and experience.

The actors were instructed to vocally express two sentences to convey 13 emotions. Both sentences were semantically neutral and were 20 syllables long, one Swedish ("En gång tvistade nordanvinden och solen om vem av dem som var starkast") and the other a nonsense sentence resembling Swedish ("Enocken lär sjölva, så marginen har ett visserlag mot såteng ferup"). The 13 expressions were selected to achieve a wide variety of emotions, some largely agreed to represent basic emotions (anger, disgust, fear, happiness, and sadness; e.g., Ekman, 1992), and some that are considered as basic by several but not all emotion theorists (contempt, interest, lust, relief, and tenderness; see Tracy & Randles, 2011). Two expressions representing self-conscious emotions (pride, shame; see Tangney & Tracy, 2012) and the low-arousal positive state serenity were also included. The actors were also instructed to read the sentences in a neutral prosody.

To avoid confusion with regard to individual differences of how emotion words are interpreted, the actors were given definitions and descriptions of typical situations that would elicit each emotion based on appraisal theory (Ellsworth & Scherer, 2003; Lazarus, 1991; Ortony et al., 1988; see Chapter 2). The actors were encouraged to use the definitions and descriptions to remind them of a similar self-experienced situation and to re-enact the situation in an attempt to elicit the emotional state. They were also instructed to express all prosodies (except neutral) with more and less intensity. For low arousal emotions such as serenity for which more and less intensity were not intuitively applicable, the actors were encouraged to express more or less of the emotion (for example moderately versus completely serene).

All actors were recorded in a low reflection soundproof room with a Brüel & Kjær type 4190 mono microphone (NEXUS Brüel & Kjær type 2690 A 0S4 amplifier, RME Babyface Pro soundcard) onto a computer with a sampling frequency of 48 kHz (16 bits per sample). The distance between the microphone and the actor's mouth was held approximately constant at 15 cm with the aid of a pop filter. The actors also recorded other materials that were not used in this thesis.

This procedure resulted in 756 recordings; 14 actors, 14 prosodies, 2 intensity levels (except neutral), and 2 sentences. The duration of the recordings ranged from 3.1 to 11.4 seconds with a mean of 4.7 s (SD = 1.0).

Listening experiment

One hundred and two participants (82 women), mean age 27.0 years (SD = 8.4, range 18-56), were recruited via advertisements on Stockholm University bulletin boards and a recruitment web page. Participation was based on informed consent and was rewarded with course credits or movie vouchers. No participants reported any hearing difficulties.

Participants were instructed to categorize each recording in a 14-alternative forced choice task by selecting the label they thought the actor in the recording was trying to express (anger, contempt, disgust, fear, happiness, interest, lust, neutral, pride, relief, sadness, serenity, shame, tenderness). They also rated how confident they were that they had selected the intended emotion label, and the intensity and the validity of the expression. To avoid confusion with regard to the interpretation of emotion labels, participants were given a sheet of paper with the same definitions and scenarios that the actors were given during the recording. Before the experiment began, participants were instructed to read the definitions and were encouraged to use them during the experiment if needed. The experiment began with four practice trials that were used as examples to aid the experimenter to describe the procedure. Participants were tested individually in a sound attenuated room and recordings were presented over headphones (Sennheiser HD 280 pro) at a comfortable listening level that could be adjusted by the participant during the practice trials. If they wished, participants could press a button to hear each recording again to reach a decision but were instructed not to dwell too long on each recording. Recording presentation order and the order of response options were randomized for each participant. PsychoPy software (Peirce, 2007) was used to present recordings and to collect responses. The experiment lasted approximately one hour including the instructions and the participants continued to judge recordings until the hour ended. On average, each participant judged 124.5 (SD = 45.3) recordings resulting in a mean number of 16.8 (SD = 45.3, range 13-25) judgments for each recording. In total, 12697 judgments were collected.

Acoustical and statistical analysis

All the 756 recorded emotion portrayals were acoustically analysed using the openSMILE software (Eyben, Weninger, Gross, & Schuller, 2013) to extract the 88 parameters included in the extended version of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS; Eyben et al., 2016) described above. The ten measures that GeMAPS computes for fundamental frequency (F0) expressed in semitones were excluded before any further analysis because they are practically identical with the ten measures of F0 expressed in hertz (leaving 78 acoustic parameters).

To control for differences in baseline parameter values between individual actor's voices, the raw values for each parameter were centred on each

actor's neutral voice. To allow for comparisons between parameters on different scales (and to aid the statistical analysis and presentation of the results) while preserving the difference between emotion expressions within each actor, all parameters for each actor were transformed to have a standard deviation of one. This standardization of the mean and standard deviation of each parameter has a similar effect as a z-transformation suggested by for example Banse and Scherer (1996) and Juslin and Laukka (2001) except that it is not sensitive to what other emotions happen to be included in the study (i.e. the z-transformed acoustic patterns for anger would look different if the low arousal expressions of sadness and serenity alone were included in the study compared with if the study also included the high arousal expressions of fear and happiness). Therefore, the results presented below are expressed as changes of the voice in relation to every speaker's neutral voice. This will allow future research to compare their results with these regardless of what emotions they include.

To obtain the most important acoustic parameters and parameter-patterns that the speakers used to express, and listeners used to classify expressions to each emotion category, a random forest approach was implemented with the "randomforest" package in R (Breiman, 2001). Random forest is a machine learning technique that builds a large number of decision trees on a set of training data and then uses the information in all trees to make an "ensemble vote" in an attempt to classify a set of test data into a number of categories (targets) specified by the type of data fed into the model. A decision tree is a recursive learning algorithm that searches for optimal values of the predictors that can be used to split the training data into "branches" corresponding to the targets that the model is trying to predict. A suggested predictor value (a split of one branch of one tree) is found by selecting a small number of predictors at random and then use the best of these to make a split. When a split has been suggested, the predictive accuracy of the tree is compared with the accuracy before the split on the test data. If the accuracy increased, the branch is kept and the algorithm continues to search for a value of some other predictor, or another value of the same predictor, to split the branch into sub-branches. This process continues either until all categories of the test data have been split into separate branches or until splitting no longer adds predictive accuracy to the model; this is the point when one decision tree is finished. Repeating this process so that many trees are "grown" creates a random forest. When the forest consists of many trees (number specified by the user), each tree classifies the unseen data in the test data set into the target categories. The category that a majority of the trees suggest (the ensemble vote) will be the suggested category of the random forest.

In the current study, three random forest models were fitted. The first model was aimed to test how well the expressions could be classified as the

intended emotion based on the acoustic parameters (classification-model). This model used the intended emotion of each expression as the classification (target) variable. The second model aimed to find the acoustic patterns that the actors used to express each emotion and intensity (expression-model). This model used the intended emotion and intensity of each expression as the target variable. Because the neutral expressions were used as a baseline for the other emotions (and because the acoustic patterns of a neutral voice is irrelevant in the context of the study), neutral expressions were excluded from this model. The third model was aimed to find the acoustic patterns that the listeners used to classify expressions to an emotion label (perception-model). This model used the selected emotion label of each judgment of each listener as the target variable. To minimize the influence of guessing (correct or incorrect), judgments for which listeners rated that they were uncertain were removed. Also, judgments for which "neutral" was selected were also removed because this would force the model to look for the acoustic patterns characterizing the perception of neutral expressions (which may be confused with other emotions). However, because the intended emotion was irrelevant in this model, expressions intended as neutral were not removed (for example, one expression intended as neutral was classified as serenity by a majority of listeners). Removing uncertain and neutral judgments left 6932 of the 12697 judgments for the model to classify (N judgments per emotion: anger: 511, contempt: 505, disgust: 215, fear: 521, happiness: 662, interest: 851, lust: 431, pride: 606, relief: 352, sadness: 776, serenity: 841, shame: 311, tenderness: 350). All three models were first fitted with the 17 acoustic parameter suggested by Scherer (2018), and then with all the 78 acoustic parameters in eGeMAPS as predictor variables.

To get a robust measure of the prediction accuracy and acoustic patterns of each model (i.e. to avoid overfitting), a cross validation-approach was used. To obtain independent train and test sets, the classification-model and the expression-model were trained on 80% of the actors and predicted what emotion (*and* intensity for the expression-model) the excluded actors were trying to express. The same procedure was used for the perception-model except that it tried to predict what emotion label the excluded listeners would select on each expression. Because actors may express emotions differently and listeners vary in how they judged the expressions, the cross validation-approach was repeated 40 times for each model, with randomly selected training and test sets in each repetition. The predictive accuracy and misclassification patterns, the parameter-patterns, and the most important acoustic parameters for each emotion presented below are based on mean values across the 40 iterations of training and testing the models with different sets of data.

Results and discussion

Recognition and model-classification accuracy

Accuracy and misclassification of all expressions

The listeners' recognition accuracy and the misclassification patterns for each emotion intended by the actors are presented in Table 1. As seen in the bold numbers in the diagonal, results show that all emotions were correctly identified by a larger proportion of the listeners than would be expected by chance (which is 7.14 percent in a 14 forced alternative choice task). However, the recognition accuracies were generally low (36% across all emotions), especially for disgust, relief, shame, and tenderness. Also, many expressions were misclassified as other emotions than the one intended by the actor. It should be noted that there was no pre-selection of expressions, which is common in other studies. The recognition accuracies and misclassifications presented in Table 1 are thus based on both "good" and "bad" expressions (see Table 3 for the corresponding results for "good" or "well-recognized" expressions only). The rows in Table 1 show the percentages of misclassification for each emotion. These misclassifications can be thought of as measures of conceptual and/or expressive closeness between different emotions.

Looking at emotions misclassified by ten percent or more, results show that anger was often misclassified as contempt; contempt was misclassified as pride; disgust as contempt, pride, and sadness; fear as sadness; happiness as pride; lust as serenity; pride as happiness and interest; relief as lust and serenity; sadness as fear and shame; serenity as tenderness; shame as sadness; and tenderness was misclassified as serenity. Also, many emotions were misclassified as neutral, and expressions intended as neutral were often classified as interest or serenity. The bottom row of Table 1 shows the sum of each column and can thus be interpreted as the listeners' bias for or against selecting each emotion label. A sum of 100 would mean that there was no bias for or against selecting this emotion label (i.e. that it was selected as many times as there were expressions of that emotion). These numbers show that listeners had a bias towards selecting interest, sadness, serenity, and especially neutral. They also show that listeners had a bias against selecting disgust, lust, relief, shame and tenderness.

Table 1. The listeners' recognition accuracy, misclassifications (in percent), and bias for the actors' intended emotions (all expressions and judgments)

Emotion Intended	Selected ang	con	dis	fea	hap	int	lus	pri	rel	sad	ser	sha	ten	neu
ang	**49**	16	4	3	2	8	0	7	1	1	0	2	0	6
con	5	**28**	7	2	2	10	1	14	3	1	4	3	1	19
dis	3	16	**16**	6	4	6	1	11	3	11	3	9	1	11
fea	8	6	3	**47**	2	3	1	2	2	13	1	6	1	5
hap	4	3	2	3	**45**	17	1	10	6	2	1	1	1	4
int	0	2	0	1	9	**43**	0	6	5	1	4	1	3	22
lus	2	5	3	3	2	7	**37**	4	4	4	14	3	7	6
pri	2	6	1	0	13	18	0	**26**	4	1	4	1	2	22
rel	1	5	2	3	4	9	11	8	**19**	3	11	3	6	14
sad	1	2	2	14	1	1	0	1	2	**55**	4	10	2	6
ser	0	1	0	1	1	3	3	1	6	9	**39**	5	12	18
sha	1	6	3	6	0	2	2	2	4	22	10	**19**	4	20
ten	0	0	1	1	4	7	6	2	6	7	28	4	**20**	15
neu	1	2	0	2	2	10	1	3	2	3	12	1	5	**57**
Bias	78	98	43	93	90	144	66	98	67	130	136	68	64	225

Table 2 shows the accuracy for all expressions of the "classification-model" fitted with the 17 acoustic features described as especially important for emotion classification by Scherer (2018). This model was intended to study how well each intended emotion could be classified from these acoustic features. Also, the accuracy and the patterns of misclassifications can be compared with those of the listeners in Table 1. Looking at the diagonal of this table shows that all intended emotions except disgust and lust were classified correctly more often than would be expected by chance by the model. The overall classification accuracy was 34%.

Looking at the row presenting the bias of the model shows that there was a bias toward classifying expressions as anger, happiness, interest, relief, serenity, tenderness, and neutral. The model's biases were thus more spread out across the emotions compared to the listeners' biases. To get an estimate of the similarities between the listeners' and the model's misclassifications, the correlations of each column (leaving out the accuracy-cell) of Table 1 and Table 2 are presented in the row below the biases. Correlations were generally high (mean r = .47), except for lust and relief, suggesting that classification based on the 17 acoustic features in general renders similar misclassifications as those made by the listeners.

Adding the additional 61 acoustic parameters included in eGeMAPS increased the overall accuracy somewhat to 39%. The accuracies, biases, and correlations with the listeners' misclassifications of the model with all 78 parameters are presented in the three bottom rows of Table 2. These numbers show that the accuracy for individual emotions generally increased, especially for disgust, serenity, and shame, and that all emotions were cor-

rectly classified more often than expected by chance. The correlations between the listeners' and the model's misclassifications also suggest a higher degree of similarity compared with the model with 17 parameters.

Although the model with all parameters performed better and had more similar misclassifications compared with the listeners, it seems as if the 17 parameters in the first classification-model were adequate to capture most of the variation in how the actors expressed the emotions and how the listeners perceived them. However, because the classification increased for some emotions, there might be additional acoustic information that listeners use to recognize them, especially for disgust, serenity, and shame, that is not captured by the 17 acoustic features suggested by Scherer (2018).

Table 2. The "classification-model's" classification accuracy, misclassifications (in percent), and bias for the actors' intended emotions (all expressions)

Emotion Intended	Selected ang	con	dis	fea	hap	int	lus	pri	rel	sad	ser	sha	ten	neu
ang	**67**	1	3	3	9	3	2	6	4	0	1	0	0	1
con	14	**17**	8	0	5	3	8	6	9	0	9	11	3	6
dis	3	10	**5**	6	10	7	11	5	13	5	5	8	6	5
fea	10	1	1	**36**	14	6	3	0	6	10	0	3	9	1
hap	18	1	1	23	**26**	13	1	7	5	3	0	0	0	1
int	5	0	3	11	9	**50**	0	5	3	2	1	0	4	6
lus	4	10	10	0	3	3	**7**	9	11	0	14	11	15	3
pri	11	6	8	1	23	15	6	**8**	9	1	2	0	3	8
rel	7	6	3	1	2	5	4	9	**33**	0	4	4	13	9
sad	6	1	8	14	9	11	1	2	9	**14**	6	7	11	2
ser	0	10	4	0	0	1	8	0	4	0	**29**	6	36	2
sha	0	8	6	0	1	3	11	1	13	1	18	**22**	14	4
ten	0	3	2	5	0	2	7	0	4	1	27	8	**39**	2
neu	0	0	0	0	0	0	0	1	0	0	0	0	4	**95**
Bias	146	75	63	101	109	121	68	60	124	38	117	79	157	144
Correlation	0.65	0.22	0.38	0.38	0.66	0.48	0.17	0.69	-0.05	0.39	0.80	0.52	0.87	0.65
78 Parameters	**69**	**15**	**18**	**40**	**31**	**46**	**9**	**12**	**31**	**16**	**45**	**41**	**47**	**100**
Bias	148	67	86	85	110	118	58	62	110	38	139	94	133	158
Correlation	0.80	0.14	0.55	0.60	0.68	0.50	0.10	0.76	0.05	0.16	0.75	0.71	0.81	0.58

Accuracy and misclassification of "well-recognized" expressions

The listeners' recognition accuracy and the misclassification patterns for expressions that were perceived as a specific emotion (regardless of what the actor was trying to express) by at least 48% of the listeners' "confident" judgments are presented in Table 3. The cut-off at 48% was selected because this was the highest cut-off that included at least one expression from each emotion. As described in the methods section, "confident" judgments were based on the listeners' self-reports. In total, 317 of the 756 expressions were "well recognized" as a specific emotion based on this requirement.

Table 3. The listeners' recognition accuracy, misclassifications (in percent), bias, and number of expressions and judgments for expressions perceived as one emotion by at least 48% of the listeners (“well-recognized” expressions and "confident" judgments)

Emotion Intended	Selected ang	con	dis	fea	hap	int	lus	pri	rel	sad	ser	sha	ten	neu
ang	**76**	12	2	1	0	3	0	0	0	0	0	0	0	4
con	6	**58**	18	0	1	1	2	1	0	1	0	0	4	8
dis	0	7	**57**	0	0	0	0	0	7	7	0	7	0	14
fea	3	2	2	**69**	2	1	2	1	11	1	2	0	2	0
hap	1	1	0	0	**72**	12	0	3	1	0	0	0	3	7
int	1	2	1	0	7	**65**	2	3	0	3	1	2	9	5
lus	0	2	0	0	0	2	**74**	6	1	8	0	3	1	1
pri	0	0	3	0	3	0	5	**73**	0	0	3	3	3	8
rel	0	1	1	10	1	0	0	0	**73**	3	7	0	3	0
sad	0	0	0	1	0	2	3	4	2	**63**	2	9	13	2
ser	0	5	0	0	0	0	5	5	10	0	**75**	0	0	0
sha	0	0	0	0	9	0	0	0	0	14	0	**64**	14	0
ten	1	1	1	1	2	9	0	1	1	9	0	3	**65**	4
neu	1	9	1	0	9	10	1	1	1	1	1	0	4	**63**
Bias	89	101	86	82	107	106	95	98	108	109	91	92	121	116
N expressions	33	14	1	28	36	32	25	11	3	48	25	2	2	57
N judgements	379	198	70	293	425	399	262	212	98	479	301	78	84	527

The two bottom rows of Table 3 show the number of well-recognized expressions and the number of confident judgments for these expressions for each emotion. Very few expressions were well recognized as disgust, relief, shame, and tenderness. However, the accuracies for each emotion are based on at least 70 confident judgments.

Because Table 3 shows the listeners recognition accuracy for confident judgments only, and only for the “best” expressions from each emotion, the purpose of this table is to show that even though the recognition accuracy was generally low for all of the intended emotion expressions, there were still many expressions that a large proportion of the listeners perceived as specific emotions. The mean accuracy for these expressions was 68% with the lowest accuracy for disgust at 58%. Also, the misclassification patterns of these judgments reveal that even though the listeners reported that they were confident in their judgment, some emotions were still more commonly misclassified than others. The patterns presented in this table are similar to, but perhaps a bit clearer than, those shown in Table 1. Anger was often misclassified as contempt; contempt as disgust; disgust as neutral; fear as relief; happiness as interest; relief as fear; sadness as tenderness; serenity as relief; shame as sadness; and neutral was often misclassified as interest. The ex-

pressions of interest, lust, pride, and tenderness were not misclassified in more than ten percent of the judgments. Among these expressions and judgments there were less bias toward or against selecting a specific emotion label compared with those observed for all expressions and judgments. However, there was still a bias towards selecting neutral.

Acoustic parameter patters communicating discrete emotions in speech

This section presents the results from the two models intended to obtain the acoustic parameter patterns that the actors used to express each emotion and intensity (expression-model) and those that listeners used to classify emotions (perception model). The results from both models, for each emotion on separate pages, are presented together with the predicted parameter patterns in Figure 3 (pages 41-53).

The two blue lines show the acoustic parameter patterns that the expression-model used to classify an expression to the relevant emotion category and intensity. Because the acoustic parameters were centred on each actor's neutral voice, deviations from zero reflect how much the emotion-expression influenced the acoustic parameters, measured in standard deviation from the actor's normal tone of voice. Thus, the patterns across the 17 acoustic parameters reflect how the actors manipulated their voices to express the intended emotion with either more (dark blue) or less (light blue) intensity.

The red line shows the acoustic parameter patterns of the expressions that the perception-model used to predict that a listener would select the relevant emotion category. Because uncertain judgments were removed before fitting the model, these parameter patterns reflect how the actors manipulated their voices in those cases where the listeners perceived the expression as the relevant emotion (whether it was intended by the actor or not).

The theoretical predictions derived from Scherer (2018) for each acoustic parameter are shown as grey areas and the results from the three previous empirical studies are shown as colored areas. In the title on each page of Figure 3, the number of expressions perceived as the relevant emotion by at least 48% of the listeners is shown. These numbers can be used as an estimation of the generalizability of the acoustic patterns. For some emotions, only a few expressions were perceived as the relevant emotion. In these cases, the results are more influenced by the acoustic patterns of the few expressions that listeners could recognize and are therefore probably less generalizable. The subtitle presents how many of the expressions that the listeners perceived as the relevant emotion were intended as that emotion, and also how many were intended as another emotion.

Comparison of emotions expressed with more and less intensity

Comparing the dark- and light blue lines for each emotion suggests that the actors manipulated their voices differently for expressions with more and

less intensity for most of the emotions. More intense emotions were generally expressed with larger deviations from the actors' normal voices, with more loudness with a larger range, higher pitch (F0) with a larger range, and higher first formant (F1) with a smaller bandwidth. As indicated by the relatively low parameter-values for hammarberg index and harmonic difference, and the high values for harmonic to noise ratio (HNR), the actors generally expressed intense emotions with a sharper and less "breathy" (i.e. more clear pitched) voice. This was true not only for the negative emotions, but also for the more arousing positive ones; happiness, interest, lust, and pride, although these patterns were especially accentuated for expressions of anger, fear, and sadness. Exceptions to these general effects of expression intensity were observed for relief, serenity, and shame, which showed practically the opposite patterns to those described above. This seems to make sense because these low-arousing emotions ought to be expressed with even less arousal when expressed in a "more intense" version.

Comparing the acoustic parameter-patterns derived from the expression-model (blue lines) with those from the perception-model (red line) suggests that the listeners' perception of anger fear, interest, pride, sadness, and serenity mostly corresponded to the actors more intense expressions of these emotions. Listeners' perceptions of disgust and happiness however, corresponded more to the acoustic patterns of the less intense expressions. For the other emotions, it was not as clear which intensity level the listeners perceived as the relevant emotion. Often there was a more or less clear pattern that listeners' perception corresponded to the more intense expressions for some acoustic parameters but not for others (look at relief for example).

Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

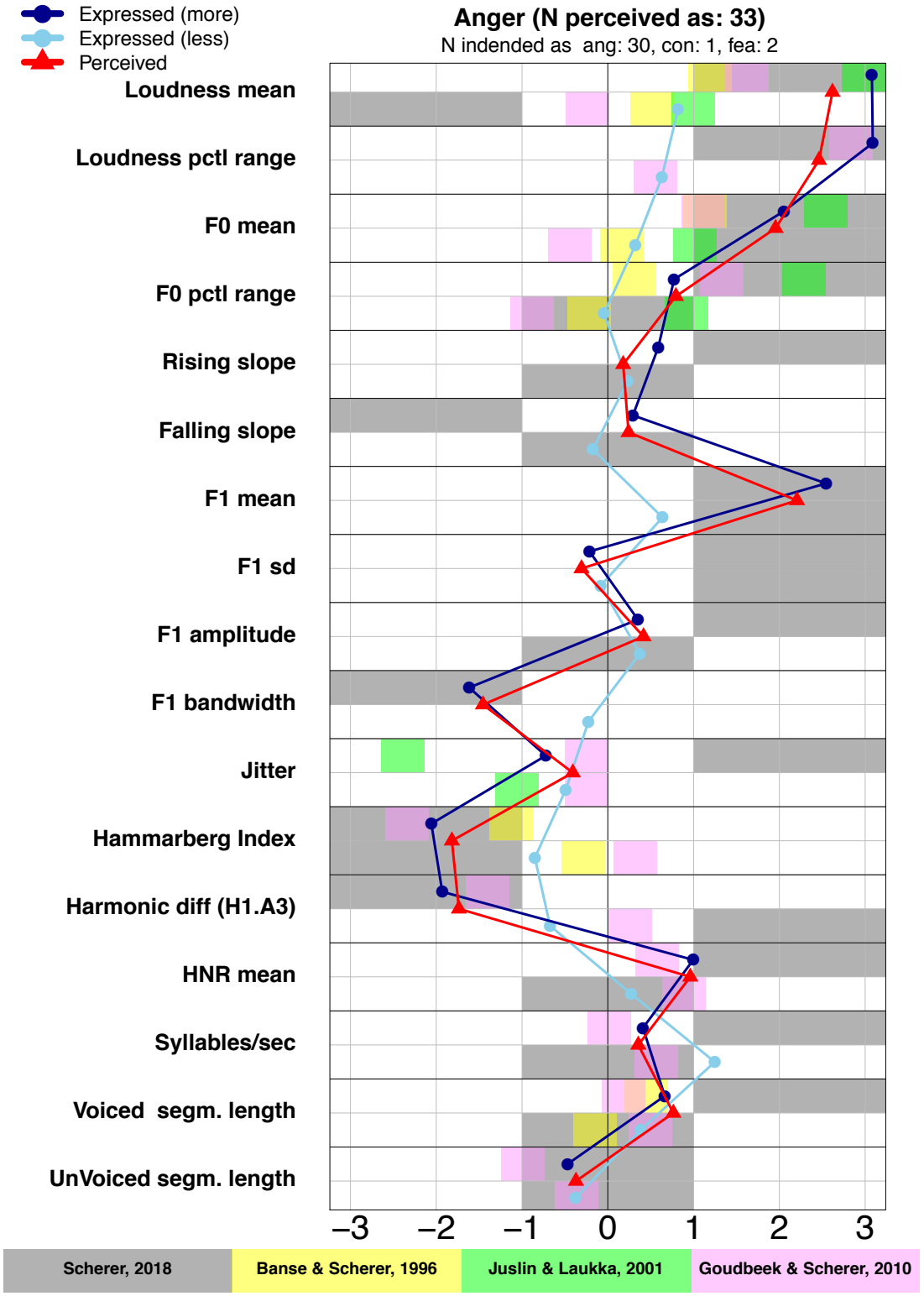


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

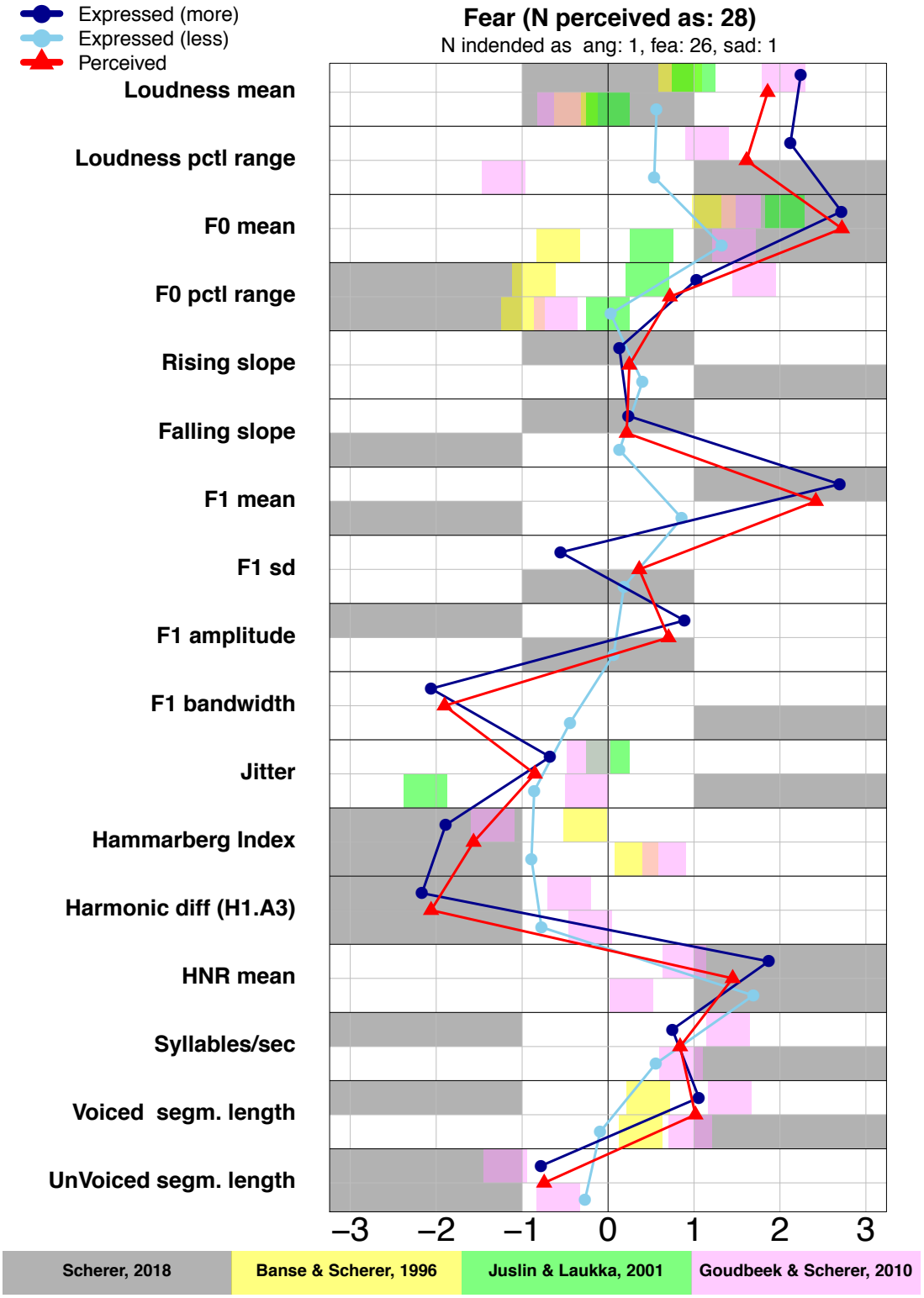


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

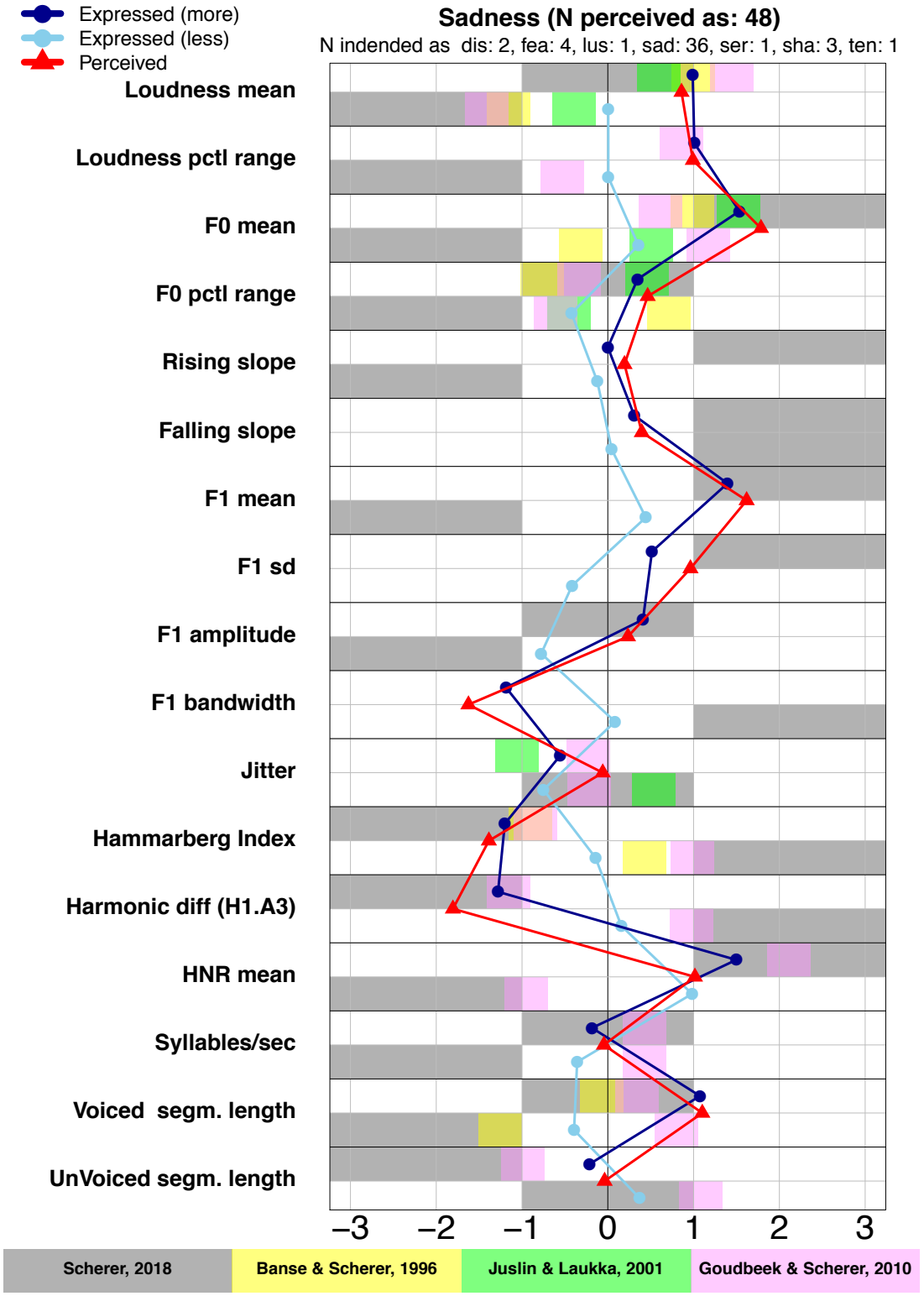


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

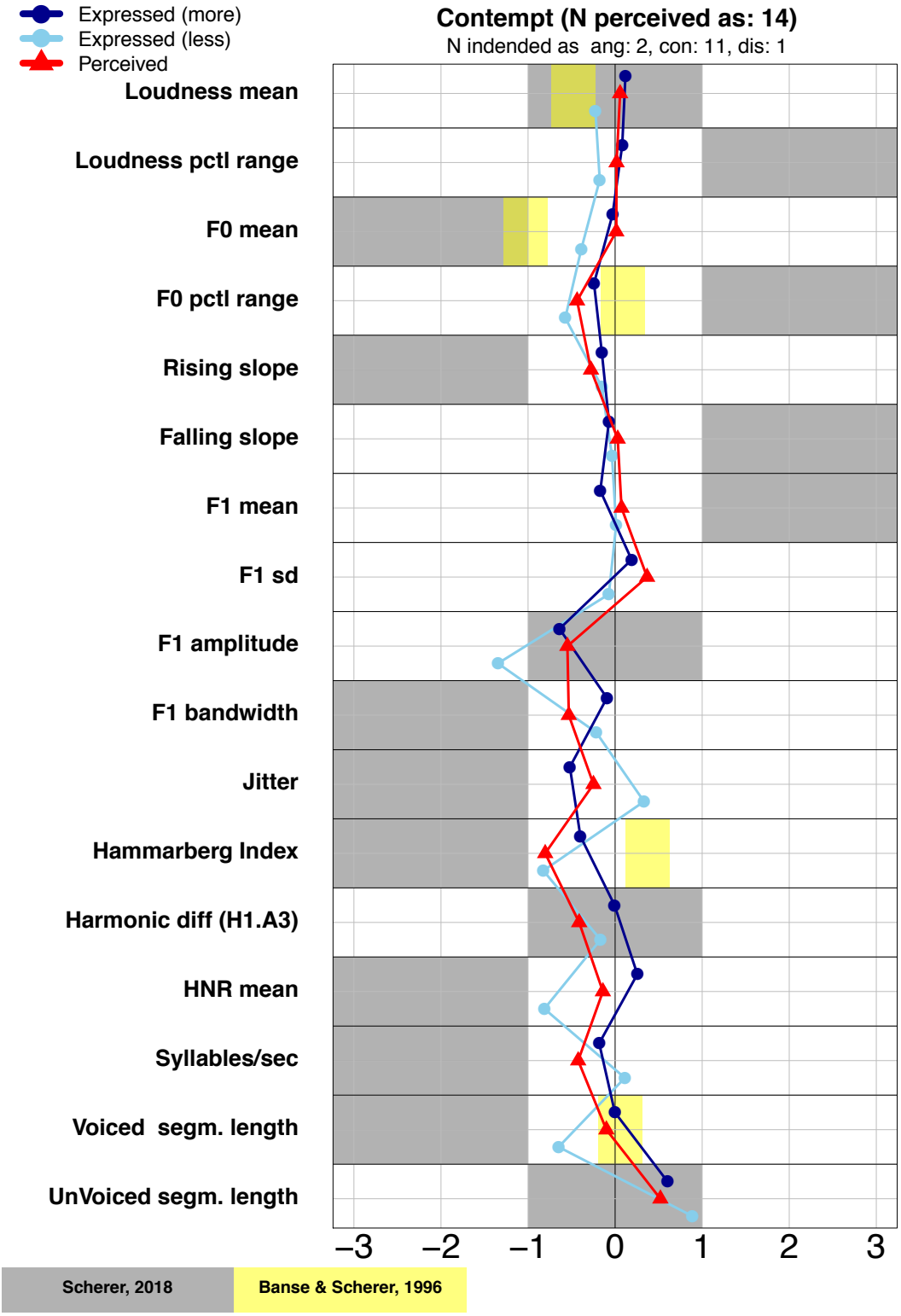


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

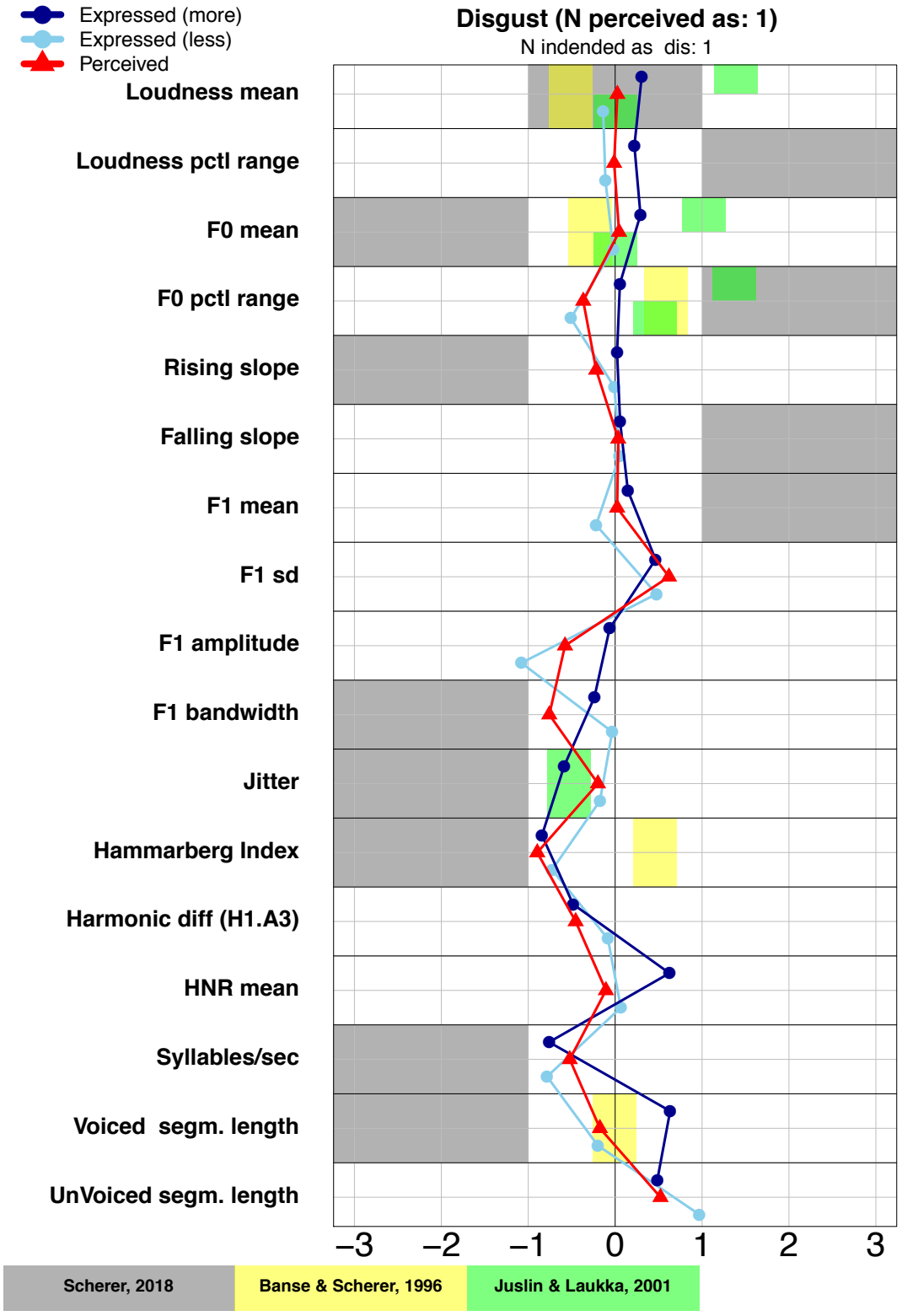


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

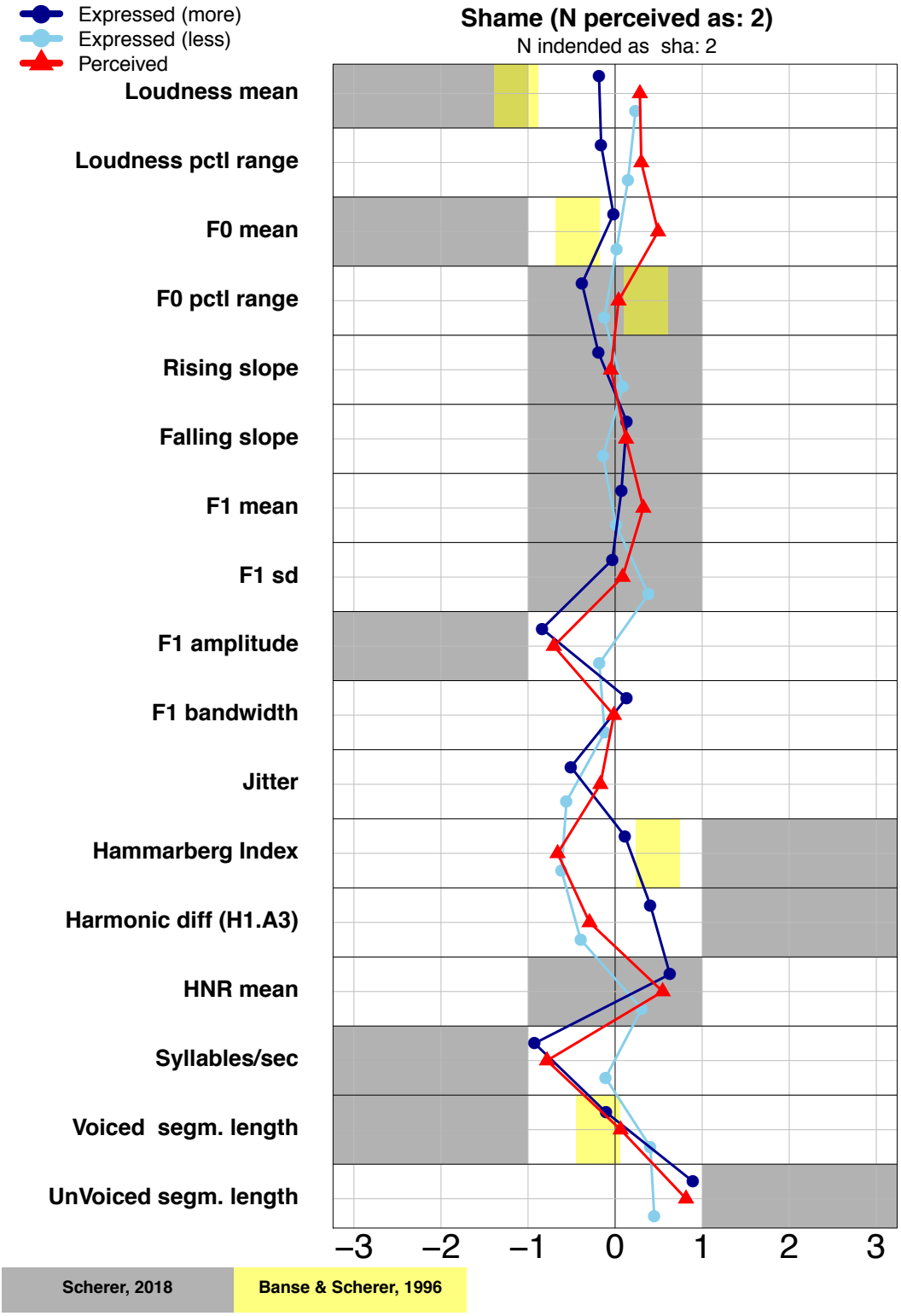


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

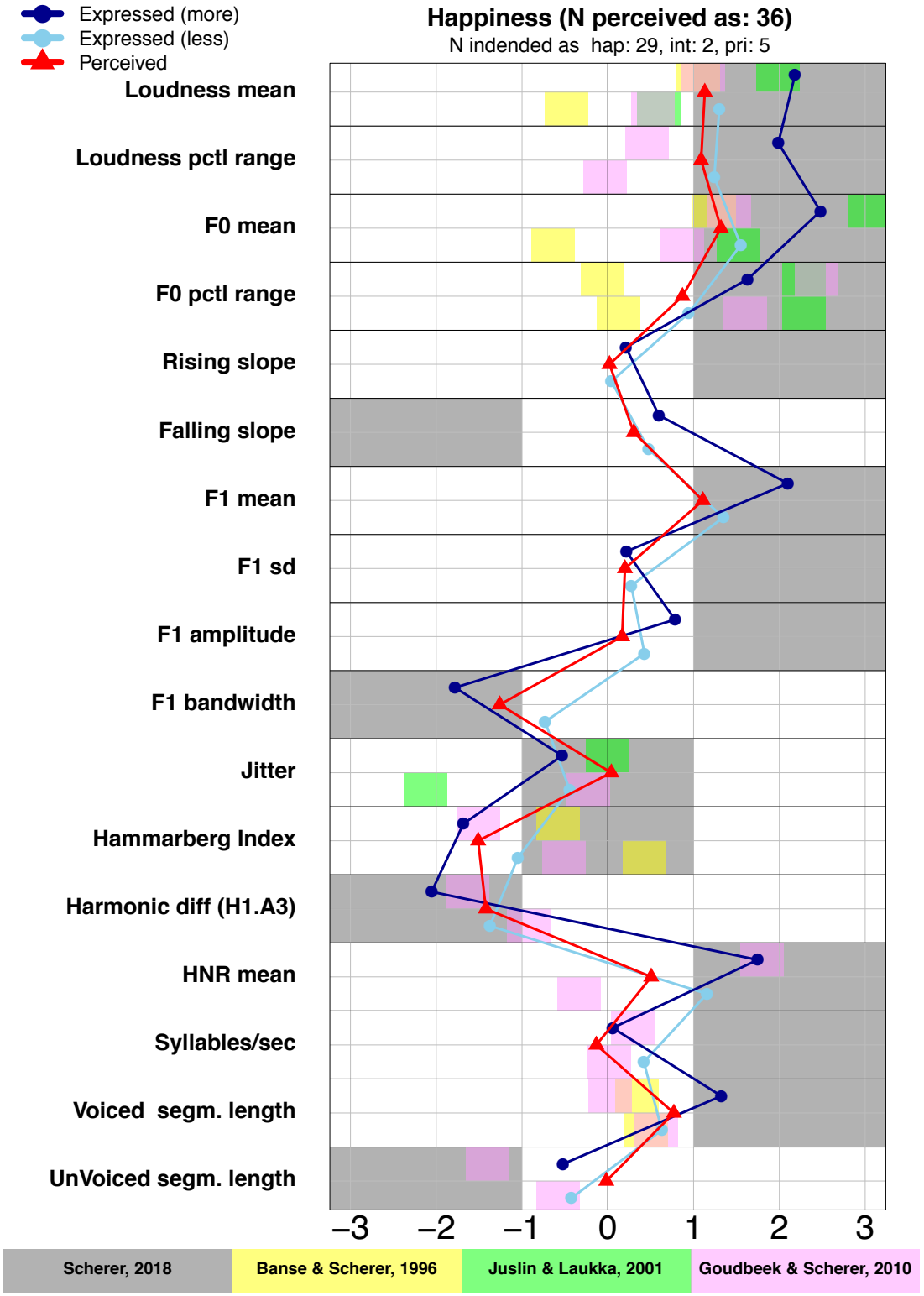


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

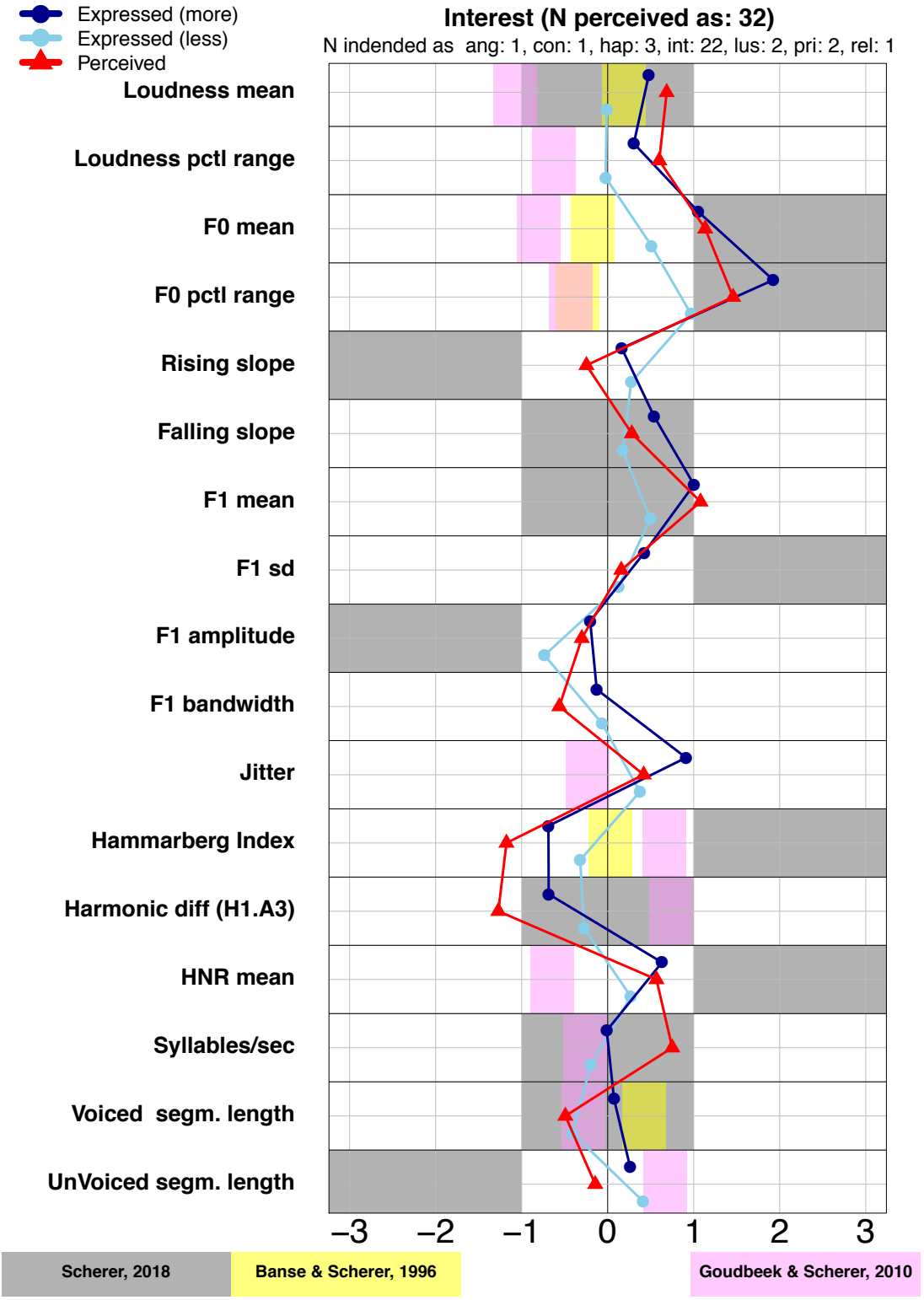


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

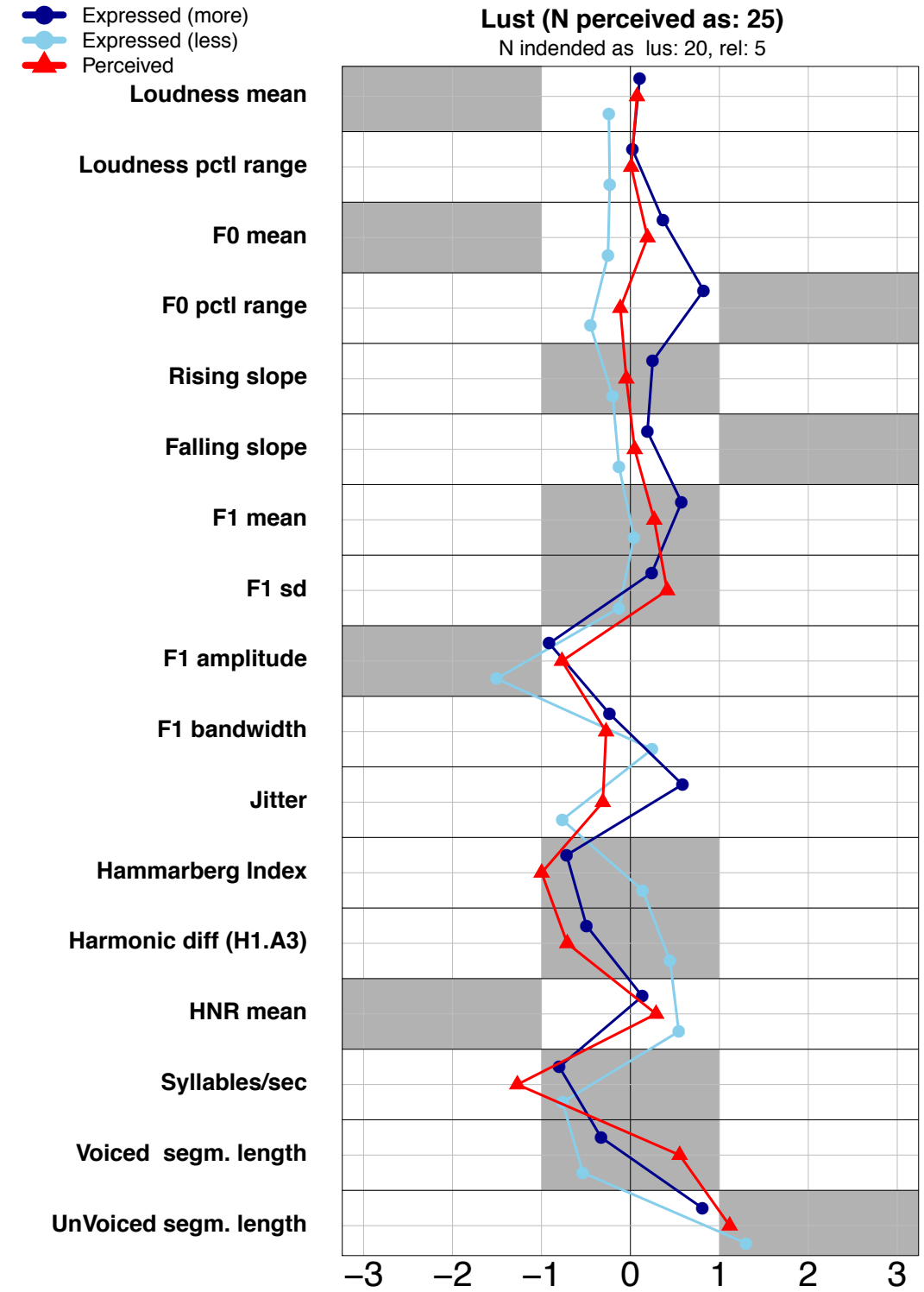


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

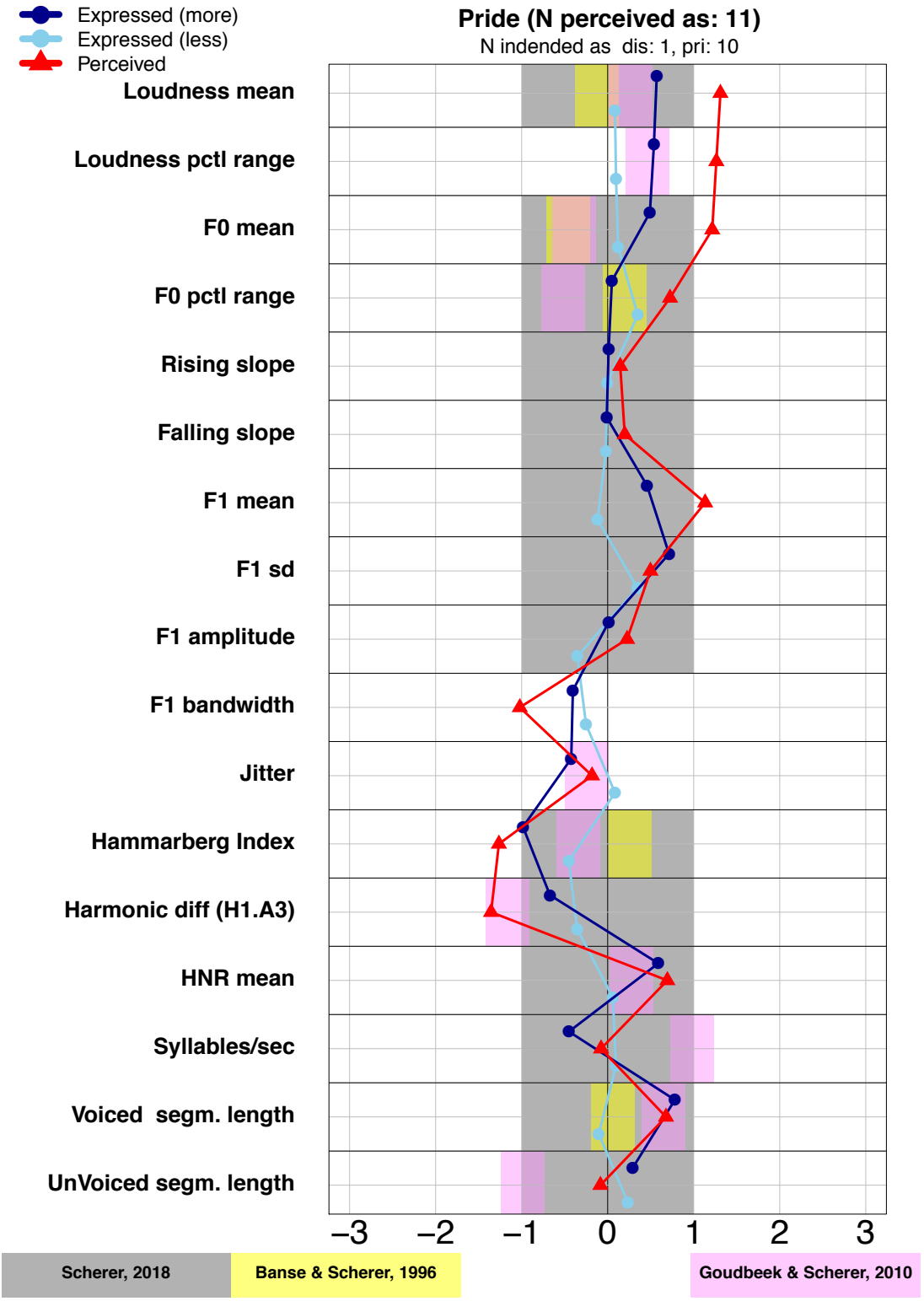


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

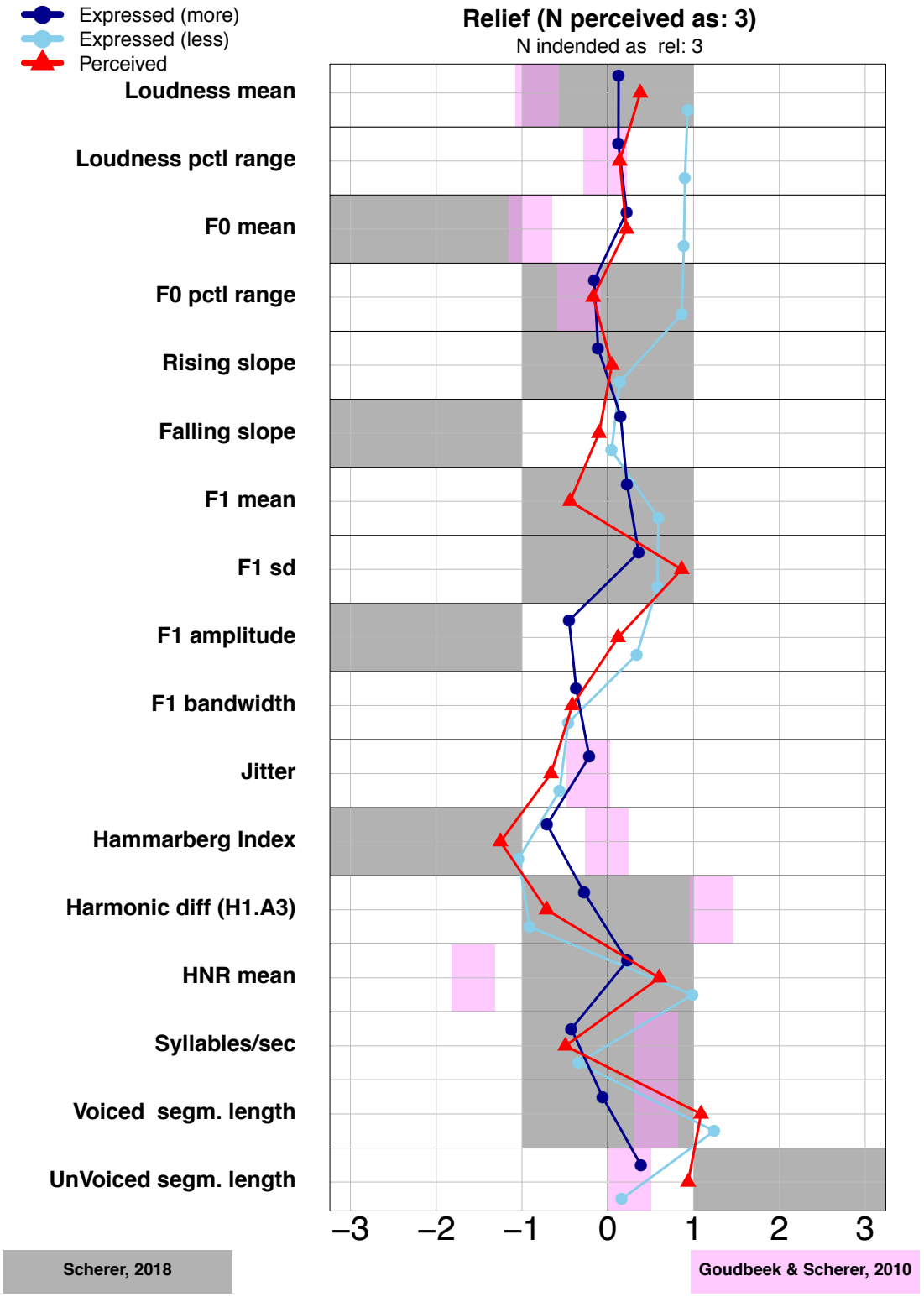


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.

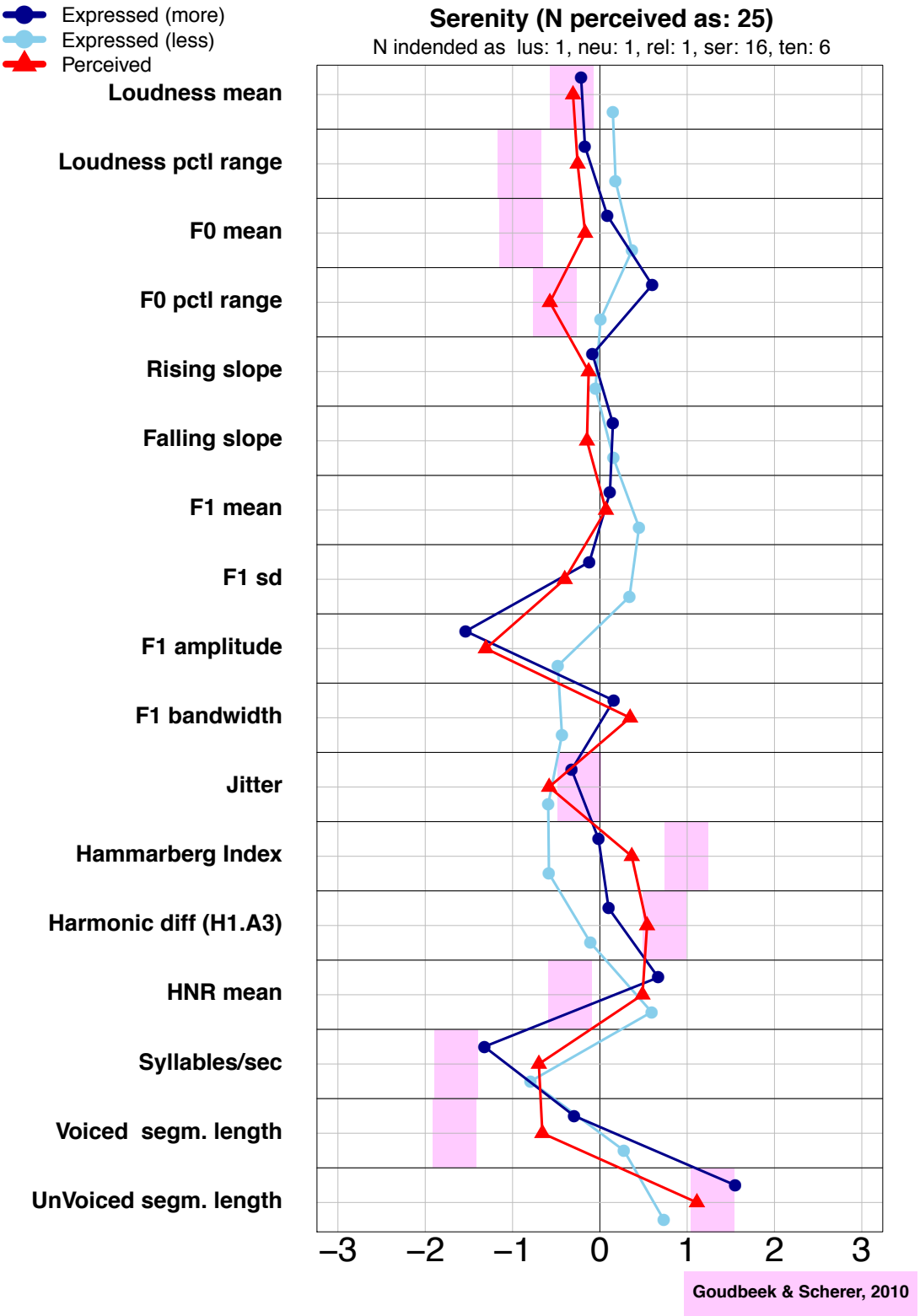
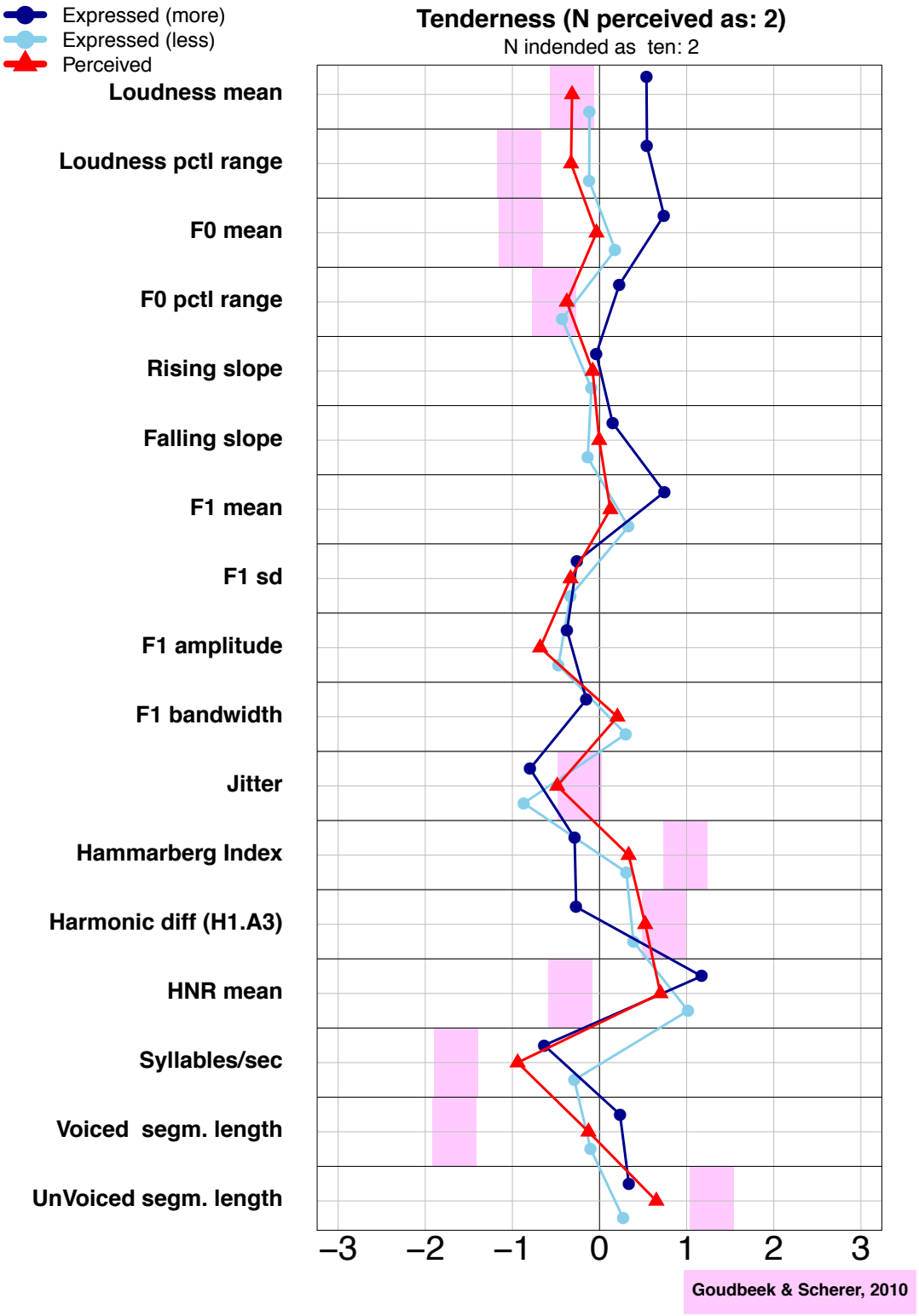


Figure 3 (pages 41-53). Acoustic parameter patterns for each emotion.



Evaluation of the theoretical predictions

Comparing the acoustic parameter-patterns that the actors used to express (blue lines), and listeners perceived as (red lines) different emotions with the theoretical predictions (grey areas) suggest a high degree of correspondence for many emotion-parameter combinations. However, the deviations observed may be important to further advance our knowledge about how these parameters are used in emotional communication.

The theory predicts that **anger** with less intensity should be expressed and perceived as having low loudness. Both current results and the results of two of the previous empirical studies instead suggest that less intense anger is expressed with higher loudness than normal. It could be argued however, that anger with less intensity is qualitatively different from what Scherer (2018) describes as "cold" anger. Another deviation between the theoretical predictions compared with the current results and previous empirical findings, was observed for intense anger and jitter, for which the predictions seems to go in the opposite direction than the empirical findings.

The theoretical predictions suggest that **fear** is expressed and perceived with medium loudness regardless of intensity. Current results rather correspond to the previous empirical findings that intense fear is associated with high loudness. Also, the percentile range of F0 is predicted to be low for fear whereas the current results suggest that it is medium to high. The previous studies have shown differing results for this parameter, ranging from low to high regardless of intensity. Similar results seem evident for jitter, hammarberg index, and speech rate. The inconsistent results regarding expressions of fear may indicate that fear might be expressed with and recognized from many different acoustic patterns. As suggested by the large differences observed between expressions of fear with more or less intensity in the current results might shed some light on the inconsistent results. Perhaps these expressions represent different types of fear that could be differentiated in future studies by instructing the actors differently. If the emotion "family" of fear contain many types of expressions it is probable that the type of instructions that the actors get will have a greater influence on the type of expressions they produce compared with other emotions.

Similar to fear, the previous findings for **sadness** differ a lot between the three empirical studies. The theoretical predictions for more intense sadness corresponded well to the results of the current study. However, the predictions for less intense sadness differed a lot from the predictions. The large differences observed between expressions with more and less intensity might, again similar to fear, suggest that sadness can be expressed in many ways and could be differentiated into several expressions in future studies.

For **contempt**, **disgust** and **shame**, it is difficult to evaluate if the theoretical predictions corresponded to the actors expressions/listeners perception because these patterns seem to deviate very little from the actors' normal

voices. However, the ranges of loudness and F0 seem to go in the opposite direction than the predictions for contempt and disgust, and perhaps also for shame. Also, the amplitude of the first formant (F1), predicted to be around zero for contempt and disgust seems to be one of the few parameters that actually changed for these expressions compared with the actors normal voice (which was predicted for shame only). All three emotions were expressed and perceived as having a relatively slow speech rate and long pauses (which was not predicted for disgust).

The results for **happiness**, **interest**, **lust**, and **pride** generally corresponded to the predicted directions for all parameters except falling slope for happiness, jitter for interest, and harmonics to noise ratio for lust. Predictions for hammarberg index and harmonic difference, both reflecting sharpness of the voice, suggest that these four emotions are expressed and perceived with about similar sharpness as a neutral voice, or softer voice for interest (although the theoretical prediction for harmonic difference is contradictory for happiness). Both the current results and the empirical predictions rather suggest that at least happiness and pride, but perhaps also interest and lust, are expressed with a sharper voice than normal. It should also be noted that the theoretical predictions for pride suggest medium values for *all* parameters (i.e. suggesting a similar acoustic pattern as the normal/neutral voice). These non-informative predictions may reflect that the results of previous studies have shown both high and low values for the parameters making general conclusions difficult. The current results suggest that pride is expressed and perceived with moderately high loudness and loudness range, pitch and pitch range, and F1 with a small bandwidth, a sharper voice than normal, and about the same speech rate as normal.

Predictions for **relief** corresponded well with the predictions for the measures of loudness, voice quality, speech rate, and pauses but not for mean F0. Similar to pride, theoretical predictions suggest "medium" level for many parameters. From the results however, one may argue that perception of relief correspond to a voice with low F1 with more variation, with slower speech rate and a bit more sharpness than the person's normal voice.

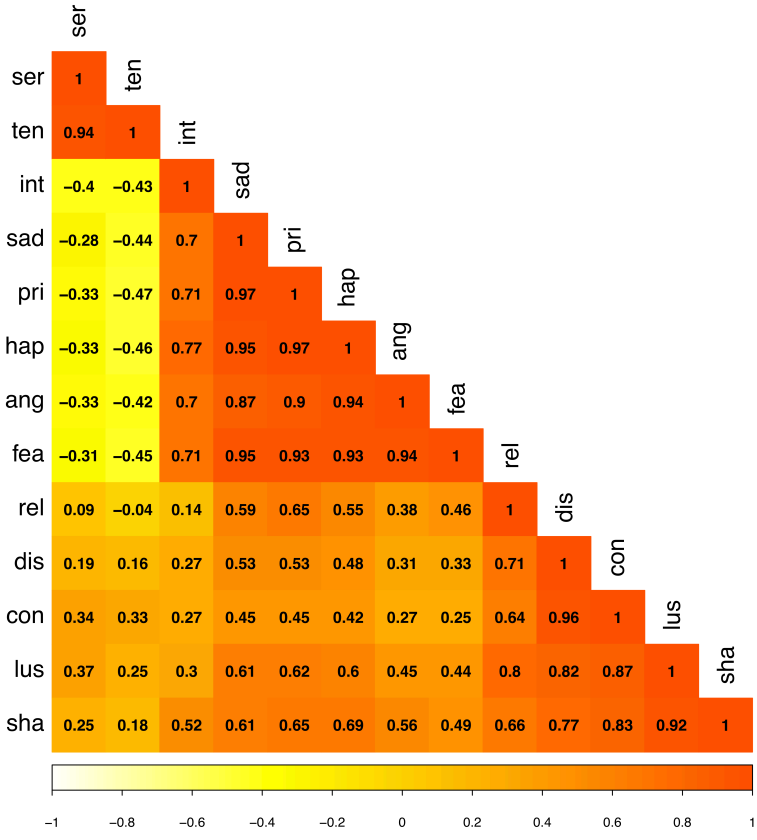
There were no predictions for **serenity** and **tenderness** but the study by Goudbeek and Scherer (2010) presented acoustic patterns for "pleasure". To have something at least remotely similar to compare with, their results for "pleasure" are shown together with the acoustic patterns for serenity and tenderness from the current study. Because they do not define "pleasure" or give a description of how the actors were instructed to express this emotion, I will not discuss potential similarities and differences from their results. The acoustic pattern for serenity differed from the actors normal voices in that these expressions and perceptions had a lower amplitude of the first formant (F1) and that the speech rate was slower than normal. Tenderness was also expressed and perceived with a slower speech rate than normal but was also

characterized by a more clear-pitched voice as indicated by the high values for harmonic to noise ratio (HNR).

Apparent similarities in acoustic patterns of different emotions

Looking at the acoustic parameter-patterns that the listeners used to classify emotions gives the notion that many are recognized with apparently similar patterns. Figure 4 quantifies this notion by showing the correlations across the 17 acoustic parameters across all emotions (patterns derived from the perception-model). These correlations suggest that the acoustic patterns that listeners perceived as the thirteen different emotions can be categorized into three more or less distinct groups of patterns. Serenity and tenderness were perceived with very similar patterns and they deviated the most from the other emotions. Interest, sadness, pride, happiness, anger, and fear make up a second group of emotions, which also were perceived with seemingly similar acoustic patterns. A third group comprise relief, disgust, contempt, lust, and shame.

Figure 4. Correlation matrix (Pearson r) across the acoustic parameter-patterns for expressions that listeners perceived as different emotions.



Within the second group of emotions with seemingly similar patterns it may be noted that the listeners seldom confused anger with any of the other emotions in this group; sadness was only confused with fear; and happiness, pride, and interest were only confused with each other and seldom with the other emotions (see Tables 1 and 3). Because all these emotions have highly correlated acoustic patterns, one may conclude that the listeners' used some other information in the actors' voices that was not captured by these 17 acoustic parameters.

Another interpretation might be that there are a few important parameters that go in opposite directions and that these parameters allow listeners to differentiate these expressions. For example, anger and happiness were almost never confused with each other; still the acoustic patterns had a correlation of .94. Looking at the theoretical predictions (grey areas) for anger and happiness, it is evident that the only postulated differences in the predictions between the emotions concern jitter, hammarberg index, and the length of pauses. Therefore, one may assume that these parameters are thought to be especially important for the ability to differentiate anger and happiness. However, direct comparisons of the acoustic patterns that the listeners in the current study used to differentiate anger and happiness reveal that these parameters do not differ much between the emotions. Instead, although both emotions were recognized from a loud voice with much loudness variation, anger was almost 1.5 standard deviations further from the speakers' normal voice compared with happiness. Also, for all parameters that did not overlap exactly, anger deviated more from zero than happiness. Thus, even if the acoustic patterns are very similar with regard to the direction of the parameters (as the strong correlation suggests), they instead seem to differ in how much each parameter deviates from the actors normal voices (i.e. from zero in the figures). Such absolute differences may be more important for listeners' judgments than the direction of specific acoustic parameters, especially between emotions with similar activation level but different valence.

However, it should also be noted that many of the emotions with highly correlated acoustic patterns were commonly confused with each other. For example, direct comparison of the acoustic patterns between **serenity** and **tenderness** (r = .94) reveals that they are almost identical except that serenity has slightly lower F1 amplitude and slower speech rate with longer pauses. Therefore, the listeners' common confusions between these emotions are not surprising.

Conclusions

This chapter gave an overview of how emotions are communicated via nonverbal aspects of a speaker's voice and presented data from a yet unpublished study evaluating one of the most influential frameworks in the

field. Comparisons between the current results with those of three previous empirical studies (Banse & Scherer, 1996; Goudbeek & Scherer, 2010; Juslin & Laukka, 2001) and the theoretical predictions suggested by Scherer (2018) reveal a high degree of similarities but also differences that may be important to further our understanding of how emotions are communicated in speech. Most notably, the framework that the theoretical predictions were based upon (Scherer; 2018, Scherer & Juslin 2005; Juslin & Laukka, 2003; Scherer, Johnstone, & Klasmeyer, 2003; Scherer, 1986; Bänziger, Mortillaro, & Scherer, 2012; Laukka et al., 2016), present predictions for specific acoustic parameters in a "high/medium/low"-format. Although this is a practical way to present the complex and sometimes contradictory findings of many empirical studies, I believe that the mechanisms allowing listeners to infer discrete emotions from nonverbal aspects of a speakers voice should be described as absolute differences, both between the speakers normal voice and between conceptually similar emotions and emotions with seemingly similar patterns, rather than focusing on the direction of the acoustic parameters.

Results suggest that the coarse patterns presented by these theories do not capture the acoustical aspects that the listeners in the current study used to infer the emotional message of the speakers. The results of the current study may of course have been caused by arbitrary factors regarding the actors and listeners that participated in this specific study. Even so, the theoretically predicted acoustic patterns are very similar between some emotions that listeners clearly do differentiate. The fact that the predictions are presented this way either means that there are other information not captured by these parameters or that there is more information to be extracted from the suggested parameters. Because adding additional acoustic parameters only slightly improved the classification accuracy of the perception-model, the current results suggest the latter.

It may seem contradictory that both the literature reviewed, and the data presented in this chapter strongly suggest that humans can differentiate many emotional expressions from nonverbal aspects of the voice but that the acoustic patterns allowing this ability seem so difficult to describe. Because the emotional information is conveyed via such complex channels it might be the case that researchers simply have not yet been able to capture what the listeners are doing with the available information. In the literature aiming at automatic classification of emotions, it is often stated that a "golden set" of acoustic parameters is still yet to be identified. A known problem of these attempts is that they have led to a "proliferation in the variety and quantity of acoustic features employed, amounting often to several thousand basic (low-level) and derived (functionals) parameters" (Eyben et al., 2016, p. 191). Such development may lead to more reliable recognition by automatic classifier systems but they are not likely going to aid our understanding of

how humans use the information to classify emotions because the mechanisms are difficult to interpret from such analyses.

Although many acoustic parameters have been successfully used to predict the arousal-dimension of emotional communication, the valence-dimension seems to be more difficult to capture from the acoustic signal (Belyk & Brown, 2014; Eyben et al., 2016). The fact that listeners usually are good at differentiating positive and negative valence while the acoustic correlates of this ability seem so difficult to find is puzzling. The suggested acoustic parameters that could potentially capture the valence dimension usually measure different aspects of the voice quality or temporal changes such as speech rate and pauses. Although these parameters may be important, they usually explain very little of the variance in how emotions are expressed and perceived. Again, closer inspection of the absolute levels of those parameters that usually explain most of the variance (those related to loudness and pitch) may be a potentially productive way to increase predictability of how listeners perceive the valence dimension of emotions. However, the importance of other parameters should not be reduced. It might be the case that these parameters interact with the absolute levels of other parameters that allow listeners to differentiate the expressions.

Another possible reason that the parameter patterns of different emotions have proven so difficult to find may be that the task of recognizing emotions itself is difficult, even in its simplest form, that is, for listeners performing a forced choice task. Though much research has been devoted to show that humans can perform such tasks, "recognition" is usually defined as some proportion of successful listeners, often as the mean across all expressions in an emotion category, and as a larger proportion of listeners than would be expected by chance. Thus, even though most listeners recognized most expressions in an emotion category, there may still be expressions that listeners were not able to classify. Performing acoustic analyses on the complete set of expressions may thus confound the patterns of acoustic features thought to carry the emotional information. Therefore, the use of confidence judgments could prove beneficial in future studies aiming to refine the knowledge about the acoustic patterns that listeners use to recognize discrete emotions. The listeners in the current study made many errors even when they said they were confident, again, indicating that the task of differentiating emotions is difficult. These observations may lead to questions regarding the main assumptions motivating the quest to understand how emotions are communicated; that emotions change the voice in predictable ways and that listeners have knowledge about these changes. If the acoustic patterns allowing listeners to recognize discrete emotions are still not adequately described in future research, it may become increasingly difficult to argue that these assumptions are justified. However, I believe that the field would benefit from finding ways to develop more fine-grained predictions for many differ-

ent types of expressions and to test these predictions directly in a more cumulative approach than previously done.

A development towards more fine-grained predictions beyond high/medium/low for different types of expressions will require more methodological alignment in how studies are designed, how actors are instructed, and how the acoustic analyses are performed. A recent initiative in this direction has been taken by leading researchers in the field suggesting a standardized set of acoustic parameters (i.e. Eyben et al., 2016). By using them in the current thesis, and by suggesting detailed acoustic patterns for each emotion, I hope that the results presented here may be more easily replicable in future studies. However, because the (more or less) subtle changes conveying the emotional message of a voice need to be understood in relation to that person's normal voice, the acoustic parameters need to be standardized to be comparable with other speakers. In the study presented in this chapter, I opted to standardize the acoustic features from each actor's voice when they read the standard sentences in a "neutral" prosody. A potential problem with this approach is that different studies may have different methods to record the speaker's normal voice. In the current study, the actors were given a neutral scenario and were instructed to imagine being engaged in doing some activity they do on an everyday basis. However, by reading the sentence rather than speaking freely, the actors may have manipulated their voices in ways that do not represent their normal voice.

Even though some studies have standardized the acoustic features from the speakers voices when they were not expressing any emotion (e.g. Pell, Paulmann, Dara, Alasseri, & Kotz, 2009), a more common approach is to standardize across the different expressions included in the study. A benefit of this method is that any manipulations that the speakers make between expressions will be enhanced but the drawback is that the acoustic patterns produced will not generalize to other sets of expressions. Another approach is to use multilevel regression on the unstandardized parameters, but even so, the results of such analyses would have to be presented in a standardized format (as the mean-changes across the actors), which would yield similar problems as the two approaches mentioned above.

Although the acoustic patterns derived in the current study are based on one sample of actors and listeners, and that many degrees of researcher-freedom needs to be taken into account, I hope that this work will contribute to the field by presenting testable hypotheses of what the acoustic feature-patterns conveying emotions may look like. If these results are tested and compared with other speakers and listeners, replicated or not, they will allow for a more cumulative science within the field of emotional communication in the voice.

4. How much acoustic information do listeners need to infer emotions from speech and music?

Everyday experience tells us that we are quick to hear if a person is speaking in an emotional tone of voice. The same goes for music; the emotional expression is usually the first thing capturing our attention. The ability to quickly and accurately identify emotional expressions is essential to social interaction because it gives us information about the relationship between the expresser and their environment. Fast emotion recognition enables the perceiver to respond in appropriate ways, to detect and avoid negative outcomes, and to promote their personal goals (Keltner & Haidt, 1999).

The emotional message of both speech and music is conveyed to perceivers via acoustic cues that develop over time. Research suggest that different emotions are associated with relatively distinct patterns of acoustic-perceptual cues that may unfold at different rates in time (Banse & Scherer, 1996; Hammerschmidt & Jürgens, 2007; Laukka et al., 2016; Sauter, Eisner, Calder, & Scott, 2010). Some cues, such as intensity level or high-frequency energy, are available to perceivers almost instantaneously while others, such as F0 variability and speech rate/tempo, require acoustic information that develops over a longer time. Perception of the complete emotional message thus requires that perceivers attend to acoustic cues that vary over shorter and longer time periods. It is possible that some emotions are recognized faster than others and that this could be explained by the pattern of cues that convey the message for each emotion. Investigation of the time course of emotion recognition in speech and music is therefore an important step in our quest to understand how emotional information is conveyed from expresser to perceivers.

An indirect approach to study the time course of emotion recognition is to record electrophysiological responses to emotional stimuli with an electroencephalogram (EEG) during a mismatch negativity paradigm. In this paradigm, listeners are presented with a "prime" that is either congruent or incongruent with a "target". If the participant noticed an incongruence, a negativity is expected in the electrophysiological response at about 400 milliseconds (ms) after the target has been presented. When applied to the study of the time course of emotion recognition, short fragments of vocal emotion

expressions (primes) have been presented followed by either an emotionally congruent or an incongruent facial expression (targets). These studies show that participants react to incongruent targets when the prime is 400 ms but not when they are 200 ms, indicating that emotion recognition occurs somewhere in this range (Chen, Zhao, Jiang, & Yang, 2011; Paulmann & Kotz, 2008; Paulmann & Pell, 2010; Sauter & Eimer, 2010; Spreckelmeyer, Kutas, Urbach, Altenmüller, & Münte, 2009).

A more direct approach that allows for a more detailed description of the time course of emotion recognition is to adopt a method that was originally used to study the time course of word recognition, namely the gating paradigm (Grosjean, 1980). In this paradigm, listeners are presented with segments or "gates" of a word read aloud. Starting from the onset of the word, listeners may begin with hearing the first syllable or the first few ms, and then get to hear an increasingly long duration until the complete word is heard. For each gate, listeners are asked to identify the word, and then the identification point (IP) is defined as the point in time when the listener identifies the word correctly without changing their response when hearing a larger portion of the word. When applied to the study of the time course of emotion recognition, the same procedure is used except that listeners are asked to report the expressed emotion instead of a word.

Previous gating studies on vocal expressions

Only a few previous studies have applied the gating paradigm to study the time course of emotional expressions in speech. One of these studies compared the recognition time course of emotions (anger, sadness, and happiness) with what they called "attitudes" (obviousness, irony and doubt) (Grichkovstova, Lacheret, Morel, Beaucousin, & Tzourio-Mazoyer, 2007). Their results showed that anger and sadness had shorter IPs (range = 800 – 1600 ms) than happiness (range = 1000 ms – full sentence) and attitudes. The authors argued that expressions of anger and sadness might be identified earlier because they can be seen as more direct and involuntary while happiness, much like attitudes, might be used more deliberately by the speaker to color social interaction.

Another study reported that listeners had shorter IPs for neutral expressions (444 ms) compared with angry (723ms) and happy (802 ms) expressions (Cornew, Carver, & Love, 2009). In turn, Pell and Kotz (2011) reported that neutral (510 ms), fear (517 ms), sadness (576 ms), and anger (710 ms) had shorter IPs than happiness (977 ms) and disgust (1486 ms). A follow-up study hypothesised that the longer IPs for happiness and disgust could be explained by the fact that the acoustic cues conveying these emotions might not be available in the expressions until the end of the utterance

(Rigoulot, Wassiliwizky, & Pell, 2013). They tested this hypothesis by reversing the gating procedure so that the last gate of the expression was presented first, followed by the two last gates and so on. Results showed that the duration listeners needed to reach the IP indeed decreased for happiness (-173 ms) and disgust (-289 ms) but did not change as much for fear (+90 ms), anger (+36 ms), sadness (-39 ms), and neutral (-144).

Finally, the gating paradigm was used to compare the time course of emotion recognition for vocal expressions uttered by speakers from the listeners own language, compared to a foreign language group (Jiang, Paulmann, Robin, & Pell, 2015). Native speakers of English or Hindi were asked to identify emotion expressions performed by speakers from one or the other language groups. Results showed that emotion recognition occurred faster if the speaker and listener were native speakers of the same language. Across languages, the shortest IPs were reported for neutral (range of average IPs across listener and speaker groups = 490-542 ms) and anger (range = 447-621 ms), followed by fear (range = 591-738 ms) and sadness (range = 622-755 ms), with the longest IPs for happiness (range = 654-1634 ms).

Previous gating studies on musical expressions

Two studies used the gating paradigm to investigate how fast listeners judge music as moving or not moving. Results showed that listeners could identify subjectively moving pieces of music from excerpts as short as 250 ms, which was the shortest gate duration tested (Bigand, Filipic, & Lalitte, 2005; Filipic, Tillmann, & Bigand, 2010). Bigand and colleagues (2005) argued that the quick judgments could be explained by the fact that skilled musicians are trained to shape the acoustic cues of the very first note of a musical performance so that its acoustic characteristics prefigures the main mood of the piece.

Two gating studies have investigated the recognition time course of musical expressions of emotion. The first showed that listeners could differentiate between happy and sad expressions for musical excerpts that were 500 ms, the shortest gate included in the study (Peretz, Gagnon, & Bouchard, 1998). The other study divided expressions of happiness, sadness, threat, and peacefulness into gates comprised of one note for each gate (Vieillard et al., 2008). Results showed that happy expressions were recognized faster (average duration = 483 ms) than the other emotions: peacefulness (1261 ms), sadness (1446 ms) and threat (1737 ms)

Study 1: The Time Course of Emotion Recognition in Speech and Music

Results from the gating studies of vocal expressions reviewed above suggest that negative emotions such as anger, fear, and sadness are recognized faster than happiness (Cornew et al., 2009; Grichkovstova et al., 2007; Jiang et al., 2015; Pell & Kotz, 2011; Rigoulot et al., 2013). However, happiness was the only positive emotion included in these studies and thus, the advantage for negative expressions might have been caused by the fact that there were more negative than positive emotions to choose from. This is unfortunate because studies have shown that several positive emotions can be conveyed by the voice (Laukka et al., 2016; Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, & Abramson, 2009). It is therefore unclear whether the advantage for negative expressions would remain if other positive emotions were included in the forced choice-task.

In contrast, the one study investigating musical expressions suggest that happiness is recognized faster than the other expressions (Vieillard et al., 2008). The lack of studies that have investigated the recognition time course of musical expressions and the fact that they have used a very limited selection of emotions indicates that more research is needed for an adequate description of how listeners use acoustic information in emotional musical expressions. Also, in the two studies that have used the gating paradigm to study emotion expressions, they were identified already at the first gate suggesting that we do not know how fast musical expressions can be identified.

In this study we present two gating experiments that investigate the time course of emotion recognition from vocal- (Experiment 1) and musical expressions (Experiment 2). These two experiments have used a wider range of emotions and shorter gate-durations with smaller increments than previous studies. This allows for a more fine-grained description of how emotion recognition unfolds in time in speech and music. Also, to allow for comparison between speech and music, we have used the same gating procedure in both experiments. The study also expands on previous research by analysing acoustic parameters at the point of identification. This allowed us to investigate what acoustic cues convey the emotional message for very brief excerpts of vocal and musical emotion expressions. The study thus gives new insights into the investigation of how listeners can identify emotions so quickly, and it describes in detail how speech and music is recognized with limited acoustic information.

Methods

To select expressions with the highest possible recognition rate for each emotion, listeners were recruited to validate the vocal and musical expres-

sions. For vocal expressions, 20 listeners were asked to identify expressions (from four of the actors from the recording procedure described in Chapter 3) in a forced choice-task with as many alternatives as there were intended emotions (i.e. 14). On the basis that the intended expressions were identified by at least 60 % of the listeners, we selected four recordings of anger, fear, happiness, interest, lust, neutral, sadness, and serenity, and three expressions of tenderness and relief. Vocal expressions of pride, shame, disgust and contempt were excluded because too few listeners could identify the intended emotions. For musical expressions, 28 listeners rated 291 expressions from three professional musicians (violin, viola, and cello) with the same forced choice-procedure. On the basis that the intended expressions were identified by at least 30 % of the listeners, we selected four recordings of happiness, interest, serenity, fear, sadness, anger and neutral, and three recordings of tenderness. Musical expressions of lust, relief, pride, shame, disgust and contempt were excluded. Thus, 38 vocal expressions of 10 emotions and 32 musical expressions of 8 emotions were selected for the study. These expressions were sliced into 11 gates. The shortest gates had 50 ms increments (50, 100, 150, 200, and 250 ms), the next five gates had 250 ms increments (500, 750, 1000, 1250, and 1500 ms) and the duration of the last gate was 2500 ms.

In two experiments, groups of listeners (n = 31 and 32 respectively) were presented with either vocal- or musical expressions. Starting with the shortest gate-duration, listeners were instructed to select the emotion label they thought best fitted each expression in a forced choice-task with as many alternatives as there were intended emotions (i.e. 10 and 8 respectively). When all expressions of the shortest gate-duration had been presented, expressions of the second shortest gate-duration were presented and so on until the complete version of the expressions were presented. The order of presentation was randomly unique for each gate-duration. In both experiments, results are presented for each emotion in terms of a) the shortest gates at which a larger proportion of the listeners than would be expected by chance could identify the intended emotion, and b) as the identification points (IPs) which represent the amount of acoustic information that listeners require to recognize each emotion in a stable manner. In addition, results are presented as c) the most prevalent misclassifications for each gate duration and emotion. Study 1 also presents detailed information about acoustic cue patterns for emotions, for both speech and music, but these results are only briefly discussed in this summary.

Figure 5. Recognition accuracy for vocal expressions as a function of gate duration for each emotion in separate panels.

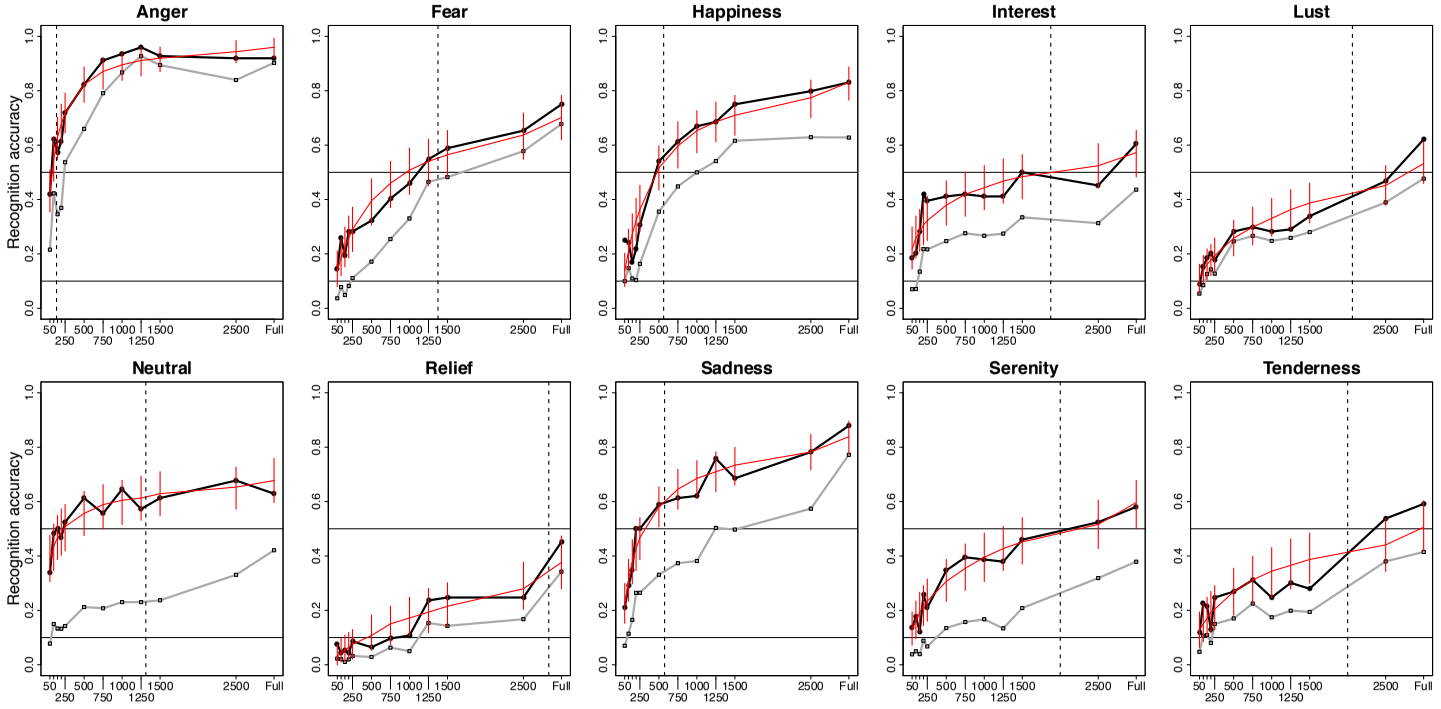
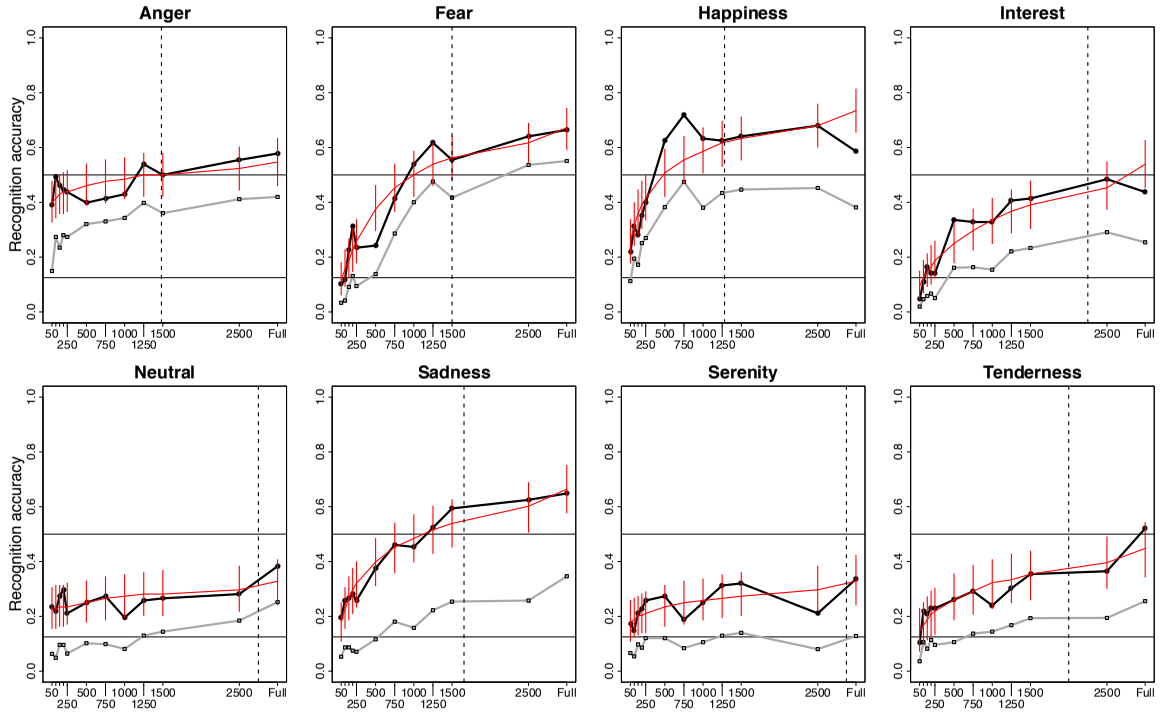


Figure 6. Recognition accuracy for musical expressions as a function of gate duration for each emotion in separate panels.



Results and Discussion

For each gate-duration and emotion, Figure 5 and Figure 6 present results from the generalized linear mixed models (GLMMs) (red lines and error bars) and the proportion of listeners who selected the intended emotion (black lines), and scores of unbiased hitrate (grey lines) for vocal- and musical expressions respectively (see Study 1 for a detailed description of the statistical analyses). The error bars illustrate the uncertainty of the estimated recognition accuracy as 95% credible intervals. The identification points (IPs) for each emotion category are presented with thick vertical dashed lines.

Generally, these figures show that emotion recognition can occur with very limited acoustic information, presumably from low-level acoustic cues, such as F0, intensity, or energy distribution across the frequency spectra. These acoustic cues seem to be available to the listener already at the very onset of a vocal or musical expression. They also show that recognition accuracy continues to improve with increasing gate duration. This reflects the increased amount of relevant acoustic information that becomes available to the listener as the expression unfolds, either via psycho-acoustic cues that need longer time to develop such as F0 contour, and speech rate/tempo, or via low-level cues that become clearer as more information is available. Anger, fear, happiness and sadness were the best recognized emotions at full-length for both modalities. Even though there were fewer musical expressions to choose from in the forced choice-task than for vocal expressions (i.e. 8 vs. 10), the overall recognition rate for the complete musical expressions was lower than that of the vocal expressions (M = .69/.52, SD = .20/.20).

For vocal expressions, results show that anger, interest, neutral, and sadness were recognized by a larger proportion of listeners than would be expected by chance at the shortest gate-duration (50 ms). Happiness, fear, lust, and serenity all reached above chance level at the second shortest gate (100 ms), and tenderness at the third shortest gate (150 ms). Relief reached this level much later at 1000 ms. **Anger** reached the point where a majority of the listeners selected the intended emotion at 100 ms and had an average IP of 138 ms and was thus the quickest vocal expression to be recognized in a more stable manner. Even though listeners had to choose from a wider range of emotions in the current study, anger was recognized much faster compared with previous studies (Cornew et al., 2009; Grichkovstova et al., 2007; Jiang et al., 2015; Pell & Kotz, 2011). This might have to do with methodological differences such that we used shorter gate durations with smaller increments but also that we chose to preserve the original sound level-difference between emotion expressions by not normalizing the recordings as some previous studies had done (e.g. Pell & Kotz, 2011; Rigoulot et al., 2013). **Happiness** and **sadness** were also quickly recognized reaching ma-

jority at 500 ms and the IPs at 562 and 575 ms respectively. Similar to anger, happiness was recognized somewhat faster than in previous studies. This seems to go against the suggestion that negative emotions are recognized faster than positive ones (e.g. Pell & Kotz, 2011), but it should be noted that happiness was the only exception to this suggestion as all other positive emotions were recognized later than the negative ones. Similar to happiness and sadness, **Neutral** expressions also reached a majority at 500 ms but had a longer IP at 1312 ms, which is a bit later than in previous studies. The fact that the current study used more (and different) response options in the forced choice-task than previous studies may have made it more difficult for the listeners to "guess" the correct answer by exclusion of other emotion alternatives that may be confused with neutral expressions. Accordingly, an inspection of the misclassification patterns in Table 4 suggests that neutral was often misidentified as serenity, which is a low arousal state that was not included in any of the previous studies. Also, recognition of neutral expressions should be interpreted with caution because the unbiased hitrate suggest that listeners guessed more frequently on neutral and serenity than any of the other emotions. Because listeners in a gating study have to make decisions under very uncertain conditions, they may be more prone to guess on neutral, a behaviour that will inflate recognition rates for neutral expressions. This pattern was evident in the results of the current study and probably in previous studies. **Fear** reached a majority a bit later than neutral at 1250 ms but had about the same IP at 1375 ms. **Serenity** and **tenderness** were slower to be recognized and did not reach a majority until the second longest gate, both at 2500 ms and the IPs at 2000 ms. Even though **interest** and **lust** did not reach a majority until the complete utterance was heard they had about the same IPs as serenity and tenderness at 1875 ms and 2062 ms respectively. **Relief** did not reach a majority even for the complete utterance and had the longest IP at 2833 ms and was thus the slowest of the included emotions to be recognized from vocal expressions.

For musical expressions, results showed that anger, happiness, and neutral were recognized by a larger proportion of listeners than would be expected by chance at the shortest gate-duration (50 ms). Sadness reached this level at the second shortest gate (100 ms) and serenity on the third shortest gate (150 ms). Fear and tenderness reached above chance level from expressions with 200 ms and 250 ms durations respectively. Interest was the slowest to reach above chance at 500 ms. **Happiness** reached a majority of correct responses at 500 ms and had an average IP of 1281 ms and was thus, on average, the quickest of the included musical expressions to be recognized reliably. Although the IP was much longer in the current than in the previous gating study on emotion recognition, this finding seems to be in line with the

Table 4. Most common confusions for each gate and emotion for vocal expressions. Bold font indicates that a larger proportion of the participants selected this emotion label than the intended emotion label.

Gate (ms)	Anger	Fear	Happiness	Interest	Lust	Neutral	Relief	Sadness	Serenity	Tenderness
50	hap 0.21	**sad 0.20**	fea 0.22	**neu 0.19**	**sad 0.26**	ser 0.16	**neu 0.34**	**ser 0.23**	**neu 0.33**	**neu 0.26**
100	fea 0.14	sad 0.24	fea 0.23	**neu 0.24**	**ser 0.23**	ser 0.13	**neu 0.30**	fea/neu 0.24	**neu 0.36**	**neu 0.23**
150	fea 0.16	**sad 0.29**	**fea 0.29**	neu 0.24	**ser 0.20**	ser 0.15	**neu 0.44**	fea 0.23	**neu 0.42**	**neu 0.24**
200	fea 0.15	sad 0.17	**fea 0.29**	neu 0.22	ser 0.19	ser 0.19	**neu 0.33**	fea 0.19	**neu 0.34**	**neu 0.30**
250	hap 0.10	**sad 0.28**	fea 0.24	neu 0.24	**ser 0.22**	ser 0.15	**neu 0.37**	fea 0.17	**neu 0.45**	neu 0.20
500	fea 0.06	sad 0.31	fea 0.11	neu 0.22	ser 0.27	ser 0.18	**neu 0.40**	neu 0.15	neu 0.30	neu 0.17
750	hap 0.03	sad 0.31	fea 0.12	neu 0.24	ser 0.23	ser 0.29	**neu 0.45**	neu 0.13	neu 0.31	neu 0.22
1000	fea/int 0.02	sad 0.33	int 0.10	neu 0.27	ser 0.27	ser 0.23	**neu 0.48**	neu 0.14	neu 0.38	neu 0.20
1250	int 0.02	sad 0.29	int 0.11	neu 0.27	ser 0.22	ser 0.27	**neu 0.41**	ser 0.08	neu 0.35	ser 0.26
1500	int 0.04	sad 0.23	int 0.10	neu 0.24	ser 0.23	ser 0.17	**neu 0.41**	neu 0.11	neu 0.31	ser 0.22
2500	hap/int/neu 0.02	sad 0.23	int 0.11	neu 0.19	ser 0.15	ser 0.14	**neu 0.33**	neu 0.08	neu 0.28	ser 0.15
Full	hap 0.02	sad 0.14	int 0.08	neu 0.19	ser 0.14	ser 0.20	hap 0.18	fea 0.08	ten 0.15	ser 0.16

Table 5. Most common confusions for each gate and emotion for musical expressions. Bold font indicates that a larger proportion of the participants selected this emotion label than the intended emotion label.

Gate (ms)	Anger	Fear	Happiness	Interest	Neutral	Sadness	Serenity	Tenderness
50	fea 0.19	**neu 0.28**	**fea 0.23**	**ang 0.34**	**sad 0.28**	**neu 0.26**	**neu 0.27**	**hap 0.20**
100	fea 0.20	**neu 0.18**	int 0.16	**ang 0.27**	**ser 0.28**	**neu 0.28**	**neu 0.31**	sad/ser 0.19
150	fea 0.21	neu 0.18	int 0.22	**ang 0.24**	**sad 0.32**	**neu 0.32**	**neu/sad 0.24**	sad 0.19
200	fea 0.24	sad 0.23	int 0.16	**neu 0.20**	**sad 0.31**	**neu 0.30**	**neu 0.27**	neu 0.18
250	fea 0.20	ang 0.15	int 0.20	**ang 0.21**	**sad 0.33**	**neu 0.29**	sad 0.24	**neu 0.25**
500	fea 0.20	sad 0.20	int 0.19	**hap 0.36**	**sad 0.38**	ser 0.19	**sad 0.28**	sad 0.25
750	int 0.13	int 0.21	int 0.20	**hap 0.38**	**sad 0.29**	neu 0.20	**sad 0.34**	sad 0.28
1000	int 0.24	ang/int 0.13	int 0.25	**hap 0.41**	**sad 0.34**	neu 0.19	**sad 0.39**	**sad 0.43**
1250	int 0.21	ang 0.19	int 0.24	hap 0.35	**ser 0.41**	ser 0.18	**sad 0.43**	**sad 0.33**
1500	int 0.17	ang 0.20	int 0.22	hap 0.34	**ser 0.34**	ten 0.13	**sad 0.41**	sad 0.32
2500	fea/int 0.13	ang 0.23	int 0.22	hap 0.37	**ser 0.32**	ser/ten 0.15	**sad 0.51**	sad 0.30
Full	int 0.12	ang 0.27	int 0.28	hap 0.37	ser 0.31	ser 0.16	**sad 0.41**	ser 0.26

notion that happiness is the fastest musical expression to be recognized (Vieillard et al., 2008). The longer IP for happy expressions compared with the previous study might, again, be due to the inclusion of other emotions that were not present in the previous study. The misclassification patterns in Table 5 suggests that this might be the case because happy expressions were commonly confused with interest which was not included in the previous study. **Fear** and **anger** first reached a majority at 1000, and 1250 ms and the IPs at 1500 and 1481 ms respectively (although anger dropped temporarily below majority to .5 at 1500 ms). It should be noted, however, that two of the four expressions of anger were recognized by a majority faster and had shorter IPs than any of the expressions of happiness or fear (see supplementary material of Study 1). This might indicate that anger is generally recognized slower than happiness and fear but it also shows that at least some expressions of anger can be recognized faster than all other musical expressions included in the study. Also, two of the expressions of fear had considerably higher recognition rates for gates above 750 ms than the two others, suggesting a similar case for musical expressions of fear. **Sadness** reached a majority at the same time as anger, at 1250 ms, and had a rather similar IP as both fear and anger at 1656 ms. However, it should be noted that the deviations between the observed rates of recognition and the unbiased hitrate suggests that listeners guessed more frequently on sadness than other emotions. This may have inflated recognition rates for sadness. **Tenderness** reached a majority when the complete expressions were heard and had an IP at 2000 ms. **Interest**, **neutral**, and **serenity** did not reach a majority even for the complete expressions and had the longest IPs of the included emotions at 2250, 2750, and 2875 ms respectively.

Comparisons of the time course of emotion recognition for vocal and musical expressions revealed that anger, happiness, and sadness were recognized better than chance in both modalities already at the shortest or second to shortest gates. Although several other vocal expressions reached above chance level faster than musical expressions, this suggests that emotion discrimination is very fast in both modalities. Thus, low-level physical characteristics that are immediately accessible to the listener might be sufficient for emotion discrimination in both speech and music.

Stable recognition, as indicated by the IPs, generally occurred faster for vocal (M = 1424, SD = 979) than for musical (M = 1973, SD = 960) expressions. This observation, together with the generally low recognition rates for music (both in the selection study and in the experiment), may suggest that emotions are not expressed as directly in music as in speech. Instead, although the general expressivity of music is undeniable, music may operate on a more symbolic level, making discrimination of discrete emotion expressions more difficult than in speech. Neutral expressions deviated most between the modalities, being reliably recognized among the fastest for vocal

expressions but second to last for musical expressions. This may reflect the fact that musicians seldom have the intention to express neutrality in their music, and that the quality of a musical performance is typically judged by its expressivity. Therefore, listeners may have been less prone to interpret the musical expressions as neutral. On the other hand, as shown in the results from the experiment with vocal expressions, listeners were generally more prone to guess neutral than other expressions. Presumably, neutral was the go-to selection when listeners were uncertain about the vocal expressions whereas sadness was the go-to selection when they were uncertain about the musical expressions.

The IPs together with the point in time when a majority of listeners could identify the intended emotions reveal that anger, happiness, fear, and sadness were recognized the fastest in both modalities (although in slightly different order). This, together with the relatively late recognition of serenity, tenderness, and interest for both modalities implies that the time course of emotion recognition in speech and music might rely on similar psycho-acoustic cues, some that are available very early and some that need longer time to develop. The full article (Study 1) also presents detailed results concerning acoustic patterns for each emotion for both speech and music (see Table III and Table V in Study 1), measured at both the IP and at the shortest gate with above chance recognition accuracy. These cue patterns were positively correlated across speech and music, which suggests that speakers and musicians used acoustic cues in a similar way to express emotions. These results are in line with the proposition that communication of emotion in speech and music is based on similar patterns of acoustic cues (Juslin & Laukka, 2003; Scherer, 1995).

Study 1 used shorter gate durations and smaller increments to allow for a more fine-grained temporal resolution, and also included a wider set of expressions of both positive and negative emotions than previous studies. Despite this, two limitations can be attributed to the expressions used in the study. First, the small number of actors and musicians who produced the expressions makes it difficult to draw general conclusions about the recognition trajectories for different emotions. Results from future studies using the same methodology and set of emotions expressed by other actors and musicians will thus have to be evaluated to determine if the results described here will generalize to other vocal and musical expressions. Second, even though we tried to minimize the variability in recognition rates for the full-length expressions by selecting only the most well recognized expressions of each emotion category in the two validation studies, there was still a large variability in general recognition rates both within and between emotion categories. The fact that some emotions and of course also some individual expressions are recognized more easily than others is not controversial. However, this circumstance makes it difficult to draw conclusions about the time

course of emotion recognition because the trajectories of different expressions/emotions will have different endpoints (i.e. the maximum recognition rate for the full-length expression). Therefore, the trajectories of different expressions (even within the same emotion category) may have different slopes and therefore different points in time when they reach above the level of chance, the level of majority, and the IP. The remedy for this limitation would, again, be to use more expressions so that the mean trajectories for each emotion category would be less influenced by single expressions. In the present study however, this was not feasible because the experiment would be too long and exhausting for the listeners and because we only had recordings of three musicians and wanted to keep the two experiments as similar as possible.

Conclusions

To summarize, with a variety of positive and negative emotions, the gating paradigm was used to study the time course of emotion recognition from both speech prosody (experiment 1) and music performances (experiment 2). The two experiments showed that anger, happiness, sadness and neutral were recognized by a larger proportion of listeners than would be expected by chance already from the shortest or second shortest gates (≤ 100 ms). In both modalities, most of the other emotions were also recognized from very brief excerpts, at gate durations of 250 ms or shorter. More stable recognition of emotions, as indicated by the IPs and when a majority of listeners perceived the intended emotion, again suggested that anger, happiness and sadness were recognized with the least amount of acoustic information in both modalities. Positive emotions such as interest, relief, tenderness, lust, and serenity required more acoustic information to reach stable recognition.

As argued by previous gating studies (e.g. Grichkovstova et al., 2007; Jiang et al., 2015) one may interpret these results as that negative emotions are generally recognized faster than positive ones. This conclusion would also be in line with evolutionary arguments suggesting that individuals need to detect and avoid potentially harmful events faster than potentially beneficial ones (e.g. Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). However, though the current study included more positive emotions than previous studies and because one of them, happiness, was recognized among the fastest, one could also interpret the results in terms of basic emotions rather than positive-negative valence. From this perspective, it could be argued that the findings supports the notion that a few emotion categories have largely automatic and biologically driven responses, including expressions, evolved as adaptations to specifically important events (e.g. Tracy & Randles, 2011).

Basic emotion theory predicts that anger, fear, happiness, and sadness should be recognized at shorter gate durations than the other affective states

because they are thought to have more distinct expressions. It could be argued that the current results support this notion. However, although fear was generally among the fastest recognized emotions, the expressions produced by the actors and musicians in the current study varied a lot in recognition accuracy between expressions. This may, of course, be explained by the performances of these specific actors and musicians, but it could also be interpreted as an indication that fear does not (at least always) have as clear expressions as the other emotions called basic. For example, studies on emotion recognition often report that fear is commonly confused with both anger and sadness (Juslin & Laukka, 2003). Also, basic emotion theory predicts that disgust should be coupled with clear expressions (e.g. Ekman, 1992). Although disgust is usually well recognized in facial expressions, these findings have been difficult to replicate for speech prosody and music, suggesting that disgust may not be as distinguishable for auditory communication as for facial cues (Juslin & Laukka, 2003; Pell et al., 2009). In the current study, neither actors nor musicians were able to produce expressions of disgust that listeners could recognize reliably enough to pass the validation study. Again, this could be an artefact caused by the performances but could also be interpreted as that disgust is not communicated well by prosody or music.

The experiment investigating musical expressions presents novel findings that several emotions can be recognized from very brief excerpts of music. This demonstrates that communication of emotions in music, similar to vocal expressions, is at least partly based on low-level acoustic cues such as F0, intensity and spectral balance, available to listeners almost instantaneously. However, recognition of musical expressions was generally lower and stable recognition occurred later than those of vocal expressions. This suggests that, although low-level acoustic cues are used to communicate emotions in music, acoustic cues that need some time to develop such as F0 variability and tempo may be especially important for musical expressions. Accordingly, perception of melody and tonality in a musical phrase requires that listeners attend to pitch-relations between consecutive notes (Parncutt, 2014) and such relations are obviously not available in very brief excerpts.

Although differences were observed, the many similarities between speech and music could be interpreted as support for the notion that these modes of communication have a shared code of expression (Juslin & Laukka, 2003; Scherer, 1995). Both trajectories for specific emotions and the acoustic patterns observed when the expressions were recognized showed clear similarities between modalities. These observations also support the theories suggesting that speech and music may have evolved from a common proto-language that our ancestors may have used to facilitate social living by means of communication (Fitch, 2006; Thompson, Marin, & Stewart, 2012).

By showing that listeners can differentiate and recognize emotions from very brief excerpts of vocal and musical expressions, this study demonstrates the importance of the acoustic channel for emotional communication. Further exploration of the time course of emotion recognition from these two modes of expression will potentially broaden our understanding of the mechanisms and processes underlying human communication of emotional information to other individuals.

5. What emotions can listeners infer from non-linguistic vocalizations?

The concept of "non-linguistic" vocalizations is intended to capture all the sounds that humans make when we are not speaking. These sounds may be produced by sighs, yawnings, crying, hums, grunts and laughter, and they are thought to convey meaningful information about the internal state of the person making them. Under conditions of strong emotional arousal, both humans and other animals produce "affect bursts" that have many similarities across species (Briefer, 2012; Owren et al., 2011). Darwin (1872) used this observation when he first made his claim that humans were evolved from ancestors common to other animals.

Darwin argued that affect burst might have been evolutionary adaptations to specific circumstances and that they, in some animals, also had a communicative purpose. For example, he argued that the behaviour of blowing out air through the mouth and nostrils when presented to foul food did not only serve the purpose of clearing the respiratory passage from the bad odour but also that the sounds produced could be interpreted by other individuals. These sounds would communicate to others that the individual felt disgust toward the object, and thus that it was not edible (Darwin, 1872). Affect burst can thus be thought of as a specific kind of non-linguistic vocalizations reflecting the emotional state of the individual. Affect bursts usually take the form of single or repeated sounds that, intentional or not, convey information about the emotional state of the individual to an observer.

Because affective burst likely predates language as a channel of communication, it has been argued that these signals have played an important role in the development of speech and language in the human species (Scheiner & Fischer, 2011). Once the communicative function of an affect burst has been established, for example that you should stay away from food that others responded to by blowing out air through the nose, expressing and interpreting them could clearly have a survival value. Non-human primates have vocalizations to signal predators differentiating leopards, eagles, and snakes (Wheeler & Fischer, 2012), and specific signals for, food, affiliation, care, sex, and aggression (Snowdon, 2002). In humans, such proto-language may then have evolved into more complex sequences of sounds, gradually acquiring syntax and melody-like intonations, and eventually developed into speech (Mithen et al., 2006; Scheiner & Fischer, 2011; Wallin et al., 2001).

Semi-linguistic interjections such as “yuck” or “ouch”, that are common in many languages, have also been interpreted as support for the close connection between language and non-linguistic vocalizations (Scherer, 2013).

Compared with prosodic expressions in speech, non-linguistic expressions tend to be better recognized, especially for fear and disgust (e.g. Hawk, Van Kleef, Fischer, & Van Der Schalk, 2009; Scott et al., 1997) perhaps because this type of vocalizations, rather than speech, may be more likely to occur when a person is inflicted by these emotions. Thus, non-linguistic vocalizations may have a more direct link to the physiological changes elicited by the emotional episode and less by the contextual demands of the situation compared with speech. Also, speech is restricted by linguistic requirements that force the speaker to control the vocal apparatus with coordinated and precise articulations. Non-linguistic expressions do not require such precision and are thus less restricted in how emotions may be expressed. Non-linguistic vocalizations may therefore allow expressers to manipulate characteristics in their voice more freely than in speech. For example, support of this notion comes from studies comparing laughter versus speech, showing that laughter allows for larger pitch variations (Bachorowski, Smoski, & Owren, 2001). The possibility to express emotions with a wider range of vocal manipulations will thus influence the acoustic content allowing for clearer distinctions between different types of expressions.

Study 2: Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations

The aim of this study was to investigate the limits of what type of emotional information that listeners can perceive in a voice. To this end, the current study used a wider range of emotional non-linguistic expressions than previous studies and investigated how these expressions were recognized in a cross-cultural setting.

Although much research has been devoted to study how well emotions are communicated across languages and cultures via nonverbal aspects of speech (e.g. Barrett & Bryant, 2008; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009; Scherer, Banse, & Wallbott, 2001; Thompson & Balkwill, 2006; Van Bezooijen, Otto, & Heenan, 1983), relatively few studies have investigated non-linguistic emotion expressions in a cross-cultural setting.

When the current study was published, there were only two studies that had examined cross-cultural recognition of nonlinguistic expressions. Sauter, Eisner, Ekman, and Scott (2010) investigated how well nine emotions expressed by native British English individuals and individuals from a culturally isolated village in Namibia could be recognized across cultures. Their results suggested that basic emotions could be recognized across cultures but that positive emotions reached accuracy above chance mainly for within-

culture expressions. The other study investigated cross-cultural ratings of valence and arousal of nonlinguistic vocalizations expressing basic emotions across Canadian and Japanese expressers and listeners (Koeda et al., 2013). Results suggested both differences and similarities in how positive and negative expressions were rated in terms of these dimensions. Previous research on cross-cultural recognition of non-linguistic emotion expressions was thus sparse but provided initial findings of both similarities and differences across cultures. These results are in line with findings related to emotional expressions in speech suggesting that prosodic expressions can be recognized across cultures but that communication is more accurate when judges rate expressions from their own culture compared with unfamiliar cultures (Elfenbein & Ambady, 2002; Juslin & Laukka, 2003).

Methods

Professional actors from India, Kenya, Singapore, and USA were instructed to produce non-linguistic vocalizations expressing nine positive emotions (affection, amusement, happiness, interest, sexual lust, peacefulness/serenity, pride, relief, and positive surprise) and nine negative emotions (anger, contempt, disgust, distress/pain, fear, guilt, sadness, shame, and negative surprise) with a moderately intense arousal level. The actors could choose any kind of vocalization they thought fit for the emotion they were trying to express. Therefore, the type of vocalizations produced could vary between emotions for the same actor, or they could be the same for the same emotion for different actors. Examples of the kind of vocalizations the actors produced were breathing sounds, crying, hums, grunts, laughter, shrieks, and sighs. The actors were instructed not to use any words but that they were allowed to use non-linguistic interjections such as “ah,” “er,” “hm,” and “oh” as long as the interjections did not convey conventional semantic meaning such as “yuck” or “ouch”. In an initial screening of the recordings, borderline words and interjections with possibly semantic meaning were excluded.

Similar to the recording procedure described in Chapter 3, actors were provided with definitions and scenarios that are commonly associated with each emotion (the definitions and scenarios are presented in Chapter 2). To help the actors produce as reliable expressions as they could, they were also instructed to try to induce the emotional state by remembering a self-experienced situation similar to that described in the scenario.

The number of vocalizations per emotion and the number of portrayed emotions varied between actors but there were approximately equally many portrayals of each emotion from female and male actors and from each culture. The final material consisted of 109 recordings by Indian actors, 99 by Kenyan, 92 by Singaporean and 127 by American actors, resulting in a total number of 427 recordings. All recordings were normalized to have the same

peak amplitude. The normalization was made because there were large loudness differences both between recordings and recording sessions and because these differences would have caused too much loudness variation in the listening experiments (ranging from disturbingly loud to inaudibly quiet).

The positive and negative expressions were divided into two sets with 213 and 214 recordings respectively. In two experiments, one with the set of positive and one with the set of negative emotions, 29 and 28 Swedish listeners judged the expressions in a forced choice task with the emotion terms corresponding to the intended expressions in each experiment. Listeners were provided with the same definitions and scenarios as the actors in a similar manner as described in Chapter 3.

Results and discussion

Cross-cultural recognition accuracy and misclassifications for positive and negative non-linguistic vocalizations are presented in Table 6 and 7 respectively. The overall recognition accuracies in the two experiments were 39% for positive emotions and 45% for negative emotions.

Positive emotions that were seldom misclassified as any specific emotion and thus had the highest recognition accuracies were relief (70%), lust (45%), interest (44%), serenity (43%), and positive surprise (42%). Happiness (36%) and amusement (32%) were often confused with each other reaching almost 30% misclassification rates. This suggests that the listeners could not differentiate these vocalizations. However, given that these expressions are conceptually similar, this finding is not surprising. Combining them into a happiness/amusement category yielded an accuracy of 60%. Pride (22%) and affection (20%) received the lowest recognition rates. Both of these expressions were most frequently misclassified as interest.

Negative emotions that were seldom misclassified as any specific emotion and thus had the highest recognition accuracies were disgust (63%), anger (57%), fear (57%), sadness (56%), and negative surprise (53%). Contempt (44%) was most commonly confused with negative surprise. Distress (33%) was often misclassified as fear, sadness, guilt, or shame. This is not surprising given that distress may be associated with threatening or mourning-situations, or with feelings of guilt or shame. Therefore, some actors may have been leaning more towards fear and other leaning more toward sadness, guilt or shame when expressing the emotions. However the case, results suggest that vocalizations of distress may be closely related to the vocalizations of fear, sadness, guilt, and shame.

Table 6. Recognition rates and confusion patterns for non-linguistic vocalizations of nine positive emotions from four cultures.

		Intended emotion								
Judgment	**Culture**	**Affection**	**Amusement**	**Happiness**	**Interest**	**Lust**	**Pride**	**Relief**	**Serenity**	**Surprise (positive)**
Affection	India	**21***								
	Kenya	**24***			20	11	14		13	
	Singapore	**18**							10	
	USA	**16**	12			10				
Amusement	India		**16**	27			10			10
	Kenya	21	**45***	34						12
	Singapore		**36***	30					10	11
	USA		**34***	23						13
Happiness	India		18	**35***						15
	Kenya	10	40	**40***			18			25
	Singapore		23	**29***						14
	USA		17	**39***						19
Interest	India	21			**51***		29			14
	Kenya	18			**35***	10	18		10	
	Singapore	24			**49***	12	39			
	USA				**43***		25			
Lust	India	13				**61***			19	
	Kenya					**26***		13		
	Singapore	16				**41***			13	10
	USA	19				**48***		10	16	12
Pride	India	10		15			**19***			
	Kenya						**26***			
	Singapore			13			**10**			
	USA		16	10			**33***			
Relief	India							**67***	21	
	Kenya					13		**68***	26	
	Singapore			14		14	20	**75***		
	USA	24						**69***	30	14
Serenity	India					18		19	**48***	
	Kenya	11							**25***	
	Singapore	17				14		12	**49***	
	USA	30			7	24		14	**46***	
Surprise (positive)	India		43	13	29		13			**44***
	Kenya			10	20	10				**46***
	Singapore		17		19		14			**47***
	USA			12	31		18			**33***

Note: The recognition rates (percentage accuracy) for which the expression portrayed is the same as the expression judged are shown in the diagonal cells (marked in bold typeface). Asterisks denote recognition rates higher than what would be expected by chance guessing (11%), as indicated by binomial tests (ps < 0.05, Bonferroni corrected; ps < 0.001, uncorrected). Blank cells indicate misclassification rates of less than 10%.

Table 7. Recognition rates and confusion patterns for non-linguistic vocalizations of nine negative emotions from four cultures.

		Intended emotion								
Judgment	**Culture**	**Anger**	**Contempt**	**Disgust**	**Distress**	**Fear**	**Guilt**	**Sadness**	**Shame**	**Surprise (negative)**
Anger	India	**42***	14							
	Kenya	**38***	10	12						13
	Singapore	**46***								
	USA	**88***								
Contempt	India	20	**52***				18		16	
	Kenya	30	**29***	12					12	11
	Singapore		**57***		20					
	USA		**42***							
Disgust	India			**40***					12	
	Kenya		15	**58***					11	
	Singapore			**81***	12					
	USA			**67***						13
Distress	India			19	**52***	12		11		
	Kenya		12	12	**52***	12	26	16	17	22
	Singapore	16			**15**	25	10	15	24	
	USA			12	**34***	19	23		29	16
Fear	India					**60***		11		
	Kenya					**60***		16		11
	Singapore				24	**55***		15	14	
	USA				27	**58***		15		
Guilt	India						**19***		18	
	Kenya						**14**		17	
	Singapore						**29***	14	15	
	USA						**23***		15	
Sadness	India				21			**65***		
	Kenya				21			**58***		10
	Singapore							**27***		
	USA				14	12		**61***		
Shame	India						17		**22***	
	Kenya						22		**21***	
	Singapore						28	13	**16**	
	USA						26		**24***	
Surprise (negative)	India	11	18	12		11	30		20	**80***
	Kenya		18			11	18		11	**18**
	Singapore	18	26				21		17	**68***
	USA		25				12		13	**42***

Note. The recognition rates (percentage accuracy) for which the expression portrayed is the same as the expression judged are shown in the diagonal cells (marked in bold typeface). Asterisks denote recognition rates higher than what would be expected by chance guessing (11%), as indicated by binomial tests (ps < 0.05, Bonferroni corrected; ps < 0.001, uncorrected). Blank cells indicate misclassification rates of less than 10%.

In the same way as happiness and amusement are conceptually similar emotions and therefore often confused, shame (21%) and guilt (20%) were most commonly confused with each other among the negative emotions. With misclassification rates reaching almost the same level as their accuracy rates results suggest that listeners could not differentiate these vocalizations.

A combined shame/guilt category yielded an accuracy of 40% indicating that these vocalizations may be recognized as a negative self-conscious emotion category. However, they were also frequently misclassified as distress and negative surprise, instead indicating that such a self-conscious emotion category may not be differentiated from other negative vocalizations.

Looking at the recognition accuracies and misclassifications of different speaker cultures suggest similar patterns as for the joint recognitions and misclassifications, both for positive and negative emotions. Nevertheless, some deviations may be noted. The listeners performed poorly on the Indian actors expressions of amusement and instead classified them (besides happiness) as pride or surprise. The same goes for American actors' expressions of affection, which were commonly confused as amusement or lust, and Singaporean actors' expressions of pride commonly confused with happiness. Also, vocalizations intended to express distress by Indian and Kenyan actors were confused with sadness (and guilt for Kenyan actors) whereas Singaporean and American actors' expressions of distress were more commonly confused with fear.

However, these culture-specific deviations should be interpreted with caution because they may be caused by other factors than the culture of the speakers. One group, regardless of culture, may simply have been less successful in communicating certain emotions. The probability that the deviations were caused by arbitrary variations between speaker groups is further increased by the small number of expressions in each emotion by culture cell. To determine if the deviations were examples of poor communication or depending on culture-specific ways of communicating these emotions one would have to conduct a study with both within- and cross-cultural conditions.

Another caveat of the design was that we assessed positive and negative emotions in separate experiments, which increased listeners' probability of guessing the intended emotion label and thus may have inflated the recognition accuracy. This design was used in order to avoid fatigue and to keep the number of response alternatives in the forced choice task at a manageable level. With these considerations in mind, the separation of vocalizations to sets of positive and negative emotions was based on the assumption that positive emotions are more likely to be confused with other positive emotions and vice versa. However, based on the current study, we do not know if this assumption is justified or not.

Conclusions

The study aimed to explore the limits of listeners' ability to recognize emotions from nonverbal aspects of the voice. In summary, we assessed cross-cultural recognition rates of non-linguistic vocal expressions intended to communicate 18 emotions and showed that many of these expressions could be reliably perceived by listeners in a forced choice task. Because non-linguistic vocalizations are unbound by the syntax of speech and language they may represent a more direct channel of emotional communication (Briefer, 2012; Owren et al., 2011). Across the two experiments, results showed that listeners can recognize a wide range of both negative and positive emotions from non-linguistic vocalizations, even if the expresser is from another culture than the listener. Therefore, results confirm the notion that non-linguistic vocalizations are a rich and nuanced channel of information that humans can use to communicate discrete emotions (Simon-Thomas et al., 2009).

Further, the study established that the voice can communicate several positive emotions other than happiness. A large proportion of listeners recognized relief, lust, interest, serenity, and positive surprise without misclassifying them as other emotions. This suggests, for the first time in a cross-cultural setting, that these emotions can be communicated nonverbally in the voice. Expressions of negative emotions were also recognized in a stable manner. Disgust, anger, fear, sadness, negative surprise, and contempt could be communicated reliably. This study thus gives new insights into the capabilities of the voice as a carrier of emotional information.

Findings that emotions can be reliably communicated across cultures are usually interpreted as support for the notion that emotion expressions are based on signalling mechanisms related to specific events that have been important in our evolutionary past (Ekman, 1992). Thus, the current results could be interpreted as support for this notion.

When the current study was published, there was only one other study that had obtained recognition accuracies of non-linguistic vocal expressions in a cross-cultural setting (Sauter et al., 2010). They included two positive emotions in their study and results suggested that amusement, but not relief, could be reliably recognized by listeners. However, in accordance with the findings of the current study, more recent findings suggest that many emotions, including several positive ones, can be reliably recognized and differentiated both within and across cultures (Cordaro, Keltner, Tshering, Wangchuk, & Flynn, 2016; Cowen, Elfenbein, Laukka, & Keltner, 2018; Lima, Anikin, Monteiro, Scott, & Castro, 2018; Shiota et al., 2017)

The current study thus gave new insights into the capabilities of the voice as a carrier of emotional information that later studies have continued to build upon. It also highlighted that the generally low recognition rates of

some emotions in the previous literature perhaps could be remedied by the use of non-linguistic expressions rather than speech.

However, not all emotions included in the study could be clearly recognized. Only modest recognition rates were obtained for pride, affection, and distress, suggesting that these emotions may not have clear signals in the voice or at least that they are difficult for actors to portray. The finding that affection was not clearly recognized was surprising because other studies suggest high recognition rates for "tenderness" (e.g., Juslin & Laukka, 2003) and because common sense tells us that affection have clear nonverbal signals (the sound people make when seeing a cute pup). Amusement could not be separated from happiness suggesting that this emotion does not have a specific signal. Instead, it seems as if both happiness and amusement are signalled with the same type of nonverbal cues. Similarly, shame and guilt could not be separated from each other. Even when combined into a negative self-conscious category, they merely showed modest separation from other negative emotions. This result, together with previous within-cultural studies (Hawk et al., 2009; Simon-Thomas et al., 2009) indicates that self-conscious emotions may not have a specific signal other than the signal of a more general state of low aroused sadness. In natural settings, in which people also have contextual knowledge, one could speculate that this signal may be enough to separate shame and guilt from other types of negative emotions, and the same might be the case for closely related positive emotions such as happiness and amusement.

6. Can listeners infer appraisal dimensions across cultures from speech?

As reviewed in Chapter 2, emotions are commonly defined in terms of several components including cognitive appraisals, central and peripheral physiological responses, action tendencies, expressive and instrumental behavior, and subjective feelings (e.g., Moors et al., 2013). According to emotion theories focusing on cognitive appraisal, emotions are elicited and differentiated by a person's appraisal of the significance of an object, situation, or event. If the event is appraised as significant to the person's needs and goals it might initiate a response in the other components. This appraisal is thought to occur more or less automatically along a small number of theoretically postulated dimensions. The postulated dimensions differ among theories but most suggest that people use aspects related to novelty, intrinsic pleasantness, goal conduciveness, urgency, power, responsibility, and compatibility with norms when appraising a potentially emotion-eliciting event (e.g., Ellsworth & Scherer, 2003).

Influenced by the early work of Tomkins, Ekman, and Izard (Ekman et al., 1969; Izard, 1971; Tomkins, 1962), most research on vocal expressions, and emotion expressions in general, have investigated perception of emotion categories such as "happy" or "sad" (for a review, see Scherer et al., 2011). In these studies, participants are typically presented with recordings of vocal portrayals intended to express various emotion categories and are then asked to select an emotion label from a list of alternatives. Meta-analyses show that listeners in such experiments can identify emotion categories with accuracy rates well above what would be expected by chance, both within and across cultures (Elfenbein & Ambady, 2002; Juslin & Laukka, 2003).

However, the strong focus on perception of emotion categories has been criticized because of its poor ecological validity. Everyday experience of social interactions tells us that we very seldom think in terms of emotion categories. Although categorical inferences might be one way of figuring out an appropriate response to a person's emotional state, it is likely that people adapt their behaviour by judgments based on several other inferences too. Because listeners in most previous studies have been limited by the forced choice-task when expressing what they are able to perceive from an emotional voice, they might have been able to infer several other aspects of the speaker's emotional state, if only they had been asked. From a perspective where emotions are thought to consist of several components, it has also

been argued that categorical judgments may "lack clarity with regard to whether one, some, or all emotion components are perceived" (p. 47, Shuman, Clark-Polner, Meuleman, Sander, & Scherer, 2017). Accordingly, research suggests that listeners are able to perceive aspects of emotional voices that would not be expected if they only made categorical judgments or if judgments were based on valence and arousal alone (Banse & Scherer, 1996; Juslin & Laukka, 2001; Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, & Abramson, 2009).

When interacting with other people, we do not only make inferences about the person's emotional state, but we may also try to infer aspects of the situation that elicited that state. Such inferences may not only help us to adjust our responses appropriately, but may also contribute to the high accuracy rates observed in emotion-labelling experiments because listeners may use "reverse engineering" when performing such tasks. In other words, listeners may use inferences about the situation when they select an emotion label they think correspond to a speaker's emotional state. It might thus be argued that the categorical judgments captured in the studies using the forced choice-task could have been inferred by first thinking of the situation that might have elicited the emotion and then inferred the emotional state of the speaker.

Appraisal theories indeed propose that perception of a small number of appraisal dimensions inferred from vocal expressions may allow listeners to understand the emotional message both in terms of categorical emotion-labels *and* a wide palette of other, more subtle, affective states. Within this theoretical framework, Scherer (1988; 1986) proposed that vocal expressions may contain information about the cognitive appraisals of the speaker, and went on to hypothesise that this "should allow the listener to reconstruct the major features of the emotion-producing event in its effect on the speaker" (p. 94, Scherer, 1988). Laukka and Elfenbein (2012) designed a study to test this hypothesis. They instructed actors to portray a wide range of emotion-eliciting situations and then let listeners judge recordings of these enactments in terms of several appraisal dimensions. Results showed that listeners could infer many aspects of the enacted events such as how novel it was, how urgently the speaker had to react, if the event corresponded to the speakers goals and values, and if the speaker possessed power to influence the outcome of the event. Other studies suggest that such judgments are not limited to vocal expressions but could also be inferred from both enacted (Hess & Hareli, 2016) and synthetic facial expressions (Sergi, Fiorentini, Trznadel, & Scherer, 2016), as well as multimodal expressions (Shuman et al., 2017).

It has been argued that inferences of this kind give support for a symbolic function of emotion expressions. In a framework first presented by Bühler (1934, as cited in Scherer, 1988), emotional expressions are suggested to

have three functions; (1) a symptom of the expressers emotional state, (2) a signal to an observer, and (3) a symbol that might be viewed as an abstraction of the object or event that elicited the emotional response. The strong focus on the perception of emotion categories has led research to study emotion expressions mainly as symptoms or signals of the speaker's internal states and largely neglected the symbolic meaning of the expression.

Understanding the symbolic meaning of emotional expressions might even shed light on the development of human language. Human language is characterised by the fact that we use sounds to represent objects in a symbolic way. In other words, the meaning of a specific combination of vocalisations is not depending on its resemblance of the object. If it can be shown that emotion expressions have a symbolic meaning, it could be theorized that early humans used such expressions to communicate meaningful information about their environment.

Investigation of the symbolic meaning of emotional expressions may also contribute to our understanding of cross-cultural recognition of emotion expressions. In the categorical tradition, results from cross-cultural studies suggest that listeners perform better than chance when judging categorical emotion-expressions from unfamiliar cultures, but also that there is an in-group advantage where listeners usually perform better judging expressions from their own group verses expressions from another culture (Laukka et al., 2016; Sauter et al., 2010; Scherer, Banse, et al., 2001; Thompson & Balkwill, 2006; Van Bezooijen et al., 1983). Research suggests that this in-group advantage might result from a better match between expression and perception styles within cultures. Even though the general patterns of the expressions might be similar across cultures, speakers may emphasize certain aspects of an expression that listeners from the same culture will notice while listeners from another culture may not (Elfenbein, 2013; Laukka et al., 2016). Letting listeners judge expressions in terms of appraisal dimensions instead of categorical emotion labels might be especially beneficial in cross-cultural settings because the emotion labels might have slightly different meanings in different cultures and languages (Hess & Thibault, 2009). Also, because it seems reasonable that the symbolic meaning of an expression is phylogenetically older than language and emotion words, judgments of other aspects than emotion categories might be more consistent across cultures and languages than categorical perceptions. Appraisal scales may thus allow researchers to assess perceptions of more subtle aspects of an expression that do not fit well into traditional emotion categories and that may correspond better to everyday social interactions (Juslin, Laukka, & Bänziger, 2018; Laukka, Neiberg, Forsell, Karlsson, & Elenius, 2011)

Study 3: Emotion appraisal dimensions inferred from vocal expressions are consistent across cultures: A comparison between Australia and India

In Study 3, we present the first investigation of cross-cultural judgments of emotion-eliciting situations described in terms of appraisal dimensions from vocal emotion expressions. Instead of judging emotion categories, listeners were instructed to rate aspects of emotion-eliciting situations, described in terms of appraisal dimensions (i.e., novelty, intrinsic pleasantness, goal conduciveness, urgency, power, and norm compatibility). Such judgments represent the listeners' cognitive understanding of the characteristics of different situations that may elicit emotional responses in every day social interactions. Using a similar method as Laukka and Elfenbein (2012), we investigated how listeners from Australia and India perceive appraisal dimensions from vocal enactments of emotion-eliciting situations performed by actors from the same two nations. We chose to compare these two nations because they exhibit different profiles on Hofstede's cultural dimensions in which Australia scores high on individualism and low on power distance whereas India show the opposite pattern (Hofstede, 2003).

The study had three aims. First, we investigated whether appraisal ratings for different enacted situations thought to elicit certain emotions were consistent with predictions based on appraisal theory as specified by Ellsworth and Scherer (2003). If listeners are able to infer aspects of the emotion-eliciting situation, as indicated by meaningful interrater agreement between listeners, this would suggest that these expressions have a symbolic function as discussed above. Second, we investigated the effects of culture by testing if listeners from Australia and India differed in their appraisal judgments. Such differences would suggest that the symbolic meaning of emotion expression might differ between cultures. Third, we investigated the extent to which judgments were associated with various acoustic characteristics. Based on the notion that the expression outcome of an emotion-eliciting event is influenced by physiological changes in the muscles that control voice production, detailed theoretical predictions of the acoustic patterns associated with cognitive appraisals could be made (Scherer, 1986). These predictions are listed below. In the categorical tradition, research suggests that several emotion categories are associated with relatively distinct patterns of acoustic characteristics (e.g. Juslin & Laukka, 2003). If appraisal ratings are consistent with the associations suggested for categorical judgments, this may indicate that categorical judgments, at least in part, could be aided by the cognitive evaluation of the emotion-eliciting situation. Such associations have been reported in a within-cultural setting in one previous study (Laukka & Elfenbein, 2012), but no study has investigated the association of appraisal ratings and acoustic parameters in a cross-cultural setting.

Results from the current study will thus provide additional clues to how various acoustic characteristics might be utilized by speakers and interpreted by listeners from different cultures. We argue that investigation of these three aims will increase our conceptual understanding about what kind of information emotional voices may carry, and that the cross-cultural aspect will further our understanding of universality and cultural specificity in emotion expression and perception.

Following Laukka and Elfenbein (2012), and based on theoretical predictions from the literature on emotion appraisal (Ellsworth & Scherer, 2003; Scherer, Schorr, & Johnstone, 2001), we set up the following predictions about the expected outcome of the listeners appraisal ratings (these predictions are also presented as grey areas in Figure 7):

1. Novelty:
   high = anger, fear, happiness
   low = sadness, serenity
2. Pleasantness:
   high = happiness, pride, relief, serenity
   low = anger, fear, sadness, shame
3. Goal conduciveness:
   high = happiness, pride, relief
   low = anger, fear, sadness, shame
4. Urgency:
   high = anger, fear
   low = happiness, relief, sadness, serenity
5. Power:
   high = anger, happiness, pride
   low = fear, sadness, shame
6. Norm compatibility:
   high = happiness, pride
   low = anger, shame

Methods

Professional actors from Australia and India were instructed to vocally enact scenarios that are commonly associated with the elicitation of anger, fear, happiness, pride, relief, sadness, serenity, and shame (Ellsworth & Scherer, 2003; Lazarus, 1991; Ortony et al., 1988). The scenarios, as described to the actors, are presented in the definitions of the emotion terms in Chapter 2. We selected 8 recordings from each of the 8 emotion categories/scenarios based on the criterion that they were recognized with sufficiently high accuracy in a previous study (Laukka et al., 2016). This selection resulted in 64 recordings from each culture.

Listeners from Australia and India were then instructed to try to imagine what type of situation the speaker in the recording was reacting to, and then judge the imagined situation on six appraisal dimensions. The questions presented to the listeners were: “Did the event occur suddenly and abruptly?” (Novelty), “Was the event pleasant?” (Pleasantness), “Did the event help the speaker to reach a goal or satisfy a need?” (Goal Conduciveness), “Did the event require the speaker to respond urgently?” (Urgency), “Could the outcome of the event be modified by the speaker's actions?” (Power), and “Was the event compatible with the speaker's norms?” (Norm compatibility). The scale ranged from 1 (“No, not at all”) to 5 (“Yes, absolutely”) for all appraisal dimensions.

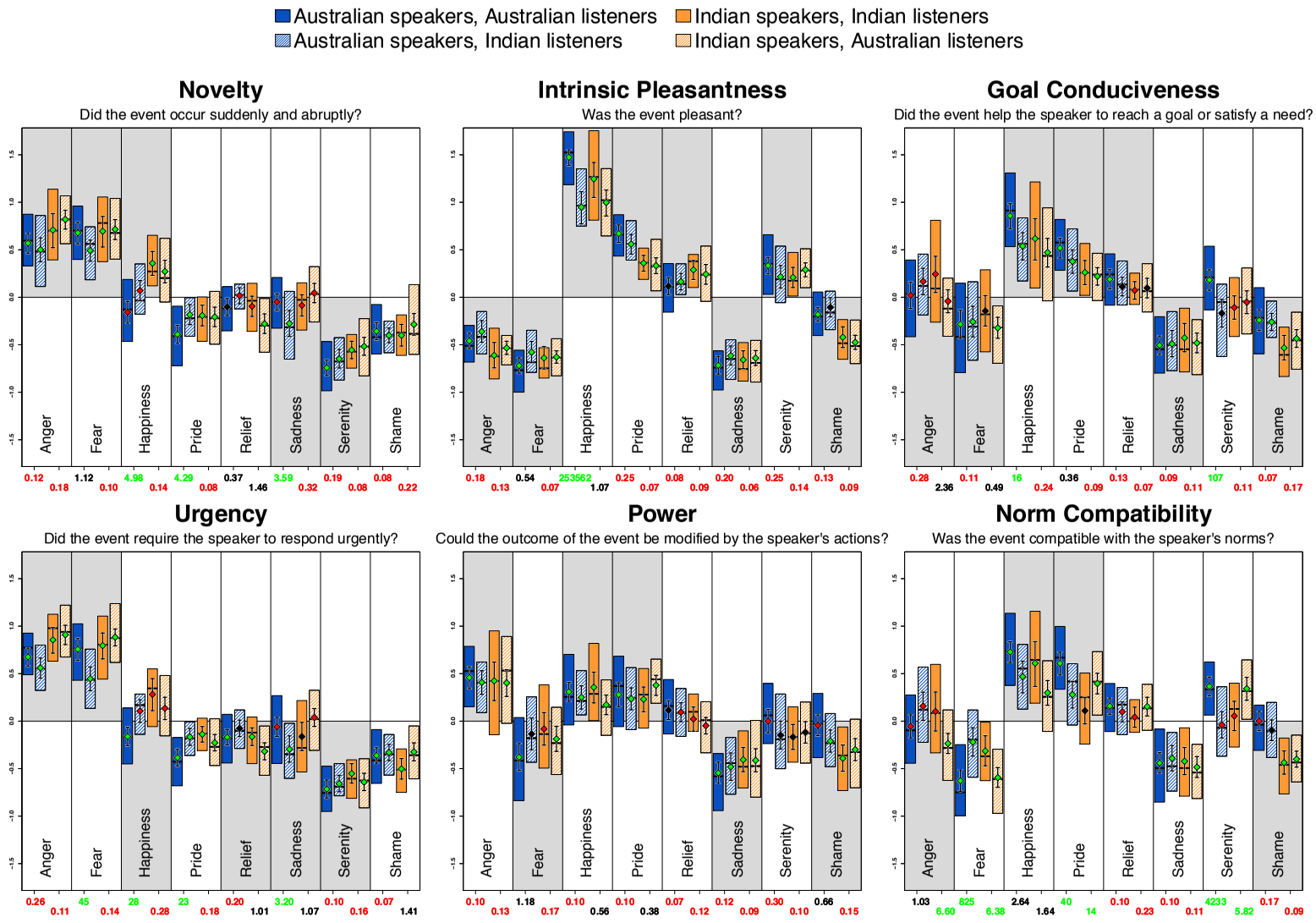
Results and discussion

Figure 7 presents a graphical summary of the appraisal ratings of Australian and Indian listeners. The boxes show the first, second (median), and third quartiles for each appraisal dimension and emotion for each combination of listener and speaker culture. The boxes can thus be used to get a notion of the direction of the appraisal ratings made both across and within cultures for each emotion expression. Blue boxes denote that speakers were Australian and orange boxes denote that speakers were Indian. Filled boxes denote that listeners and speakers were from the *same* culture, and striped boxes denote that listeners and speakers were from *different* cultures. The grey background areas denote the direction of the predictions based on appraisal theory presented above. Boxes in grey areas that do not overlap zero thus mean that at least 75% of the listeners made ratings in the predicted direction.

The green, red, and black diamonds and the adjacent error bars superimposed on the boxes in Figure 7, can also be used to estimate the direction of the appraisal ratings. The diamonds show the mean ratings and the error bars show 95% Credible Intervals (CIs) of the mean ratings. The color of the diamonds indicates if the associated Bayes Factor (BF) supports the prediction (green, $BF > 3$) or a population mean close to zero (red, $BF < 1/3$), or if both hypotheses are equally likely (black, $3 > BF > 1/3$). See the method section of the article (Study 3) for a detailed description of how the null and alternative hypotheses were defined and how the CIs and BFs were computed.

Concerning the first aim of the study, **whether ratings are consistent with predictions based on appraisal theory**, results showed that 113 out of 144 (78%) mean ratings were in the predicted direction. This result replicates the findings of Laukka and Elfenbein (2012) suggesting that listeners are able to infer several features of the emotion-eliciting situations from the

Figure 7. Mean ratings (and boxplots) of appraisal dimensions (z scores) as a function of culture and the intended emotion of the vocal expressions.



nonverbal aspects of a speaker's voice. Further, some appraisal and emotion combinations for which no theoretical predictions could be made showed consistent ratings across the two studies. Future studies could use these results to predict low ratings of novelty and urgency for shame, low ratings of urgency for pride, and low ratings of norm compatibility for fear and sadness. Such studies can provide valuable input for the refinement of appraisal theories' efforts to link emotions with specific appraisal patterns (Ellsworth & Scherer, 2003; Scherer, et al., 2001). However, some mismatches between ratings and predictions were also consistent across the studies. Such examples included ratings of novelty and urgency for happiness and sadness. This may indicate that listeners found it difficult to make the distinction between novelty and urgency, perhaps because these theoretically separate appraisals were interpreted as the general level of arousal elicited by the situation rather than specifically how novel and urgent the situation was thought to be. If listeners cannot separate between these aspects of the situation, appraisal theories might benefit from reducing these dimensions into the more general dimension of arousal. Additional studies are needed to determine whether this difficulty is specific for happiness and sadness or if it generalizes to vocal expressions in general.

Concerning the second aim of the study, **whether ratings differed between Australian and Indian listeners**, results showed very similar appraisal judgments in both in-group and out-group conditions. Out of the 113 supported predictions, 28 and 27 were from Australian and Indian participants in in-group conditions, and 27 and 31 were from Australian and Indian participants in out-group conditions. Directly comparing the ratings between listener cultures also showed relatively few group differences. The numbers presented on the horizontal axis in Figure 7 show BFs for each comparison of ratings between Australian and Indian listeners. The color-coding is the same as those used for the mean ratings; green ($BF > 3$) indicates that the BF may be interpreted as support for a difference between participant cultures, red ($BF < 1/3$) as support for no difference, and black ($3 > BF > 1/3$) that both hypotheses are equally likely. Again, the method section of the article describes in detail how the null and alternative hypotheses were defined. Out of the 96 comparisons, only 17 (18%) of the BFs gave support for a difference between listener cultures. These data suggest that appraisal inferences of emotion-eliciting situations are relatively independent of the cultural backgrounds of listeners and speakers.

Findings from cross-cultural studies in the categorical tradition show that listeners perform better than chance when discriminating between emotion categories expressed by speakers of unfamiliar cultures (Laukka et al., 2016; Sauter et al., 2010; Scherer et al., 2001; Thompson & Balkwill, 2006; Van Bezooijen et al., 1983). Although the current study did not measure recognition accuracy, we argue that these results suggest a similar pattern for ap-

praisal inferences. Mean ratings were in the predicted direction for both in-group and out-group conditions, and there were few differences between listener groups. Similar and systematic appraisal inferences across cultures suggest that cross-cultural transmission of nonverbal emotion signals is not limited to emotion categories but may also contain a symbolic representation about the characteristics of the situation that elicited the emotional response in the speaker (Scherer, 1988).

However, ratings of norm compatibility might be a notable exception. Seven out of the 16 comparisons (44 %) suggested a difference in how Australian and Indian listeners rated the expressions. Similar results for ratings of norm compatibility have been obtained in one previous study (Scherer, 1997), in which participants from 37 countries made appraisal ratings of scenarios described to them in written format. This may suggest that some appraisal dimensions might be more affected by cultural differences in expression and perception styles than others. Future cross-cultural studies could compare traditional categorical judgments and appraisal ratings of emotion expressions directly. Such studies could further our understanding of cultural effects on the perception of specific appraisal dimensions.

Concerning the third aim of the study, **the extent to which appraisal judgments were associated with acoustic characteristics**, the results are presented as an exploratory description about which acoustic cues might be associated with which appraisal ratings in both in- and out-group conditions. Table 8 presents the correlations between ratings of each appraisal dimension and acoustic cues for each combination of speaker and listener culture. These correlations indicate which acoustic cues listeners may use to make inferences about the emotion-eliciting situations.

Correlations between appraisal ratings and voice intensity (IntM), and proportion of high- vs. low-frequency energy in the voice (Hammarberg index) suggest that these cues may have been used for all appraisal judgments except intrinsic pleasantness and norm compatibility. Inferences of pleasantness and norm compatibility may instead have relied more on energy in the region of the first formant (F1A), which correlated with ratings of both these appraisals. Also, higher ratings of intrinsic pleasantness were associated with slower speech rate (i.e., low values of VoicedSegPerSec), indicating that listeners may have used slow speech rate as an indication that the speaker thought the situation was more pleasant. However, ratings of intrinsic pleasantness, which is linked to valence (positivity or negativity), generally showed smaller associations with acoustic cues than other appraisal ratings. This observation corroborates results from previous studies that report a stronger effect of arousal on the voice compared to valence (Belyk & Brown, 2014). Accordingly, ratings of novelty and urgency, which are linked to arousal, showed stronger associations with all types of acoustic cues (i.e., frequency, intensity, spectral balance, and temporal cues).

Table 8. Correlations (Pearson r) between selected acoustic parameters and listeners' mean ratings of appraisal dimensions for each combination of listener and speaker culture. Note: N = 64. Bold type indicates r ≥ 0.30.

	listener culture	novelty		urgency		intrinsic pleasantness		goal conduciveness		norm compatibility		power	
acoustic cue	speaker culture	Aus	Ind	Aus	Ind	Aus	Ind	Aus	Ind	Aus	Ind	Aus	Ind
frequency cues													
F0M	Aus	**0.46**	**0.45**	**0.50**	**0.45**	−0.05	−0.11	−0.03	0.02	−0.18	−0.09	−0.12	0.04
	Ind	**0.56**	**0.51**	**0.55**	**0.47**	−0.26	−0.15	−0.18	0.08	**−0.39**	0.00	−0.06	0.09
F0SD	Aus	−0.25	−0.24	−0.27	**−0.32**	−0.13	−0.10	−0.13	−0.22	−0.10	−0.19	−0.12	−0.24
	Ind	0.25	0.15	0.20	0.11	−0.27	−0.21	−0.23	−0.11	**−0.36**	−0.11	−0.15	−0.05
F1FreqM	Aus	0.21	0.23	0.20	0.16	−0.10	−0.05	0.01	0.02	−0.16	−0.01	−0.12	0.02
	Ind	**0.56**	**0.54**	**0.55**	**0.54**	−0.23	−0.15	−0.03	0.18	−0.25	0.14	0.22	**0.32**
intensity cues													
IntM	Aus	**0.60**	**0.70**	**0.61**	**0.69**	0.15	0.05	**0.32**	**0.47**	0.07	**0.36**	0.28	**0.50**
	Ind	**0.71**	**0.75**	**0.70**	**0.78**	0.08	0.09	**0.36**	**0.60**	0.06	**0.49**	**0.55**	**0.71**
IntSD	Aus	0.00	−0.02	0.00	−0.01	0.13	0.10	0.19	0.21	0.11	0.08	0.14	0.12
	Ind	0.00	−0.14	−0.02	−0.14	−0.13	−0.13	−0.07	−0.10	−0.14	−0.16	0.04	−0.06
spectral balance cues													
F1amplitude	Aus	0.07	0.15	0.08	0.20	**0.34**	0.21	0.29	0.26	0.23	0.26	**0.30**	0.22
	Ind	0.11	0.15	0.11	0.18	**0.33**	**0.36**	**0.34**	**0.47**	**0.34**	**0.44**	**0.40**	**0.39**
Hammarberg	Aus	**−0.41**	**−0.58**	**−0.40**	**−0.55**	−0.11	−0.09	**−0.36**	**−0.47**	−0.15	**−0.38**	**−0.32**	**−0.55**
	Ind	−0.24	−0.25	−0.18	−0.25	−0.18	−0.11	**−0.30**	−0.29	−0.17	**−0.34**	**−0.31**	**−0.38**
spectral slope	Aus	0.19	0.18	0.19	0.18	−0.03	0.03	−0.03	0.14	−0.14	0.06	0.01	0.12
	Ind	**0.34**	0.26	**0.33**	0.26	**−0.35**	**−0.32**	**−0.30**	−0.20	**−0.42**	−0.27	−0.19	−0.25
spectral flux *M*	Aus	**0.63**	**0.70**	**0.65**	**0.69**	0.01	−0.12	0.18	**0.35**	−0.05	0.20	0.16	**0.40**
	Ind	**0.75**	**0.77**	**0.74**	**0.80**	0.02	0.04	**0.30**	**0.57**	−0.01	**0.45**	**0.52**	**0.69**
spectral flux s.d.	Aus	0.18	0.25	0.20	0.22	0.08	0.07	0.20	**0.30**	0.09	0.12	0.21	0.29
	Ind	−0.09	−0.21	−0.09	−0.20	−0.13	−0.11	−0.13	−0.17	−0.19	−0.25	−0.07	−0.13
temporal cues													
VoicedSegPerSec	Aus	0.17	0.15	0.15	0.12	**−0.32**	**−0.31**	**−0.30**	−0.21	**−0.31**	−0.18	**−0.39**	−0.25
	Ind	0.21	0.20	0.29	0.28	−0.22	**−0.33**	−0.15	−0.13	−0.15	−0.21	−0.05	−0.05

It should also be noted that the correlation patterns were very similar for ratings of novelty and urgency, which matches the observation that listeners seem to have had difficulties separating between these two dimensions in the rating task. Norm compatibility, as well as novelty and urgency, were also associated with mean spectral flux (i.e., the rate of change of the power spectrum) indicating that cues of spectral balance may have been used both for appraisals related to valence and arousal.

The correlations presented in Table 8 seem consistent across speaker and listener cultures, and are also consistent with the findings of results reported in Laukka and Elfenbein (2012) for American speakers and listeners. This suggests that more detailed predictions about the associations between appraisals of emotion-eliciting situations can be made in future studies, even across cultures. Although correlations were generally consistent across cultures, some differences should be noted. For example, negative correlations were observed for fundamental frequency variability (F0SD) and ratings of novelty and urgency for Australian speakers whereas the opposite pattern was observed for Indian speakers. Also, reflecting the results of differences in how Australian and Indian listeners rated norm compatibility, there was more variability in acoustic correlations between listener cultures for these, compared with other, appraisal dimensions.

Conclusions

The general aim of this chapter was to investigate what type of information listeners can infer from an emotional voice. Appraisal theory predicts that emotional speech may contain more information than the commonly investigated categories such as “happy” or “sad”, or the two dimensions of activation and valence. According to these theories, a limited number of appraisal dimensions make up the foundation of all conceivable emotional responses and should therefore also be inferred from expressive behaviour. The current study investigated if listeners could infer several aspects of the events that may have triggered the emotion of the speaker. Results suggest that such inferences are possible even in a cross-cultural setting.

More specifically, the predictions based on appraisal theory as specified by Ellsworth and Scherer (2003) were consistent with the listeners’ ratings suggesting that emotional expressions have a symbolic function in communication, and that this function is consistent across cultures. The acoustic analysis explored the correlates of how this symbolic meaning might be conveyed by the voice. These results were also consistent with the predictions of how emotion-eliciting events influence physiological changes in the muscles involved in voice production (Scherer, 1986). The focus on the third, and least investigated aspect of vocal expressions – their symbolic meaning – gives new insights into the communicative capabilities of the

voice. The current results thus provide additional clues in the refinement of the theories trying to explain how emotions are communicated within and across cultures. Seen from a methodological perspective, the study also contributes to the field by presenting a novel way to study emotional communication that is not tied to the traditional emotion categories. By showing that emotional expressions may be judged along several rating scales describing appraisal dimensions, the current results may encourage future work to find new ways to think about emotion recognition. This method may allow listeners to express their perceptions in a more fine-grained manner that perhaps will capture the nuanced and multifaceted emotional concepts that the voice may convey. As suggested by the current results, such methods may prove especially valuable in cross-cultural settings because emotion terms may have slightly differing meanings across cultures and/or languages.

7. General discussion and conclusions

Emotional expressions serve important functions in our everyday social interactions. They allow individuals to communicate vital information about their current state of mind to others, which facilitates coordinated interactions. The thesis investigated one channel by which this information is conveyed, namely the nonverbal aspects of the voice. Each chapter of the thesis investigated nonverbal communication of emotions from slightly different perspectives, and each was intended to answer the broad research questions presented in that chapter's title.

Chapter 3 investigated the broad question of how emotions are communicated in speech. More specifically, that chapter described how the communication process depends on the speaker's manipulations of the voice and how these manipulations can be understood by listeners and measured with acoustic analysis of the waveform. The purpose of describing the acoustic properties of emotional speech, both those reviewed and the empirical findings presented, was to further our understanding of the mechanisms that allow humans to communicate emotions nonverbally. This question reoccurs throughout the thesis, and in the three studies, but the purpose of Chapter 3 was to give a broad overview and empirically evaluate some of the fundamental ideas of previous research while the following chapters investigated more specific aspects of the communication process.

Chapter 4 and Study 1 investigated how much acoustic information listeners need to infer emotions from speech and music. More specifically, results showed that low-level acoustic features – available in the very first tenths of a second of an utterance or a musical tune – allow listeners to recognize and differentiate several emotions even with very limited information. The study also presented trajectories for the included emotions describing how the recognition accuracy increased as more acoustic information became available to the listeners.

Chapter 5 and Study 2 investigated what emotions listeners can infer from non-linguistic vocalizations. That chapter investigated the limits of emotional communication via nonverbal aspects of the voice, and showed that non-linguistic vocalizations could reliably convey several emotions that are usually difficult for listeners to recognize from speech prosody alone. This suggests that such vocalizations serve important functions in the emotional communication process and that speakers may use non-linguistic vocaliza-

tions as a complement to prosody to emphasize the emotional message so that it is more easily understood.

Chapter 6 and Study 3 investigated whether listeners can infer appraisal dimensions across cultures from speech. More specifically, that chapter investigated whether emotional expressions may carry symbolic meaning about how a speaker evaluated the emotion-eliciting situation. Results suggested that listeners do make such inferences, which give support to the notion that elicitation of emotions are connected to specific cognitive appraisals. This means that emotional communication is not only based on inferences about the speaker's inner state, but also about the event preceding that state.

Outlook and limitations

Even though decades of research has aimed to describe the acoustic properties allowing nonverbal communication of emotions in speech, one conclusion of Chapter 3 was that the predictions available in the literature were inadequate to predict listeners' judgments for several emotions, even in the controlled setting of the experiment. It is puzzling that so many studies show that listeners can recognize discrete emotions from nonverbal aspects of speech but that research has not yet been able to produce reliable predictions about the acoustic parameters explaining this ability. One reason for this may be that the patterns suggested by previous literature are rather vague and presented in relative, rather than absolute terms. The theoretical predictions based on how different emotions may influence the voice are usually described in a high/medium/low format, which does not seem to capture many of the subtle vocal manipulations that listeners used to categorize emotional expressions in the experiment presented in Chapter 3. Although many of the theoretical predictions gained support from the results in the sense that they were influenced in the expected direction for many emotions, they were too vague to explain how listeners could differentiate between emotions with seemingly similar acoustic parameter-patterns.

The acoustic parameter patters presented in Chapter 3 (Figure 3, pages 41-53) are instead described as absolute deviations from the speakers' normal or neutral voice. Thus, this figure contains the answer(s) to the main research question of Chapter 3, at least for the actors, speakers, and emotions included in the study. Although the complexity of the results may seem unsatisfactory if one wishes to get a quick grasp of how emotions are generally communicated, I hope the reader appreciates that the very specific predictions presented here are bolder, and thus more easily falsified, than those presented by previous research. I believe that this presentation style will allow for a more cumulative science because the patterns can be directly

compared with findings of future research and thus gain gradual support, be tuned or adjusted, as more empirical evidence is added. As is the case of any cumulative science, results of future studies will be needed to determine if the acoustic parameter patterns suggested here will generalize to other groups of actors and listeners. Future research will also determine which of these parameters are most important for communication of specific emotions, which can be omitted, and which additional parameters need to be added to understand the mechanisms of emotional communication.

An assumption in emotional communication is that listeners use both low-level acoustic parameters that are available in a sound almost instantaneously and other acoustic parameters that need longer time to develop. The contour of pitch – or in musical terms – the relation between consecutive notes that make up a melody, are examples of parameters that need longer time to develop, whereas the pitch or loudness of any sound can be heard instantaneously. The rationale of Study 1 was to restrict the acoustic information in an attempt to separate the acoustic parameters that need longer and shorter time to develop and study how they influence the recognition of emotions. Although the acoustic information in low-level cues was sufficient to recognize some emotion expressions, others required longer time, and perhaps other types of acoustic information to be recognized, especially for music. Also, results showed that the recognition accuracy increased as speech and music unfolds. This suggests that even though low-level acoustic parameters aid communication of emotion, other parameters such as pitch variability, contours, and speech rate/tempo may be needed for listeners to be certain in their inferences. Combining the findings of Study 1 with the acoustic patterns presented in Chapter 3 suggests that the emotion expressions that listeners could recognize with the least amount of acoustic information – anger, happiness, sadness, and fear – are also very different from the speakers' neutral voice when it comes to low-level acoustic parameters. This is especially evident for anger and fear for loudness and for all four emotions for pitch, suggesting that these parameters may be especially important carriers of these emotions.

I believe that future research will show that the acoustic parameters related to pitch, loudness, and tempo will persist as the main carriers of the emotional message and that future research will benefit more from a closer investigation of these parameters rather than trying to add more spectral balance parameters. Spectral balance parameters should of course not be neglected, but they seem to be in abundance in many parameter sets, and the vocal manipulations related to these parameters may already be well covered. Instead, I believe that parameters reflecting the contours of pitch and loudness – that is, how they develop in time during an utterance – could be improved. In the acoustic parameter set used throughout this thesis (eGeMaps), there are only a few parameters intended to capture how pitch and

loudness change during an utterance (the mean and standard deviation of rising and falling slope), and these measures may not be sensitive enough. I believe that closer inspection of the information contained in the vocal changes related to pitch and loudness contours will be important in the quest to describe the acoustic correlates of emotional communication. Future research should also pay more attention to individual differences in peoples' voices rather than focus only on mean values averaged across many individuals.

One of the main theoretical assumption behind the research aims guiding this thesis is that the voice is influenced by physiological changes of the individual that is caused by the emotional episode. This assumption highlights that a large part of the communication of emotions is based on involuntary changes of the voice. However, it is also assumed that both speakers and listeners are aware of how emotions influence the voice, which makes it possible for speakers to attenuate or emphasize the emotional message. My impression is that many studies have focused mainly on the physiological changes thought to influence the voice rather than the intentional manipulations that speakers' use deliberately to communicate emotions. This probably has to do with differing aims of different studies, some focusing on explaining the communication process while others use vocal and other expressions as a means for studying emotions in a more general sense. I believe that the mix of aims, and that these aims are not always clearly stated, could be another reason that the mechanisms allowing emotional communication are still elusive.

All studies in the thesis are limited by the use of enacted emotion expressions, rather than expressions spontaneously produced by speakers while they were experiencing emotional episodes. Recent studies have showed small but systematic differences with respect to the acoustic characteristics of enacted and spontaneous emotion expressions (e.g., Juslin, Laukka, & Bänziger, 2018), and results of this thesis should ideally be replicated also using spontaneously produced stimuli. However, it can also be argued that the distinction between acted and spontaneous expressions is not clear-cut and may instead best be viewed as a continuum, because people often up- and down-regulate their expressions during real-life interactions (e.g., Scherer & Bänziger, 2010). When using actor portrayals, it is important that the instructions and emotion terms are carefully defined. If there is no theory behind the production of the stimulus material and procedure of the experiment, diverging results between studies may be due to confusion of what emotion was expressed by the speaker and/or interpreted by the listener. If the definition of the emotion is unclear, results are difficult to compare between studies. The term anger may comprise everything from mild irritation to violent rage. Actors may have differed in their interpretation and listeners may have misclassified irritation for contempt for example. One way to mit-

igate this issue is to use descriptions of scenarios commonly associated with different emotions, such as those used in the present thesis (and presented in Chapter 2).

Listeners' ability to understand emotions from nonverbal aspects of the voice does not only rely on prosodic manipulations that people make during speech, but also on the many sounds we make in between sentences or when we are not speaking. To enhance experimental control in the studies presented in this thesis, non-linguistic vocalizations and the prosodic aspects of speech have been investigated in separate studies. In natural conversations though, listeners make use of both simultaneously, which likely enhances the ability of the voice to convey emotions. Chapter 5 and Study 2 showed that the emotional message of the voice could be emphasized by the use of non-linguistic vocalizations because they can convey emotions that are difficult to interpret from the prosodic aspects of speech alone. To enhance ecological validity and to broaden the scope of how emotions are communicated, future studies could combine these channels in single experiments. This would allow investigation of potential interaction effects between speech and other vocalizations that we make during conversations, which could give new insights into the emotional communication process. Such experiments could perhaps explain why some emotions (e.g., disgust) are poorly recognized from prosodic aspects of speech while they are clearly visible in facial expressions.

Although people make all sorts of inferences from the nonverbal aspects of a person's voice, emotional communication is usually studied either in terms of discrete emotions or along a small number of dimensions (valence and arousal). Appraisal theory suggests that the cognitive appraisal of a situation, rather than the situation itself, is the main component that elicits emotions. Therefore it is assumed that listeners should be able to infer aspects of a speaker's appraisal of emotion-eliciting situations. Chapter 6 and Study 3 showed that listeners could make several inferences, other than valence and arousal, about how the speaker appraised the emotion-eliciting situation. Such inferences facilitate emotional communication because they allow listeners to understand many aspects of speakers' cognitive evaluations – their values and goals, their ability to anticipate and cope with the consequences of the situation – which may be used to better interpret the emotional state of a speaker. These results highlight the possibility to study other aspects of emotional communication than discrete emotion categories.

The long withstanding study of discrete emotions is strongly linked to the use of the forced-choice paradigm. The use of this paradigm in all studies in this thesis except Study 3 could be seen as a disadvantage because the recognition rates of different emotions are affected by which emotion-labels are included in an experiment. Because some emotions are conceptually closer than others, including very dissimilar emotions in an experiment

could thus increase the recognition rates. This is definitely a disadvantage in studies aiming to describe the nature of emotions (e.g., which affective states are "real" emotions) because replacing one emotion term with another will lead to different conclusions. However, when studying how emotions are communicated, I believe this side effect of the design could give valuable information because it tells us if the acoustic properties allowing recognition of one emotion are similar to those of another.

The forced choice task has also been criticised for having a low ecological validity. It is obvious that real life emotion recognition is very different from the task of selecting an emotion label from a list of alternatives. However, because of the high level of control and the fact that much is unknown about how emotions are communicated in the voice, I believe that this task may still generate important findings in future research. By first understanding how emotions are conveyed in this simplest form of communication, results from such studies may then be tested in more ecologically valid settings. These more ecologically valid settings may include spontaneous rather than enacted expressions, investigate effects of the context and situation on emotion perception, and even include other channels beside the voice (e.g., facial gestures and body movement).

The studies of this thesis aimed to describe the physical properties of the sounds that people experience as expressions of emotions. Although this process is very complex and relies on many different vocal manipulations and cognitive abilities, I believe that this research could benefit by allowing more influence from the field of psychophysics. The general aim of psychophysical studies is to connect psychological experience with the physical characters of the stimuli that give rise to these experiences. Psychophysics generally deals with more basic aspects of auditory perception than those involved in emotional communication (e.g., to describe how loud a sound with a certain energy is perceived), but I believe that many of the rigorous methods used are directly applicable to perception of emotions. In the end, communication via auditory cues boils down to the production and perception of sounds and though the processes involved to interpret them may be more complex, attention to the perceptual details may reveal important aspects of the bigger picture. A better understanding of the "basic science" of vocal emotion recognition will also lead to better applications, such as training programs for improving emotion expression and recognition skills, facilitated human-computer interaction and automatic classifier systems, and even improved understanding between people from different cultures.

References

Ackermann, H., Hage, S. R., & Ziegler, W. (2014). Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*, *37*(06), 529–546. https://doi.org/10.1017/S0140525X13003099

Adolphs, R. (2016). How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences. *Social Cognitive and Affective Neuroscience*, nsw153. https://doi.org/10.1093/scan/nsw153

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

Bachorowski, J.-A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, *110*(3), 1581–1597.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*(5), 1161–1179. https://doi.org/10.1037/a0025827

Barrett, H. C., & Bryant, G. (2008). Vocal Emotion Recognition Across Disparate Cultures. *Journal of Cognition and Culture*, *8*(1), 135–148. https://doi.org/10.1163/156770908X289242

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*(1), 28–58.

Barrett, L. F. (2014). The Conceptual Act Theory: A Précis. *Emotion Review*, *6*(4), 292–297. https://doi.org/10.1177/1754073914534479

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*, 1–23. https://doi.org/10.1093/scan/nsw154

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323

Belyk, M., & Brown, S. (2014). The Acoustic Correlates of Valence Depend on Emotion Family. *Journal of Voice*, *28*(4), 523.e9-523.e18. https://doi.org/10.1016/j.jvoice.2013.12.007

Bigand, E., Filipic, S., & Lalitte, P. (2005). The Time Course of Emotional Responses to Music. *Annals of the New York Academy of Sciences*, *1060*(1), 429–437. https://doi.org/10.1196/annals.1360.036

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence: Vocal communication of emotions. *Journal of Zoology*, *288*(1), 1–20. https://doi.org/10.1111/j.1469-7998.2012.00920.x

Chen, X., Zhao, L., Jiang, A., & Yang, Y. (2011). Event-related potential correlates of the expectancy violation effect during emotional prosody processing. *Biological Psychology*, *86*(3), 158–167. https://doi.org/10.1016/j.biopsycho.2010.11.004

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, *16*(1), 117.

Cornew, L., Carver, L., & Love, T. (2009). There's more to emotion than meets the eye: A processing bias for neutral content in the domain of emotional prosody. *Cognition & Emotion*, *24*(7), 1133–1152. https://doi.org/10.1080/02699930903247492

Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2018). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*. https://doi.org/10.1037/amp0000399

Darwin, C. (1872). *The expression of the emotions in man and animals* (3rd ed.). New York, USA: Oxford University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-Cultural Elements in Facial Displays of Emotion. *Science*, *164*(3875), 86–88. https://doi.org/10.1126/science.164.3875.86

Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Review*, *5*(1), 90–96.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (Vol. 572, pp. 572–595). New York: Oxford University Press.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., … Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the munich open-source multimedia feature extractor. In A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, & R. Zimmermann (Eds.), *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 835–838). https://doi.org/10.1145/2502081.2502224

Filipic, S., Tillmann, B., & Bigand, E. (2010). Judging familiarity and emotion from very brief musical excerpts. *Psychonomic Bulletin & Review*, *17*(3), 335–341. https://doi.org/10.3758/PBR.17.3.335

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., … Güntürkün, O. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1859), 20170990. https://doi.org/10.1098/rspb.2017.0990

Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*(1), 173–215. https://doi.org/10.1016/j.cognition.2005.11.009

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, *128*(3), 1322–1336. https://doi.org/10.1121/1.3466853

Grichkovstova, I., Lacheret, A., Morel, M., Beaucousin, V., & Tzourio-Mazoyer, N. (2007). Affective speech gating. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 805–808). Saarbrücken, Germany: Saarland University.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283. https://doi.org/10.3758/BF03204386

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical Correlates of Affective Prosody. *Journal of Voice*, *21*(5), 531–540. https://doi.org/10.1016/j.jvoice.2006.03.002

Hawk, S. T., Van Kleef, G. A., Fischer, A. H., & Van Der Schalk, J. (2009). " Worth a thousand words": Absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion*, *9*(3), 293.

Hess, U., & Hareli, S. (2016). On the malleability of the meaning of contexts: the influence of another person's emotion expressions on situation perception. *Cognition and Emotion*, 1–7. https://doi.org/10.1080/02699931.2016.1269725

Hess, U., & Thibault, P. (2009). Darwin and emotion expression. *American Psychologist*, *64*(2), 120–128. https://doi.org/10.1037/a0013386

Hofstede, G. (2003). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.).

Izard, C. (1971). *The face of emotion*. New York: Appleton-Century-Crofts.

James, W. (1884). What is an Emotion? *Mind*, *os-IX*(34), 188–205. https://doi.org/10.1093/mind/os-IX.34.188

Jiang, X., Paulmann, S., Robin, J., & Pell, M. D. (2015). More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 597–612. https://doi.org/10.1037/xhp0000043

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of Emotions*, *2*, 220–235.

Jürgens, U. (2009). The Neural Control of Vocalization in Mammals: A Review. *Journal of Voice*, *23*(1), 1–10. https://doi.org/10.1016/j.jvoice.2007.07.005

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., & Bänziger, T. (2018). The Mirror to Our Soul? Comparisons of Spontaneous and Posed Vocal Expression of Emotion. *Journal of Nonverbal Behavior*, *42*(1), 1–40. https://doi.org/10.1007/s10919-017-0268-x

Juslin, P. N., & Scherer, K. R. (2005). Vocal Expression of Affect. In *Handbook of methods in nonverbal behavior research*.

Keltner, D., & Haidt, J. (1999). Social Functions of Emotions at Four Levels of Analysis. *Cognition & Emotion*, *13*(5), 505–521. https://doi.org/10.1080/026999399379168

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. (2013). Cross-Cultural Differences in the Processing of Non-Verbal Affective Vocalizations by Japanese and Canadian Listeners. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00105

Kreiman, J., & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. https://doi.org/10.1002/9781444395068

Laukka, P., & Elfenbein, H. A. (2012). Emotion Appraisal Dimensions can be Inferred From Vocal Expressions. *Social Psychological and Personality Science*, *3*(5), 529–536. https://doi.org/10.1177/1948550611428011

Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, *111*(5), 686–705. https://doi.org/10.1037/pspi0000066

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K. (2011). Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, *25*(1), 84–104. https://doi.org/10.1016/j.csl.2010.03.004

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

Lazarus, R. S. (1968). Emotions and adaptation: Conceptual and empirical relations. *Nebraska Symposium on Motivation*. University of Nebraska Press.

Lazarus, R. S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

Lazarus, R. S., Coyne, J. C., & Folkman, S. (1984). Cognition, emotion and motivation: The doctoring of Humpty-Dumpty. *Approaches to Emotion*, 221–237.

Lieberman, M. D. (2018). Boo! The consciousness problem in emotion. *Cognition and Emotion*, 1–7. https://doi.org/10.1080/02699931.2018.1515726

Lima, C. F., Anikin, A., Monteiro, A. C., Scott, S. K., & Castro, S. L. (2018). Automaticity in the recognition of nonverbal emotional vocalizations. *Emotion*. https://doi.org/10.1037/emo0000429

Mithen, S., Morley, I., Wray, A., Tallerman, M., & Gamble, C. (2006). The Singing Neanderthals: the Origins of Music, Language, Mind and Body, by Steven Mithen. London: Weidenfeld & Nicholson, 2005. ISBN 0-297-64317-7 hardback £20 & US$25.2; ix+374 pp. *Cambridge Archaeological Journal*, *16*(1), 97–112. https://doi.org/10.1017/S0959774306000060

Moors, A. (2014). Flavors of Appraisal Theories of Emotion. *Emotion Review*, *6*(4), 303–307. https://doi.org/10.1177/1754073914534477

Moors, A. (2017). Integration of Two Skeptical Emotion Theories: Dimensional Appraisal Theory and Russell's Psychological Construction Theory. *Psychological Inquiry*, *28*(1), 1–19. https://doi.org/10.1080/1047840X.2017.1235900

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

Mustafa, M. B., Yusoof, M. A. M., Don, Z. M., & Malekzadeh, M. (2018). Speech emotion recognition research: an analysis of research focus. *International Journal of Speech Technology*, *21*(1), 137–156. https://doi.org/10.1007/s10772-018-9493-x

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge, UK: Cambridge university press.

Owren, M. J., Amoss, R. T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology*, *73*(6), 530–544. https://doi.org/10.1002/ajp.20913

Parncutt, R. (2014). The emotional connotations of major versus minor tonality: One or more origins? *Musicae Scientiae*, *18*(3), 324–353.

Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language*, *105*(1), 59–69. https://doi.org/10.1016/j.bandl.2007.11.005

Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: How much is enough? *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 230–242. https://doi.org/10.3758/CABN.10.2.230

Peirce, J. W. (2007). PsychoPy — Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pell, M. D., & Kotz, S. A. (2011). On the Time Course of Vocal Emotion Recognition. *PLoS ONE*, *6*(11), e27256. https://doi.org/10.1371/journal.pone.0027256

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*(4), 417–435. https://doi.org/10.1016/j.wocn.2009.07.005

Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, *68*(2), 111–141. https://doi.org/10.1016/S0010-0277(98)00043-2

Planalp, S. (1996). Communicating emotion in everyday life: Cues, channels, and processes. In *Handbook of communication and emotion* (pp. 29–48). Elsevier.

Provine, R. R. (2012). *Curious behavior: Yawning, laughing, hiccupping, and beyond*. Harvard University Press.

Rigoulot, S., Wassiliwizky, E., & Pell, M. D. (2013). Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00367

Roseman, I. J. (2013). Appraisal in the emotion system: Coherence in strategies for coping. *Emotion Review*, *5*(2), 141–149.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual Cues in Nonverbal Vocal Expressions of Emotion. *Quarterly Journal of Experimental Psychology*, *63*(11), 2251–2272. https://doi.org/10.1080/17470211003721642

Sauter, D. A., & Eimer, M. (2010). Rapid Detection of Emotion from Human Vocalizations. *Journal of Cognitive Neuroscience*, *22*(3), 474–481. https://doi.org/10.1162/jocn.2009.21215

Scheiner, E., & Fischer, J. (2011). Emotion Expression: The Evolutionary Heritage in the Human Voice. In W. Welsch, W. J. Singer, & A. Wunder (Eds.), *Interdisciplinary Anthropology: Continuing Evolution of Man* (pp. 105–129). https://doi.org/10.1007/978-3-642-11668-1_5

Scherer, K. R. (1988). On the Symbolic Functions of Vocal Affect Expression. *Journal of Language and Social Psychology*, *7*(2), 79–100. https://doi.org/10.1177/0261927X8800700201

Scherer, K. R. (2018). Acoustic patterning of emotion vocalizations. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 61–91). Oxford, New York: Oxford University Press.

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. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, *73*(5), 902–922. https://doi.org/10.1037/0022-3514.73.5.902

Scherer, K. R. (2013). Affect Bursts as Evolutionary Precursors of Speech and Music. In G. A. Danieli, A. Minelli, & T. Pievani (Eds.), *Stephen J. Gould: The Scientific Legacy* (pp. 147–167). Springer Milan.

Scherer, K. R., & Bänziger, T. (2010). On the use of actor portrayals in research on emotional expression. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 271-294). New York: Oxford University Press.

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., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology*, *46*(6), 401–435. https://doi.org/10.1080/00207594.2011.626049

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*.

Scott, S. K., Young, A. W., Calder, A. J., Hellawell, D. J., Aggleton, J. P., & Johnsons, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, *385*(6613), 254.

Sergi, I., Fiorentini, C., Trznadel, S., & Scherer, K. R. (2016). Appraisal Inference from Synthetic Facial Expressions. *International Journal of Synthetic Emotions (IJSE)*, *7*(2), 45–61. https://doi.org/10.4018/IJSE.2016070103

Shiota, M. N., Campos, B., Oveis, C., Hertenstein, M. J., Simon-Thomas, E., & Keltner, D. (2017). Beyond happiness: Building a science of discrete positive emotions. *American Psychologist*, *72*(7), 617.

Shuman, V., Clark-Polner, E., Meuleman, B., Sander, D., & Scherer, K. R. (2017). Emotion perception from a componential perspective. *Cognition and Emotion*, *31*(1), 47–56.

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., … Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, *144*(4), 343–393. https://doi.org/10.1037/bul0000128

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, *9*(6), 838–846. https://doi.org/10.1037/a0017810

Skinner, E. R. (1935). A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness; and a determination of the pitch and force of the subjective concepts of ordinary, soft, and loud tones. *Communications Monographs*, *2*(1), 81–137.

Spreckelmeyer, K. N., Kutas, M., Urbach, T., Altenmüller, E., & Münte, T. F. (2009). Neural processing of vocal emotion and identity. *Brain and Cognition*, *69*(1), 121–126. https://doi.org/10.1016/j.bandc.2008.06.003

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

Tangney, J. P., & Tracy, J. L. (2012). Self-conscious emotions. In M. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 446–478). New York: Guilford Press.

Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proceedings of the National Academy of Sciences*, *109*(46), 19027–19032. https://doi.org/10.1073/pnas.1210344109

Thompson, W. F., & Balkwill, L. (2006). Decoding speech prosody in five languages. *Semiotica*, *2006*(158), 407–424.

Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

Tracy, J. L. (2014). An Evolutionary Approach to Understanding Distinct Emotions. *Emotion Review*, *6*(4), 308–312. https://doi.org/10.1177/1754073914534478

Tracy, J. L., & Randles, D. (2011). Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, *3*(4), 397–405. https://doi.org/10.1177/1754073911410747

Van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of Vocal Expressions of Emotion: A Three-Nation Study to Identify Universal Characteristics. *Journal of Cross-Cultural Psychology*, *14*(4), 387–406. https://doi.org/10.1177/0022002183014004001

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, *22*(4), 720–752. https://doi.org/10.1080/02699930701503567

Wallin, N. L., Merker, B., & Brown, S. (Eds.). (2001). The 'musilanguage' model of music evolution. In *The origins of music* (pp. 271–300). Cambridge, MA: MIT press.

Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews*, *21*(5), 195–205. https://doi.org/10.1002/evan.21319