The time course of emotion recognition in speech and music

Henrik Nordström, and Petri Laukka

Citation: The Journal of the Acoustical Society of America **145**, 3058 (2019); doi: 10.1121/1.5108601
View online: https://doi.org/10.1121/1.5108601
View Table of Contents: https://asa.scitation.org/toc/jas/145/5
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Listening benefits in speech-in-speech recognition are altered under reverberant conditions
The Journal of the Acoustical Society of America **145**, EL348 (2019); https://doi.org/10.1121/1.5100898

The role of pitch and harmonic cancellation when listening to speech in harmonic background sounds
The Journal of the Acoustical Society of America **145**, 3011 (2019); https://doi.org/10.1121/1.5102169

Congenital blindness enhances perception of musical rhythm more than melody in Mandarin speakers
The Journal of the Acoustical Society of America **145**, EL354 (2019); https://doi.org/10.1121/1.5100899

Audio-based deep music emotion recognition
AIP Conference Proceedings **1967**, 040021 (2018); https://doi.org/10.1063/1.5039095

Relationship between the privacy index and the speech privacy class
The Journal of the Acoustical Society of America **145**, EL435 (2019); https://doi.org/10.1121/1.5109049

Envelope regularity discrimination
The Journal of the Acoustical Society of America **145**, 2861 (2019); https://doi.org/10.1121/1.5100620

Advance your science and career as a member of the
ACOUSTICAL SOCIETY OF AMERICA

LEARN MORE

The time course of emotion recognition in speech and music

Henrik Nordström[a] and Petri Laukka
Department of Psychology, Stockholm University, 106 91 Stockholm, Sweden

(Received 2 July 2018; revised 25 April 2019; accepted 25 April 2019; published online 28 May 2019)

The auditory gating paradigm was adopted to study how much acoustic information is needed to recognize emotions from speech prosody and music performances. In Study 1, brief utterances conveying ten emotions were segmented into temporally fine-grained gates and presented to listeners, whereas Study 2 instead used musically expressed emotions. Emotion recognition accuracy increased with increasing gate duration and generally stabilized after a certain duration, with different trajectories for different emotions. Above-chance accuracy was observed for $\leq$100 ms stimuli for anger, happiness, neutral, and sadness, and for $\leq$250 ms stimuli for most other emotions, for both speech and music. This suggests that emotion recognition is a fast process that allows discrimination of several emotions based on low-level physical characteristics. The emotion identification points, which reflect the amount of information required for stable recognition, were shortest for anger and happiness for both speech and music, but recognition took longer to stabilize for music vs speech. This, in turn, suggests that acoustic cues that develop over time also play a role for emotion inferences (especially for music). Finally, acoustic cue patterns were positively correlated between speech and music, suggesting a shared acoustic code for expressing emotions.
© *2019 Acoustical Society of America*. https://doi.org/10.1121/1.5108601

[DB] Pages: 3058–3074

I. INTRODUCTION

Everyday experiences tell us that we can quickly hear if an individual's tone of voice starts to turn emotional. Likewise, when we hear emotional music, it can capture our attention within seconds. But how quickly can we recognize emotions from speech prosody and music? And are some emotions recognized faster than others? To address these questions, this paper presents two studies that use a gating paradigm to estimate how much acoustic information is needed to recognize a variety of positive and negative emotions from speech prosody (Study 1) and music performances (Study 2). In addition, this paper describes and compares the acoustic patterns that are associated with emotions for both speech and music.

A. Expression and perception of emotion in speech and music

Emotional expressions convey information about the relationship between the expresser and their environment. The ability to quickly and accurately identify emotional expressions is therefore important for social interaction and enables individuals to promote goals and to detect and avoid negative outcomes (e.g., Keltner and Haidt, 1999). Past research on nonverbal communication of emotion has had a strong focus on facial expressivity (e.g., Nummenmaa and Calvo, 2015), but recent advances in affective computing has led to a major increase of studies also on vocal expressivity (Schuller *et al.*, 2011).

Previous vocal expression studies report evidence for relatively distinct patterns of acoustic-perceptual cues which are associated with a number of emotion categories such as anger, fear, joy/happiness, and sadness (e.g., Banse and Scherer, 1996; Hammerschmidt and Jürgens, 2007; Laukka *et al.*, 2016; Sauter *et al.*, 2010). Some of the most commonly investigated cues include fundamental frequency ($F0$), voice intensity, voice quality (e.g., energy distribution across the frequency spectra), and durations (e.g., speech rate). For example, voices expressing anger usually display fairly high levels of $F0$ and voice intensity coupled with much high-frequency energy and fast speech rate, whereas voices expressing sadness instead show low levels of $F0$ and voice intensity, little high-frequency energy, and slow speech rate (for detailed reviews, see Juslin and Laukka, 2003; Scherer, 2019). Acoustic patterns such as the above are assumed to at least partly reflect emotion-related effects on the physiology of the vocal apparatus, although more research is needed to understand the mechanisms underlying these effects. Acoustic patterns may also be affected by more or less conscious regulation strategies, where expressers may wish to convey certain emotions, but not others, for strategic reasons (Scherer, 1986).

Perception studies have shown that listeners are able to recognize a variety of emotions from vocal expressions both within and across cultures (Scherer *et al.*, 2011), although recognition rates tend to be higher when stimuli come from the perceivers' own culture compared to stimuli from unfamiliar cultures (for meta-analyses, see Elfenbein and Ambady, 2002; Juslin and Laukka, 2003). In general, there is an imbalance between expression and perception studies, where perception studies are more prevalent and also include a larger number of different emotions compared to expression studies (Scherer *et al.*, 2011). It can be noted that several recent studies suggest that a variety of positive

[a]Electronic mail: henrik.nordstrom@psychology.su.se

emotions—over and above joy/happiness—seem to be well recognized from vocal expressions (e.g., Cordaro *et al.*, 2016; Cowen *et al.*, 2018; Laukka *et al.*, 2013; Simon-Thomas *et al.*, 2009), but relatively few studies have presented acoustic patterns for positive emotions (Sauter, 2017).

Like speech, music is also transmitted through sound and research has identified many similarities in the way acoustic cues are used to express emotions in speech and music. Juslin and Laukka (2003) reviewed and compared cue patterns using data from 77 vocal expression studies and 35 music performance studies. They showed that acoustic cues related to frequency, intensity, sound quality, and durations showed similar associations with various emotion categories for both speech and music. Recent empirical studies have directly compared acoustic cues in vocal expression and music performance, using a variety of musical instruments and genres, and also reveal many parallels between the two modes of expression (e.g., Bowling *et al.*, 2012; Coutinho and Schuller, 2017; Scherer *et al.*, 2015).

Indeed, the similarities in cue patterns between expressive speech and music have been proposed as one reason for why listeners experience that music expresses emotions in the first place (Juslin and Laukka, 2003). Although fewer studies have been conducted on perception of musically expressed emotions compared to studies of vocal expression, similar trends can be observed in the literature. Several emotions can be recognized with accuracy above chance from both culturally familiar and unfamiliar music (e.g., Balkwill and Thompson, 1999; Fritz *et al.*, 2009), although recognition rates are again higher when the cultural origin of the music stimuli matches the listeners' own cultural background (Laukka *et al.*, 2013).

One defining quality of both speech and music is that they are dynamic processes that develop over time. Some of the acoustic properties that are thought to convey emotion need a longer time to develop (e.g., *F*0 variability, durations), whereas others are almost instantaneous (e.g., intensity level). Perceivers therefore need to attend to acoustic properties that vary over shorter and longer time periods—which makes temporal analysis of expression and perception an important enterprise. Because different emotions are associated with different cue patterns that may unfold in time at different rates, it is possible that some emotions would be recognized faster than others. In contrast to the majority of past studies—which have investigated acoustic measures and listener impressions averaged across complete stimuli—the current study focuses specifically on the time-course of emotion recognition.

B. Previous studies of the time course of vocal emotion recognition

Indirect evidence on the time-course of vocal emotion recognition is available from studies using event-related potentials (ERPs) and the mismatch negativity paradigm. For example, Paulmann and Pell (2010) presented primes with short fragments of vocal emotion expressions followed by either an emotionally congruent or an incongruent facial expression. Results showed a N400 response to incongruent trials when the vocal primes were 400 ms long but not when the prime was 200 ms, suggesting that emotion recognition occurs somewhere in this range. Other studies report similar early differentiation (around 200 ms) in the latency of ERPs in response to processing of neutral vs emotional vocal stimuli (e.g., Chen *et al.*, 2011; Liu *et al.*, 2012; Paulmann and Kotz, 2008; Sauter and Eimer, 2010; Spreckelmeyer *et al.*, 2009). Another indirect method is to investigate the behavioral response latency in speeded classification tasks. For example, Lima *et al.* (2018) had participants perform yes/no decisions to indicate whether emotional vocalizations expressed a pre-specified target emotion or not, and reported that performance exceeded chance-level for responses with latency around 300 ms and that high recognition accuracy was reached at 500 ms after stimulus onset.

A more direct approach to study the time-course of auditory stimulus recognition is provided by the gating paradigm, which was originally used to study the time-course of lexical access in spoken word recognition (e.g., Grosjean, 1980). In a typical experiment, the word stimulus is cut into segments, or "gates," that are presented to the listener with a successively increased duration, starting from the onset of the word until the whole word is presented. Stimuli can be segmented on the basis of either absolute durations or linguistic units so that the first gate may comprise, for example, the first 50 ms or the first syllable of the word. After each gate, participants are asked to identify the word, and then the identification point (IP) is calculated (sometimes also called the isolation point). IP is usually defined as the point in time when the participant correctly identifies the word without any changes in response thereafter. Gating studies have also used an estimate of stimulus recognition based on when it is correctly identified above a pre-defined percentage-correct threshold across trials or participants.

The gating paradigm is directly applicable to the study of how much acoustic information is required for a listener to recognize vocal expressions, but only a handful of previous studies have utilized this method for studying the time-course of vocal emotion recognition. In an early study, Pollack *et al.* (1960) reported that recognition accuracy increased as a function of increased gate length, and that overall recognition of 8 "modes of expression" was above chance for utterances with a duration of only 60 ms. The modes of expression did include the emotions fear and happiness, but most of the "modes" were not emotional (e.g., statement, question, confidential communication). Pollack *et al.* (1960) did not report data for individual emotions, which coupled with limited methodological reporting makes it hard to evaluate their results. More recently, Audibert *et al.* (2007) investigated perception of various affective states by dividing single-word utterances into six gates. They reported that satisfaction was recognized with above chance accuracy already from the first gate, whereas resignation did not reach chance level until the last gate.

Five previous studies have reported IPs for individual emotions (so called emotion identification points, or EIPs). Grichkovtsova *et al.* (2007) compared the recognition time-course of emotions (anger, sadness, and happiness) vs "attitudes"

(obviousness, irony and doubt) and reported that anger (range = 800–1600 ms) and sadness (range = 800–1600 ms) had shorter EIPs than happiness (range = 1000 ms—full sentence) and the attitudes. The authors argued that happy expressions might be recognized later because they are used deliberately to color social interaction in a manner similar to attitudes, whereas anger and sadness are expressed more directly and involuntarily. Cornew *et al.* (2010) reported that participants reached the EIP for neutral expressions faster (444 ms) than for angry (723 ms) and happy (802 ms) expressions. Pell and Kotz (2011) in turn reported that neutral (510 ms), fear (517 ms), sadness (576 ms), and anger (710 ms) were recognized earlier than happiness (977 ms) and disgust (1486 ms). In a follow-up study, Rigoulot *et al.* (2013) tested if the slower recognition of happiness and disgust could be explained by the fact that the acoustic cues conveying these emotions were not available until later in the utterance. By reversing the gating procedure so that the last segment of the expression was presented first, followed by the two last segments and so on, they showed that the stimulus duration needed for recognition indeed decreased for happiness and disgust but not for the other emotion expressions.

Finally, Jiang *et al.* (2015) used the gating paradigm to show that emotion recognition was faster when perceivers (native speakers of English or Hindi) judged vocal stimuli produced by native speakers of their own language compared to a foreign language. They reported that EIPs were shortest for neutral (range of average EIPs 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 EIPs for happiness (range = 654–1634 ms). They also reported that recognition accuracy was above chance for all emotions in all language and listener group conditions (except happiness in English) already when the gate duration was 200 ms.

C. Previous studies of the time course of emotion recognition in music

Bigand *et al.* (2005) and Filipic *et al.* (2010) used the gating paradigm to study the time-course of judgments of music as moving or not moving. These studies 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 *et al.* (2005) argued that skilled musicians can shape the acoustic cues of the very first note of a musical performance so that its acoustic characteristics (e.g., intensity and spectral envelope) prefigures the main mood of the piece. The fact that very short fragments of music can contain a lot of acoustic information is further underscored by studies that have used gating to investigate melody and music recognition. For example, when participants were asked to name known melodies played with a piano timbre (Dalla Bella *et al.*, 2003) or to name famous artists/songs (Filipic *et al.*, 2010), results showed that melodies could be recognized after hearing 500 ms and that famous artist/songs could be recognized from music excerpts as short as 300 ms [see also Krumhansl (2010)].

Peretz *et al.* (1998) conducted the first gating study of emotion recognition in music. They showed that participants could reliably distinguish happy from sad excerpts when the stimuli had a duration of 500 ms (which was the shortest gate included). Vieillard *et al.* (2008) developed a set of music stimuli that conveyed happiness, sadness, threat, and peacefulness, and divided the excerpts into gates in a way such that one note corresponded to one gate. They reported that happy excerpts were recognized from shorter excerpts (average duration = 483 ms) than the other emotions: peacefulness (1261 ms), sadness (1446 ms) and threat (1737 ms).

D. Contribution of the current studies

This paper presents two gating studies which investigate the time course of emotion recognition from speech prosody (Study 1) and from music performances (Study 2), respectively. Similar methods are employed in both studies to allow for comparisons of results across expressive modalities.

Negative emotions such as anger, fear, and sadness seem to be recognized faster than happiness in all previous vocal gating studies reviewed above (Cornew *et al.*, 2010; Grichkovtsova *et al.*, 2007; Jiang *et al.*, 2015; Pell and Kotz, 2011; Rigoulot *et al.*, 2013). However, happiness is usually the only positive emotion included in previous vocal gating studies (except for Audibert *et al.*, 2007), which is unfortunate because recent research suggests that vocal expressions can convey a wide range of positive emotions (e.g., Sauter, 2017). Study 1 expands upon previous gating studies by including the widest range of vocal expressions to date: anger, fear, happiness, interest, lust, sadness, serenity, tenderness, and relief (and neutral expressions). This selection enables the investigation of whether the advantage for negative vocal expressions remains when happiness is not the only positive emotional state to be judged.

Previous studies have reported that some emotions were recognized with above-chance accuracy already at the earliest gates that were presented (e.g., at 200 ms in Jiang *et al.*, 2015). To further investigate how much (or rather how little) acoustic information is needed to allow for emotions to be recognized from speech prosody, Study 1 includes shorter gates (starting at 50 ms) with a more fine-grained temporal resolution (increments of 50 ms) compared to previous studies. Following previous studies (e.g., Pell and Kotz, 2011; Rigoulot *et al.*, 2013), Study 1 also presents EIPs to investigate how much acoustic information is needed to recognize different vocal emotions in a stable manner (albeit for a wider selection of emotions).

In Study 2, the same gating design as in Study 1 is instead applied on the recognition of musically expressed emotions. Like studies of vocal expression, recent studies of musically expressed emotions have begun to investigate a wider variety of both positive and negative emotions (e.g., Laukka *et al.*, 2013; Scherer *et al.*, 2017). However, previous gating studies of emotion recognition in music (Peretz *et al.*, 1998; Vieillard *et al.*, 2008) remain very limited in terms of the number of emotion categories under study, as well as the number and duration of gates. Study 2 will therefore provide

information about the amount of acoustic information required for recognition for several novel musical emotions (e.g., anger, fear, interest, and tenderness).

Finally, all stimuli from both Study 1 and 2 are analyzed with regard to a standard set of acoustic parameters relevant for affective computing (see Eyben *et al.*, 2016), and the acoustic patterns that are associated with very quick emotion recognition (i.e., at the shortest gate at which emotions could be recognized) and with more stable emotion recognition (i.e., at the EIP) are presented and compared. These analyses provide initial clues about the acoustic features underlying rapid emotion recognition from speech prosody and music.

II. STUDY 1: TIME COURSE OF RECOGNITION OF VOCAL EXPRESSIONS

Study 1 utilized a gating paradigm to investigate the time needed to recognize a variety of positive and negative emotions (anger, fear, happiness, interest, lust, neutral, sadness, serenity, tenderness, and relief) from vocal expressions. Results from the gating experiment are presented for each emotion in terms of the shortest gates at which each emotion could be accurately recognized, and as EIPs—which represent the amount of acoustic information that is required to recognize each emotion in a stable manner. Study 1 also presents the most prevalent biases and misclassifications for each gate duration and emotion. Finally, acoustic cue patterns associated with each emotion are presented for cues measured at both the shortest gate where emotions were recognized and at the EIP.

A. Materials and method

1. Stimulus construction

a. Recording of vocal expressions. Four native Swedish professional actors (2 females) were recruited to express two 20-syllable sentences in 13 emotional prosodies plus neutral prosody. One sentence was in Swedish (*En gång tvistade nordanvinden och solen om vem av dem som var starkast*) and the other was a nonsense sentence that resembled Swedish (*Enocken lär sjölva, så marginen har ett visserlag mot såteng ferup*). The actors also recorded other materials not used in the present study. To achieve a wide variety of emotional expressions, several states largely agreed to represent basic emotions (anger, disgust, fear, happiness, and sadness, e.g., Ekman, 1992) were included, together with states that are considered as basic by some but not all emotion theorists (contempt, interest, lust, relief, and tenderness; see Tracy and Randles, 2011). In addition, pride and shame, which are considered self-conscious emotions (Tangney and Tracy, 2012), and the low-arousal positive state serenity, were included. Importantly, this selection also included equally many positive and negative emotions.

To control for individual differences with regard to how emotion words are interpreted and following the procedure in Laukka *et al.* (2016), actors were provided with definitions and situations that typically would elicit each emotion, based on the appraisal theory of emotion (e.g., Ellsworth and Scherer, 2003; Lazarus, 1991; Ortony *et al.*, 1988). The actors were encouraged to imagine that they were in a self-experienced situation that was similar to the described situation, and were instructed to portray expressions with both moderate and strong emotion intensity.

Expressions were recorded in a low reflection soundproof room with a Brüel & Kjær type 4190 mono microphone and NEXUS Brüel & Kjær 2690A 0S4 amplifier (Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark) and RME Babyface Pro soundcard (RME Audio, Haimhausen, Germany) onto a computer with a sampling frequency of 48 kHz (16 bits per sample). The distance from the speaker's mouth to the microphone was held approximately constant at 15 cm with the aid of a pop filter.

b. Pre-selection of vocal expressions. To select recordings of expressions with the highest possible recognition rate for each emotion, all recordings ($N = 224$; 4 actors $\times$ 2 sentences $\times$ 2 levels of emotion intensity $\times$ 14 expressions) were validated by 20 participants [mean age = 25.8, standard deviation (SD) = 5.9, range = 18–39] in a forced choice emotion identification task. Recording presentation order and the order of response options (anger, contempt, disgust, fear, happiness, interest, lust, neutral, pride, relief, sadness, serenity, shame, and tenderness) were randomized for each participant. The experiment lasted until either all recordings had been rated or for 1 h if the participant was not yet finished. This procedure resulted in a mean number of 18.74 (SD = 0.99) ratings per recording. PsychoPy software (Peirce, 2007) was used to present recordings and to collect responses.

Four recordings of anger, fear, happiness, interest, lust, neutral, sadness, and serenity, and three expressions of tenderness and relief, were selected on the basis that they were correctly identified by at least 60% of the listeners. Recordings of pride, shame, disgust, and contempt were excluded because too few recordings met this requirement. Thus, 38 recordings were selected for the gating experiment. Ten and seven recordings were from the two female speakers, and ten and 11 were from the two male speakers. The duration of the selected recordings ranged from 3.2 to 10.4 s with a mean of 4.8 s (SD = 1.4).

c. Gate construction. Each recording was sliced into 11 gates resulting in a total of 456 stimuli including the full-length recordings. The five 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. All gated stimuli had a 10 ms fadeout to prevent clicking sounds at the offset (e.g., for the 50 ms stimuli, the fadeout started after 40 ms).

2. Gating experiment

a. Participants. Thirty-one participants (20 women), mean age 26.32 years (SD = 6.65, range 18–49), 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 filled in a self-report inventory about individual differences in musical sophistication after the

experiment was finished. This inventory aims to assess interest in music and ability to engage with music in a nuanced way (Müllensiefen *et al.*, 2014). The mean musical sophistication score was 7.00 out of 15 (SD = 4.22), and 13 participants reported engaging in daily practice on a musical instrument.

b. Experimental procedure. Participants were tested individually in a sound-attenuated room during a single session lasting approximately 1 h. Before starting the experiment, participants were given instructions about the procedure and performed four practice trials (which did not recur in the experiment). They also received a list with definitions of each emotion label (the same ones as previously given to the actors) to control for individual differences in interpretation of emotion labels. Stimuli were presented over headphones (Sennheiser HD 280 pro, Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) at a comfortable listening level controlled by a laptop computer running PsychoPy software (Peirce, 2007).

After each stimulus presentation, a list of response options was displayed on the screen (anger, fear, happiness, interest, lust, neutral, relief, sadness, serenity, and tenderness). The order of response options was randomized and varied for each participant. Participants were instructed to select the emotion label that they thought best fitted the expression conveyed by the stimulus in a forced choice-task. Stimuli were presented in a gate-blocked format in a unique random order for each gate and participant. The shortest gate stimuli were presented first, followed by the second shortest and so on, until all 11 gates had been presented. After that, all full-length recordings were also presented and judged. Participants could, if they wished, press a button to hear each stimulus as many times as needed to reach a decision, but were instructed not to dwell too long on each stimulus.

3. Acoustical analyses

All full-length recordings from each actor expressing the ten selected emotions were acoustically analyzed using the *openSMILE* software (Eyben *et al.*, 2013) to extract the parameters included in the Geneva Minimalistic Acoustic Parameter Set (GeMAPS; see Eyben *et al.*, 2016, for a detailed description). All acoustic parameters were calculated frame-wise with a temporal resolution of 10 ms and the values for each frame were used to calculate means and SDs for each parameter and gate length. These values were subsequently used to derive cue values present at each gate length for each recording. The raw values for each speaker and parameter were *z*-transformed across all recordings of the selected emotions from the speaker in question separately for each gate length. This was done to control for individual differences in baseline values between speakers (e.g., Banse and Scherer, 1996; Juslin and Laukka, 2001).

A selection of GeMAPS-parameters was performed, in order to get a comprehensive but manageable set of acoustic features that represents the acoustic diversity of the stimuli. The selection was based on a principal components analysis (PCA) of the full set of acoustic features ($N = 28$) and the cues with the highest factor loadings and/or interpretability were chosen to represent each factor. All stimuli that were used in both Study 1 and Study 2 were included in the PCA in order to get a set of cues that can be compared across stimulus types (speech, music) and lengths. Parallel analysis (Dinno, 2009) suggested a solution with 11 factors, but no acoustic features had loadings >0.70 for the last factor. Ten acoustic cues covering parameters related to frequency, intensity, and spectral balance were selected and are shown in Table I. Temporal parameters (e.g., speech rate) were not included because the duration of the stimuli was manipulated in the gating design.

To determine which acoustic cues were the most important predictors of the listeners' judgments (i.e., selected emotion) for each emotion and gate, the projpred package in R and the feature selection procedure recommended by Piironen *et al.* (2018) were used. First, a model was fitted with stan (Stan Development Team, 2018a) using the rstanarm package (Stan Development Team, 2018b) with selected emotion as the dependent variable and the ten selected cues from the PCA analysis as predictor variables. As recommended, regularizing horseshoe priors were used on the regression coefficients. Second, the cues were ranked by their importance using approximate leave-one-out cross-validation. This procedure was repeated for all emotions and gates.

4. Statistical analyses

First, a generalized linear mixed model (GLMM) was used to provide robust estimates of emotion recognition accuracy. The hit rates estimated by the GLMM were used as the main measure of emotion recognition accuracy because the model incorporates nesting information available in the repeated measures design of the current study. More specifically, a logistic Bayesian GLMM was fitted with stan (Stan Development Team, 2018a) using the rstanarm package (Stan Development Team, 2018b). In this model, the probability of a hit (defined as trials where the listener selected the emotion intended by the speaker) was estimated with gate length, emotion, sound file, and listener as predictor variables. Gate length (in log-transformed ms) was used as a continuous predictor and varying intercepts and slopes were estimated for each emotion, each sound file (nested within emotion), and each listener (nested within emotion). Priors for the estimated intercepts and slopes were defined as student *t* distributions with seven degrees of freedom centered on zero with standard deviations of 2.5. These priors can be thought of as "weakly informative" on the log-odds scale (Ghosh *et al.*, 2015). This means that they are regularizing but easily outweighed by the data, and mainly serve to facilitate computation of the posterior distributions. The maximum likelihood and the 95% credible intervals (CIs) for each gate length and emotion were computed by first averaging the posterior distributions across sound files and listeners, and then computing the median values and the percentiles covering 95% of these distributions (see the online supplementary R-code for details[1]).

Second, the EIP for each participant and recording was defined as the shortest gate for which the participant selected the intended emotion and then made zero, or only one error

TABLE I. Summary of selected acoustic parameters.

Feature type	Description	Factor loading
Frequency cues		
*F*0 M	Mean fundamental frequency (*F*0) on a semitone frequency scale	Factor 2: 0.91
*F*0 SD	Variability (standard deviation) of *F*0	Factor 3: 0.88
*F*1 M	Mean of first formant (*F*1) center frequency	Factor 6: 0.84
*F*1 SD	Variability (standard deviation) of F1 center frequency	Factor 8: 0.80
*F*3 bandwidth	Mean bandwidth of third formant (F3) center frequency	Factor 10: 0.80
Jitter	Mean of jitter, i.e., deviations in individual consecutive *F*0 period lengths	Factor 7: 0.88
Energy cues		
Int M	Mean voice intensity estimated from an auditory spectrum	Factor 5: 0.78
Spectral balance cues		
Hammarberg M	Mean Hammarberg index, i.e., the ratio of the strongest energy peaks in the 0–2 kHz vs 2–5 kHz regions	Factor 1: 0.88
Hammarberg SD	Variability (SD) of Hammarberg index	Factor 4: 0.83
Spectral slope SD	Variability (SD) of spectral slope (i.e., linear regression slope of the logarithmic power spectrum) for the 0–500 Hz region	Factor 9: 0.74

for longer gates (see Pell and Kotz, 2011). The EIP for each recording was then defined as the median across all participants, and the EIP for each emotion category was defined as the mean across the medians for all recordings within that emotion category.

In addition, the proportion of participants who selected each emotion category was calculated for each gate-duration for each recording as a measure of confusions between intended and selected emotions. To further account for response biases, the unbiased hit rate (Hu) was calculated for each participant and condition, using equations presented in Wagner (1993). The unbiased hit rate gives a measure of perceptual sensitivity and is defined as "the joint probability both that a stimulus is correctly identified (given that it is presented) and that a response is correctly used (given that it is used)" (Wagner, 1993, p.16). Hu ranges from 0 to 1, where a score of 1 indicates that all stimuli of an emotion have been correctly classified and that the respective emotion has never been misclassified as another emotion.

B. Results

1. Recognition accuracy, EIP, and common confusions for vocal expressions

Figure 1 gives a comprehensive picture of the time course of emotion recognition in Study 1. The estimated hit rate—which is the main measure of emotion recognition and represents the proportion of participants who selected the

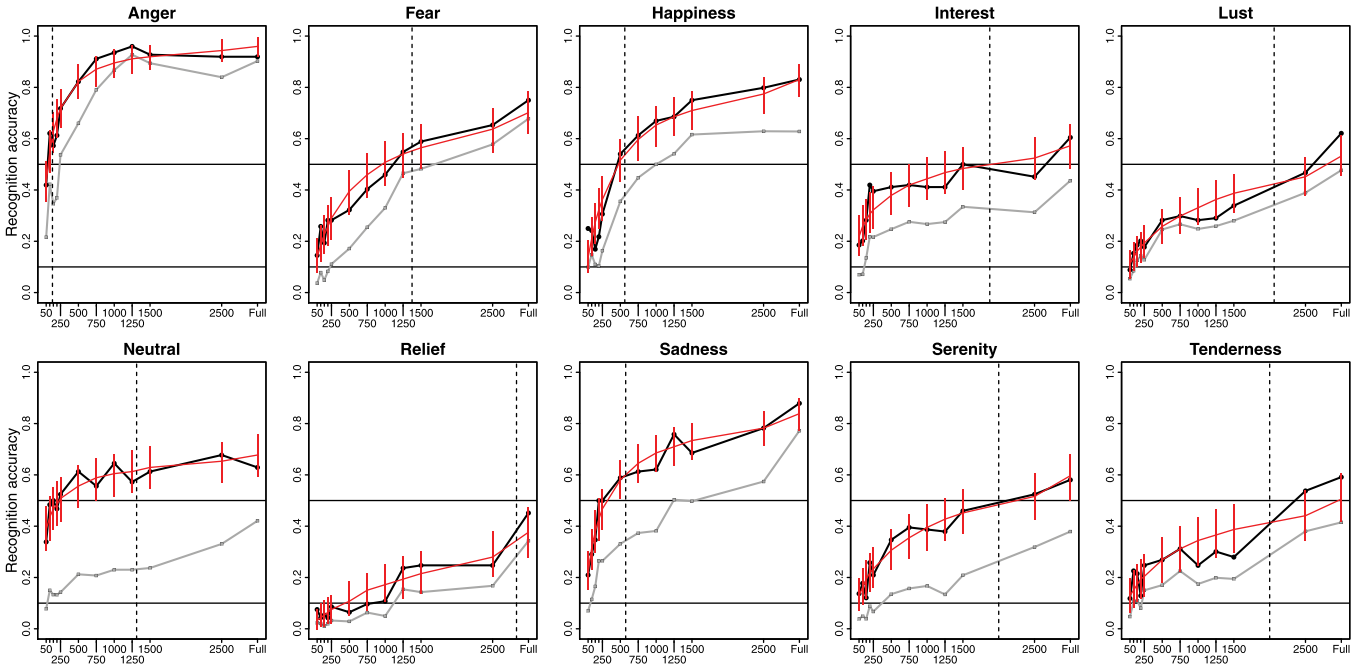

FIG. 1. (Color online) Recognition accuracy for vocal expressions as a function of gate duration for each emotion in separate panels. The thin (red) line with error bars depicts the estimated recognition accuracy and 95% CIs. The thick black line and black points depict the observed proportion of participants who selected the intended emotion label. The horizontal thin black lines mark the chance level (0.10) and the majority level (0.50). The gray line and gray points depict the joint probability measure (Hu). The vertical dashed line depicts the emotion identification point (EIP).

intended emotion label as estimated by the GLMM—is shown for each emotion as red lines with error bars depicting 95% CIs. For the sake of completeness, the observed proportion of participants who selected the intended emotion label (observed hit rate) is also shown as black points and lines. Recognition accuracy improved over the time-course of an utterance and some emotions were generally recognized at an earlier gate-duration than others. Within each emotion category, some stimuli were also recognized at earlier gate-durations than others. Recognition rates are presented separately for each speech recording in the supplementary material.[1]

Regarding the shortest gate at which each emotion could be recognized, Fig. 1 shows that the estimated hit rates for anger, interest, neutral, and sadness were higher than expected by chance (i.e., CIs not overlapping the level of chance responding, which is 0.10 in a 10-alternative forced choice task) when only 50 ms of the utterances were presented. Looking at the CIs for later gate-durations suggests that fear, happiness, and serenity had estimated hit rates with above-chance accuracy at 100 ms, and lust and tenderness at 150 ms. Relief reached above-chance accuracy much later than the other emotions at 1250 ms.

The EIPs—which represent more stable emotion recognition—are also presented in Fig. 1 for each emotion category (as vertical dashed lines). Anger had the shortest EIP (138 ms), followed by happiness (562 ms), and sadness (575 ms). Neutral (1312 ms) and fear (1375 ms) had intermediate EIPs and the longest EIPs were obtained for interest (1875 ms), lust (2062 ms), tenderness (2000 ms), serenity (2000 ms), and relief (2833 ms).

Wagner's (1993) joint probability measure (Hu) is also presented in Fig. 1, as a measure of perceptual sensitivity (gray squares and lines). Hu corrects accuracy rates for participants' response biases (e.g., the tendency to use some response alternatives more frequently than others), and Hu-values can therefore be compared with the uncorrected hit rate to get a sense of the degree of bias. The largest difference between Hu and uncorrected hit rate was observed for neutral, which suggests that recognition rates for neutral may have been affected by response bias. A more detailed look at response biases is available in Table II, which presents the most frequent confusions (or misclassifications) for each intended emotion category and gate duration. This table suggests that confusions were more prevalent at shorter gate-durations compared to longer gate-durations for most emotions. Fear was most commonly confused with sadness; happiness with fear (for short gates) and interest (for longer gates); lust with serenity; and sadness with fear, serenity, and neutral. Interest, serenity, tenderness, and relief were instead most commonly confused with neutral; and neutral was most commonly confused with serenity. Anger was rarely mistaken for other emotions, but the most common confusions were with happiness (at 50 ms) and fear.

2. Acoustic correlates of expressed and perceived vocal emotions

Table III presents an overview of the acoustic cue patterns associated with each emotion (see Table I for a description of the acoustic cues). First, the point-biserial correlations between the speakers' intended expression (dichotomously dummy-coded as 1 or 0) and the acoustic cue values (z-scores) were calculated separately for each emotion at (a) the shortest gate where recognition rates reached above-chance levels, and (b) at the EIP. These correlations give indications of the degree to which each cue was used by the speakers to express each emotion. Second, the correlations between listeners' average emotion judgments (continuous proportion of participants who selected each emotion) and cue values (z-scores) were calculated for each emotion, also separately for the shortest gate with above-chance accuracy and the EIP. These correlations in turn give indications of the degree to which each cue was used by the listeners on a consistent basis to make judgments about perceived emotions. For sake of completeness, acoustic correlates of perceived vocal emotions are presented separately for each gate length in the supplementary material.[1]

The major trends are summarized below by focusing on the cues which were used by both speakers and listeners and showed the largest effects ($r >$ approx. 0.30). $F0$M and $F0$SD (mean and SD of fundamental frequency, respectively) were positively correlated with anger and happiness, whereas $F0$M was negatively correlated with serenity and tenderness. Happiness was positively correlated with $F1$M

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

TABLE III. Acoustic cue patterns associated with each emotion for both expressed and perceived vocal expressions, and for both the shortest gate where recognition rates reached above-chance level and at the EIP. Bold font indicates correlations ≥0.3 and highlighted cells indicate that this cue was ranked among the top three cues in the feature selection analysis. See Table I for a description of the acoustic cues. *Correlation type indicates if the displayed *r*-values refer to expressed or perceived emotions. "Expressed" = the point-biserial correlations between acoustic cues and the speakers' intended expression dichotomously dummy-coded as 1 or 0. "Perceived" = the Pearson correlations between acoustic cues and the proportion of listeners who selected that emotion.

Emotion	Gate	Correlation type*	R2	*F*0 M	*F*0 SD	*F*1 M	*F*1 SD	*F*3 bandwidth	Hammarberg M	Hammarberg SD	Jitter	Int M	Spectral slope SD
anger	50	expressed		**0.35**	−0.10	0.17	0.24	**0.31**	**−0.70**	−0.14	−0.07	**0.84**	−0.13
		perceived	0.83	**0.43**	−0.07	**0.35**	0.27	**0.47**	**−0.64**	−0.06	0.02	**0.87**	−0.12
	150	expressed		**0.42**	−0.16	0.02	0.28	**0.40**	**−0.64**	**0.34**	−0.17	**0.74**	0.17
		perceived	0.84	**0.61**	−0.21	0.08	0.27	**0.35**	**−0.54**	0.27	−0.25	**0.90**	0.25
fear	100	expressed		0.11	0.00	0.19	0.22	−0.04	**0.30**	0.24	0.25	−0.09	0.02
		perceived	0.07	0.18	0.05	**0.37**	0.16	−0.26	−0.14	−0.01	0.13	0.08	0.17
	1500	expressed		0.15	0.28	0.15	0.19	0.13	−0.10	0.20	−0.03	−0.02	**−0.36**
		perceived	0.20	0.19	**0.41**	0.25	0.23	0.13	−0.12	0.06	−0.13	0.00	**−0.38**
happiness	100	expressed		**0.35**	−0.13	0.27	0.10	0.08	0.03	−0.04	−0.11	0.19	0.20
		perceived	0.18	**0.40**	−0.12	**0.35**	**0.40**	0.05	**−0.30**	0.09	−0.15	**0.35**	0.24
	750	expressed		**0.40**	0.19	0.28	0.27	0.23	−0.18	−0.13	−0.05	0.18	0.09
		perceived	0.12	**0.36**	0.16	0.29	**0.33**	0.04	−0.24	−0.11	−0.01	0.15	**0.32**
interest	50	expressed		−0.04	−0.05	−0.16	0.06	−0.08	0.12	0.14	0.09	−0.08	0.24
		perceived	0.06	0.29	−0.27	−0.27	−0.02	0.04	−0.01	−0.13	−0.13	**0.31**	0.12
	2500	expressed		−0.06	0.18	−0.07	−0.03	−0.04	0.18	0.15	**0.30**	−0.17	0.04
		perceived	0.02	−0.02	0.21	−0.06	0.00	−0.01	0.08	0.04	0.18	−0.12	0.06
lust	150	expressed		**−0.40**	**0.63**	0.00	−0.01	−0.09	−0.06	−0.14	−0.06	−0.15	0.09
		perceived	0.07	−0.07	−0.06	0.17	**0.35**	−0.08	0.13	−0.23	−0.12	−0.18	−0.09
	2500	expressed		−0.18	−0.14	0.02	−0.17	−0.18	−0.05	0.12	−0.08	−0.14	0.01
		perceived	0.02	−0.13	−0.12	0.02	−0.02	−0.17	−0.02	0.17	−0.06	−0.15	0.00
neutral	50	expressed		−0.11	−0.07	−0.15	0.00	−0.06	−0.04	−0.06	**0.46**	−0.12	−0.10
		perceived	0.18	**−0.39**	0.04	−0.28	**−0.31**	−0.05	0.28	−0.07	0.04	**−0.47**	−0.23
	1500	expressed		−0.22	**−0.41**	−0.28	−0.14	0.17	0.11	−0.09	0.21	−0.14	0.00
		perceived	0.38	**−0.49**	**−0.46**	**−0.58**	**−0.54**	−0.10	**0.38**	−0.04	0.07	**−0.34**	0.13
relief	1250	expressed		0.02	−0.07	**−0.30**	−0.22	**−0.32**	−0.05	−0.25	−0.10	0.06	**0.36**
		perceived	0.07	0.03	0.11	−0.07	−0.07	−0.26	−0.22	−0.23	−0.14	0.05	**0.39**
	Full	expressed		−0.04	−0.11	−0.11	0.17	**−0.36**	−0.10	−0.19	−0.15	0.02	**0.37**
		perceived	0.07	−0.11	−0.10	−0.14	0.17	**−0.37**	−0.11	−0.14	−0.21	−0.05	**0.30**
sadness	50	expressed		0.07	−0.09	0.01	−0.24	0.01	0.07	−0.06	−0.16	−0.15	**0.32**
		perceived	0.22	−0.21	**0.42**	−0.15	0.02	−0.12	0.17	0.17	0.23	**−0.46**	−0.03
	750	expressed		−0.12	−0.03	0.05	−0.03	−0.24	0.15	−0.04	0.22	−0.23	−0.03
		perceived	0.06	−0.23	0.05	0.04	0.02	−0.02	−0.02	0.25	**0.31**	**−0.30**	−0.10
serenity	150	expressed		−0.13	−0.18	−0.04	−0.15	−0.03	0.11	0.00	−0.15	−0.18	−0.22
		perceived	0.32	**−0.48**	0.18	**−0.39**	−0.29	−0.09	**0.45**	0.21	0.15	**−0.52**	0.11
	2500	expressed		**−0.46**	−0.08	−0.02	−0.02	0.14	0.28	0.28	−0.07	−0.28	0.07
		perceived	0.42	**−0.64**	−0.27	−0.23	**−0.33**	0.06	**0.48**	**0.41**	−0.06	**−0.46**	0.17
tenderness	150	expressed		−0.25	−0.06	0.06	−0.08	**−0.36**	0.23	0.08	−0.18	−0.28	−0.25
		perceived	0.25	**−0.39**	0.05	−0.06	−0.04	**−0.42**	**0.39**	0.02	−0.10	**−0.37**	**−0.33**
	2500	expressed		−0.27	0.20	−0.14	−0.25	−0.07	0.17	0.07	−0.29	−0.26	−0.01
		perceived	0.13	**−0.39**	0.16	−0.25	**−0.37**	−0.11	0.17	0.02	−0.23	−0.36	−0.05

and *F*1SD (mean and SD of the first formant), whereas negative correlations were observed between *F*1M and neutral. *F*3 bandwidth (the mean bandwidth of formant 3) was positively correlated with anger and negatively correlated with relief. The highest positive correlations with IntM (mean voice intensity) were observed for anger followed by happiness, whereas the highest negative correlations with IntM were observed for neutral, sadness, serenity, and tenderness.

Regarding spectral balance cues, the mean Hammarberg index was negatively correlated with anger and positively correlated with serenity. Spectral slope variability was positively correlated with relief for both gate lengths, and also negatively correlated with fear (but only at the EIP). The relation between spectral measures and perceived voice quality (and the underlying voice source variation) is not straightforward (e.g., Gobl and Ní Chasaide, 2010). However, the results for anger are in line with what would be expected from a tense (or possibly harsh) voice quality, whereas results for low-arousal positive emotions such as serenity are instead in line with a softer voice quality.

Overall, the acoustic cue patterns were similar at both the shortest gate with above-chance recognition and the EIP. For expressed emotion, the correlation between cue patterns at the "shortest gate" and the EIP was $r = 0.65$, and for perceived emotions the correlation was $r = 0.72$ (correlations were calculated across all 10 selected cues for each of the

8 emotions, $N = 80$, using values displayed in Table III). Table III also presents the results from the feature selection analyses (see Piironen *et al.*, 2018) where listeners' judgments of each emotion were predicted from acoustic cues. Usually the cues with the highest correlations (as described above) were also ranked as the most important predictors. Anger was by far the best predicted emotion, followed by serenity, neutral, and happiness. Except for anger, the proportion of explained variance was modest, and was especially low for interest and lust.

C. Discussion

Results from the recognition experiment showed converging results regarding both the shortest gate-duration for accurate recognition and the EIPs. In general, accurate recognition was observed from very brief stimuli, but recognition rates also increased with increasing gate duration. Anger was recognized the fastest with above chance accuracy already at the shortest gate (50 ms) and an average EIP of 138 ms. Interest, neutral, and sadness were also recognized from the shortest gate, and all emotions, except relief, were in fact accurately recognized from stimuli that had a duration of 150 ms or less. This suggests that communication of emotion from speech prosody is a very fast process that allows discrimination of several emotions based on low level physical characteristics that can be rapidly expressed and detected within a few glottal cycles (e.g., level of *F*0 and voice intensity, and voice quality) and do not need to develop over time. Relief was an obvious exception and did not reach above chance accuracy until stimulus duration was well over 1 s, and may thus be especially reliant on cues that either develop over time or appear toward the end of an utterance (e.g., Rigoulot *et al.*, 2013).

EIPs reflect the amount of acoustic information that is required to recognize each emotion in a more stable manner. Apart from anger, happiness (562 ms) and sadness (575 ms) also had brief EIPs, which indicates that stable recognition was achieved fast. For the other emotions, however, EIPs were not reached until stimulus duration was above 1 s (fear, interest, and neutral) or even 2 s (lust, relief, serenity, and tenderness). These results may suggest that acoustic cues that need some time to develop (e.g., *F*0 contours) also contribute to listeners' emotion judgments, at least for some emotions.

A comparison of the current EIPs with previous studies (Cornew *et al.*, 2010; Grichkovtsova *et al.*, 2007; Jiang *et al.*, 2015; Pell and Kotz, 2011), reveals that anger was recognized much faster in the current study. This may have to do with methodological differences between studies. For example, the current study employed shorter gate durations, which may have allowed a more fine-grained measurement of the EIP for anger, than previous efforts. Some previous studies have also normalized the sound level of the stimuli (e.g., Pell and Kotz, 2011; Rigoulot *et al.*, 2013), whereas the current study opted not to do that. Because (high) sound level has been reported to be particularly relevant for anger expressions in many studies (see Juslin and Laukka, 2003), the preserved between-stimulus sound level variability may have facilitated recognition of anger in particular. Happiness was also recognized somewhat faster than in previous studies—which seems to go against the proposition that negative emotions should be recognized faster than positive ones (based on evolutionary arguments about the importance of avoiding danger, e.g., Pell and Kotz, 2011). However, looking at the whole variety of positive emotions in the current study, it can be noted that EIPs for lust, interest, relief, serenity, and tenderness were all longer ($\geq$1875 ms) than for any of the included negative emotions. This seems to indicate that recognition of negative emotions was faster than for positive emotions (with the exception of happiness), but this conclusion may be premature because some negative emotions with potentially longer EIPs (e.g., disgust; see Pell and Kotz, 2011) were excluded from the study because they were not well recognized in the listening test for the initial selection of recordings.

Neutral and fear expressions instead had longer EIPs in the current study vs previous studies, whereas EIPs for sadness were in the same range as in previous studies (Pell and Kotz, 2011; Jiang *et al.*, 2015). The longer EIPs for neutral and fear could be due to, for example, the fact that the current study used more (and different) response options in the forced-choice task, which may have required different types of decisions with less opportunity for guessing the "correct" answer. There was also response bias for neutral judgments, suggesting that participants may have responded with "neutral" when they were uncertain of their choice, which calls for caution when interpreting recognition rates for neutral in the current study. In general, it is difficult to directly compare results across studies, because the results of each individual study are based on a limited number of emotional utterances which differ with regard to many characteristics (e.g., semantics, speech prosody, emotional nuances) that may affect the timing of emotion recognition. More research is clearly needed to determine if there is such a thing as the "typical" amount of information that is required to recognize different emotions.

A comparison of the cue correlations in Table III with previous review articles (Scherer, 2003; Juslin and Laukka, 2003) and more recent empirical studies (Bänziger *et al.*, 2012; Laukka *et al.*, 2016; Scherer, 2019, Table 4.3) suggests a high degree of consistency for the cues which showed the largest effects in the present study. This indicates that the cue patterns that were used by speakers and listeners, at both the shortest gate with above-chance recognition and at the EIP, were similar to those that have previously been reported for full-length expressions of anger, fear, happiness, interest, lust, neutral, relief, sadness, and tenderness (no previous studies were available for serenity or for *F*3 bandwidth and spectral slope variability). Thus, much of the emotion relevant low-level acoustic information that can be observed in full-length utterances is present already in the first few hundred milliseconds of speech utterances, and may be used by listeners to draw emotion inferences.

Although recognition accuracy was higher at the EIP vs the shortest gate with above-chance recognition, acoustic correlates did not become more pronounced as gate duration increased. It may be that low-level acoustic measures do not

capture the additional acoustic information available in the longer speech stimuli. A wider selection of cues which are sensitive to time-varying information (e.g., measures of rate of change) could shed more light on which cues lie behind the rise in recognition accuracy as stimulus duration increases. However, it may also be the case that the gating method is not well suited for studying the temporal development of specific acoustic cues over time, because cues measured at longer gates by design contain all of the information also contained in cues measured at shorter gates. To conclude, the current results give initial clues about the acoustic features that enable rapid emotion recognition from speech, but one should keep in mind that the selected cues may capture only a part of the wealth of acoustic information that is contained in even very brief speech stimuli. After all, the results from the feature selection analyses showed that the selected acoustic cues could only explain a modest proportion of variance in listener judgments for many of the included emotions.

III. STUDY 2: TIME COURSE OF RECOGNITION OF MUSICAL EXPRESSIONS

Study 2 utilized a similar gating paradigm as in Study 1 to investigate the time needed to recognize a variety of positive and negative emotions (anger, fear, happiness, interest, neutral, sadness, serenity, and tenderness) from musical excerpts played on the violin, viola, and cello. String instruments were used because they allow musicians a large degree of freedom in shaping the intonation, amplitude, and spectral balance of the tone (e.g., Askenfelt, 1991; Schoonderwaldt, 2009). Results from the gating experiment are presented in terms of (a) the shortest gate at which each emotion could be accurately recognized, (b) as EIPs, and (c) as the most prevalent confusions for each gate duration and emotion. In addition, the acoustic cue patterns that were present at both the shortest gate of recognition and the EIP are described.

A. Materials and method

1. Stimulus construction

a. Recording of music expressions. Three professional string musicians (violin, viola, cello) were recruited to express the same emotions as in Study 1. Recordings were made with the same equipment and a similar procedure as for the actors. First, the musicians were instructed to use improvised melodies that they thought best expressed each emotion. Then they were instructed to use a "standard melody" to express the same emotions. The standard melody was composed for the purposes of the current study and was intended to be emotionally neutral (notation is available in the supplementary material[1]). The musicians self-selected when they were satisfied with their recordings for inclusion in the present study. They also recorded other materials not used in the present study.

b. Pre-selection of music stimuli. To select recordings of expressions with the highest possible recognition-rate for each emotion, 291 recordings were validated by 28 participants (mean age = 24.7, SD = 5.8, range = 18.0–45.0) in a forced choice emotion identification task using the same procedure as in Study 1. This procedure resulted in a mean number of 19.4 (SD = 1.9) ratings per recording.

Four recordings of happiness, interest, serenity, fear, sadness, anger, and neutral, and three recordings of tenderness were selected on the basis that they were correctly identified by at least 30% of the listeners. Contempt, disgust, lust, pride, relief, and shame were excluded because too few recordings met this requirement. Thus, 31 recordings were selected for the gating experiment. Eight of the selected recordings were performed on the violin, 12 on the viola, and 11 on the cello. The duration of the selected recordings ranged from 3.6 to 12.9 s with a mean of 7.5 s (SD = 3.2).

c. Gate construction. The selected musical recordings were sliced into 11 gates (50, 100, 150, 200, 250, 500, 750, 1000, 1250, 1500, and 2500 ms) using the same procedure as in Study 1. This procedure resulted in 372 stimuli including the full-length recordings.

2. Gating experiment

a. Participants and procedure. Thirty-two participants (23 women), mean age 27.2 years (SD = 8.5, range 19–61), were recruited via an identical advertisement on the same bulletin board and via the same web page as in Study 1. No participants reported any hearing difficulties. Similar to Study 1, participation was based on informed consent and was rewarded with course credits or movie vouchers. The mean musical sophistication score was 6.28 out of 15 (SD = 3.90) and 18 participants reported engaging in daily practice on a musical instrument (see Müllensiefen *et al.*, 2014). Overall, participants in Study 2 were recruited the same way and had a similar distribution of age, gender, and musical sophistication scores as the participants in Study 1.

Participants judged the expression of each musical stimulus using the same procedure as in Study 1, with the exception that the number of emotion categories in the forced-choice task was eight (anger, fear, happiness, interest, neutral, sadness, serenity, and tenderness) compared to ten in the speech task.

3. Statistical and acoustical analyses

Measures of recognition accuracy—the proportion of participants who selected the intended emotion label, as estimated by a logistic Bayesian GLMM, and accompanying 95% CIs—and EIPs were calculated in the same way as in Study 1.

All music stimuli were analyzed using the same procedure (Eyben *et al.*, 2013, 2016) and set of acoustic parameters as in Study 1, to allow comparisons across studies. Although the sound production mechanisms obviously differ between the voice and string instruments, the underlying physical principles can be similarly explained by the source-filter theory (e.g., Askenfelt, 1991) and many acoustic parameters originally developed for speech analysis are also suitable for describing string instrument sounds (e.g., Mores, 2017). See Table I for an explanation of each of the ten acoustic parameters. Similar to Study 1, the raw values for each musician and parameter were z-transformed across the

two best recognized recordings of each of the selected emotions from the musician in question (to control for individual differences in baseline values between instruments).

B. Results

1. Recognition accuracy, EIP, and common confusions for musical expressions

Figure 2 gives a comprehensive picture of the time course of emotion recognition in Study 2. The estimated hit rate (i.e., the proportion of participants who selected the intended emotion label as estimated by the GLMM) is shown for each emotion as red lines with error bars (95% CIs). The observed hit rates (i.e., the observed proportion of participants who chose the intended emotion label) are also shown as black points and lines for sake of completeness. Overall, emotion recognition improved over the time-course of a musical excerpt and some emotions were generally recognized at an earlier gate-duration than others. There was also variability within each emotion category, as some recordings were recognized at an earlier gate-duration than others. Recognition rates are presented separately for each musical recording in the supplementary material.[1]

Figure 2 shows that the estimated hit rates for anger, happiness, and neutral were higher than expected by chance (i.e., CIs not overlapping the level of chance responding, which is 0.125 in an 8-alternative forced-choice task) when only 50 ms of the musical excerpt was presented. CIs for the later gate-durations indicate that sadness reached estimated hit rates with above-chance accuracy at 100 ms, serenity at 150 ms, fear at 200 ms, tenderness at 250 ms, and interest at 500 ms.

The EIPs for each emotion category are also presented in Fig. 2 (as vertical dashed lines). Happiness had the shortest EIP (1281 ms), followed by anger (1481 ms), fear (1500 ms), and sadness (1656 ms). Tenderness (2000 ms), interest (2250 ms), neutral (2750 ms), and serenity (2875 ms) had longer EIPs.

Finally, Wagner's (1993) joint probability score (Hu) is shown in Fig. 2 (as gray squares and lines) for each emotion. The largest differences between Hu and uncorrected hit rate were observed for anger and neutral, for short gates, and for sadness across all gates. This suggests that recognition rates for these emotions may have been affected by response bias. Table IV gives a more detailed picture of response biases, and presents the most frequent confusions for each intended emotion category. Main confusions included anger being confused with fear (for short gates) and interest (for longer gates); fear with neutral, sadness, and anger; and sadness with neutral, serenity, and tenderness. Happiness was confused with fear in the first gate (50 ms) but with interest in the other gates. Interest was confused with anger and happiness; serenity with neutral and sadness; tenderness with happiness, sadness, neutral, and serenity; and neutral with sadness and serenity.

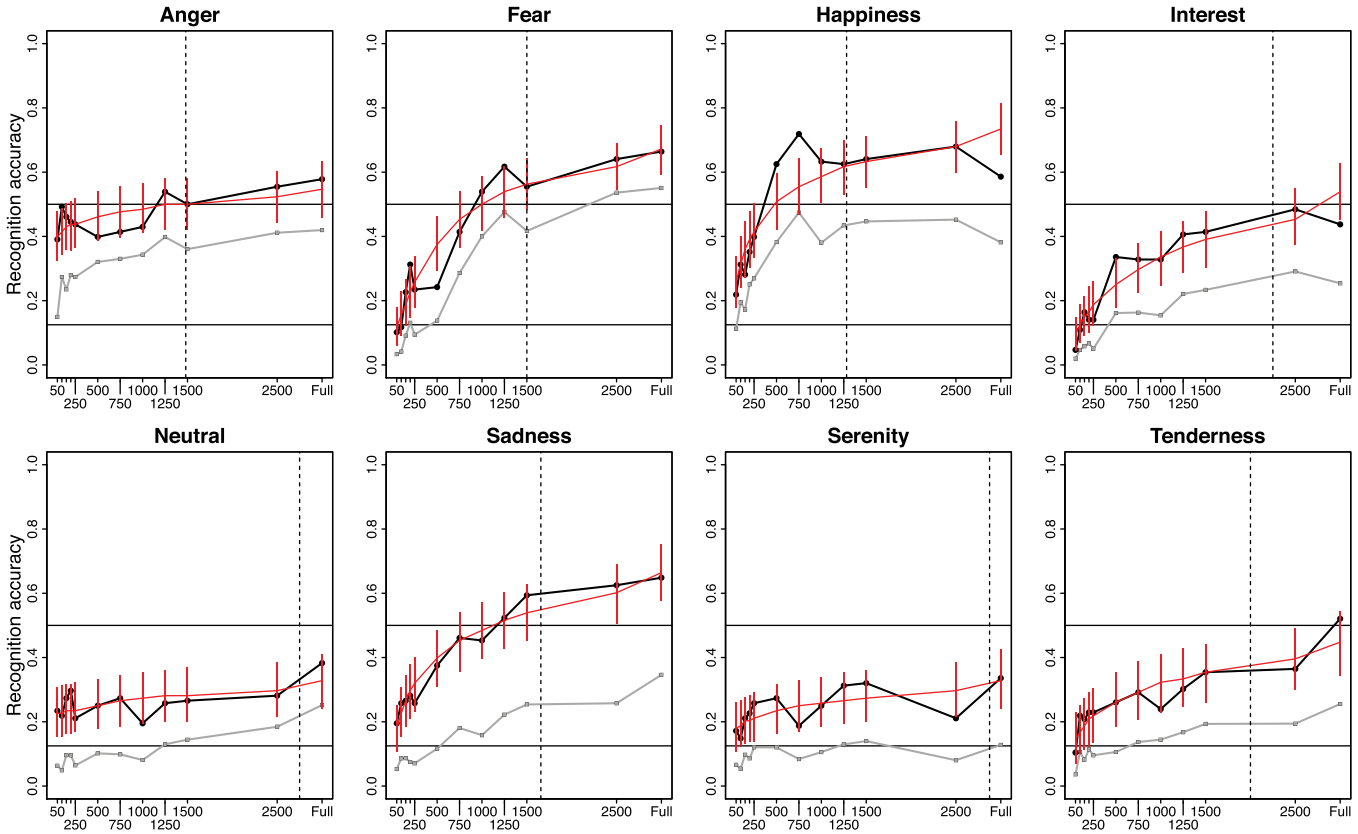

FIG. 2. (Color online) Recognition accuracy for musical expressions as a function of gate duration for each emotion in separate panels. The thin (red) line with error bars depicts the estimated recognition accuracy and 95% CIs. The thick black line and black points depict the observed proportion of participants who selected the intended emotion label. The horizontal thin black lines mark the chance level (0.125) and the majority level (0.50). The gray line and gray points depict the joint probability measure (Hu). The vertical dashed line depicts the emotion identification point (EIP).

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

2. Acoustic correlates of expressed and perceived musical emotions

Table V presents an overview of the acoustic cue patterns associated with each emotion (acoustic cues are described in Table I). Like in Study 1, correlations were calculated between the acoustic cue values (z-scores) for each musical stimulus (at both the shortest gate with above-chance accuracy and at the EIP) and the speakers' intended emotions (dichotomously dummy coded as 1 or 0) and the listeners' average emotion judgments. These correlations give indications of the degree to which each cue was used by speakers (to express emotions) and listeners (to make emotion inferences) and the main findings are summarized below (focusing on the cues which showed the largest effects, $r >$ approx. 0.30, across conditions). For sake of completeness, acoustic correlates of perceived musical emotions are presented separately for each gate length in the supplementary material.[1]

*F*0M (mean fundamental frequency) was positively correlated with happiness and negatively correlated with sadness, and *F*0SD (standard deviation of fundamental frequency) was positively correlated with interest. Happiness was positively correlated with *F*1M (mean of the first formant). For IntM (mean intensity), the highest positive correlations were observed for anger, whereas the highest negative correlations were observed for sadness and serenity. Regarding spectral balance cues, mean Hammarberg index was negatively correlated with happiness (but only for the shortest gate) and positively correlated with interest and neutral (for both gate lengths). Spectral slope variability was negatively correlated with sadness.

Compared to speech (Study 1), cue patterns for music were less consistent across gate lengths. For expressed emotion the correlation between cue patterns at the shortest gate with above-chance accuracy and the EIP was $r = 0.57$, and for perceived emotions the corresponding correlation was $r = 0.43$ (correlations were calculated for all 10 selected cues for each of the 8 emotions, $N = 80$, using the values shown in Table V). Table V also presents results from the feature selection analyses where listeners' judgments of each emotion were predicted from acoustic cues. As for speech, anger was the best predicted emotion, but also interest, sadness, and serenity were relatively well predicted. Overall, the proportion of explained variance in listener judgments was higher for music than for speech.

C. Discussion

Study 2 provided novel information about the time course of emotion recognition from music, using a larger selection of emotions and shorter gate durations than previous studies (Peretz *et al.*, 1998; Vieillard *et al.*, 2008). Anger, happiness, and neutral were recognized especially fast and reached above-chance accuracy already at the shortest gate (50 ms), followed by sadness at 100 ms. Similar to Study 1, these results suggest that very short musical excerpts contain enough low-level acoustic information to allow for discrimination between different emotions. The EIPs—which reflect the amount of acoustic information needed for more stable emotion recognition—followed a similar trajectory with the shortest EIPs observed for happiness, anger, fear, and sadness. However, acoustic cues that develop over time may play a larger role for music than for speech, as emotion recognition took longer to stabilize (indicated by longer EIPs) in Study 2 vs Study 1. A more detailed comparison of results follows below. Overall, there were many similarities both regarding emotion recognition and acoustic cues, although some differences between modalities are also noted.

1. Emotion recognition

Overall, recognition accuracy was lower for musical expressions ($M = 0.52$, $SD = 0.20$) compared to vocal expressions ($M = 0.69$, $SD = 0.20$) for the full length stimuli. A couple of emotions (i.e., lust and relief) could not be included for musical expressions due to low recognizability during the initial selection of recordings, whereas they were relatively well recognized from vocal expressions. However, anger, fear, happiness, and sadness were the best recognized emotions at full-length for both vocal (Fig. 1) and musical (Fig. 2) expressions. Looking at the shortest gate at which emotions were recognized with accuracy above chance, anger, happiness, and sadness were recognized already when

TABLE V. Acoustic cue patterns associated with each emotion for both expressed and perceived musical expressions, and for both the shortest gate where recognition rates reached above-chance level and at the EIP. Bold font indicates correlations ≥0.3 and highlighted cells indicate that this cue was ranked among the top three cues in the feature selection analysis. See Table I for a description of the acoustic cues. *Correlation type indicates if the displayed *r*-values refer to expressed or perceived emotions. "Expressed" = the point-biserial correlations between acoustic cues and the speakers' intended expression dichotomously dummy-coded as 1 or 0. "Perceived" = the Pearson correlations between acoustic cues and the proportion of listeners who selected that emotion.

Emotion	Gate	Correlation type*	R2	*F*0 M	*F*0 SD	*F*1 M	*F*1 SD	*F*3 band width	Hammar berg M	Hammar berg SD	Jitter	Int M	Spectral slope SD
anger	50	expressed		0.27	−0.14	0.08	−0.01	−0.11	−0.23	0.23	0.03	**0.87**	0.02
		perceived	0.40	**0.46**	−0.25	−0.09	0.18	0.07	−−0.16	**0.39**	0.29	**0.61**	**0.36**
	1500	expressed		−0.16	−0.08	0.03	−0.10	0.11	0.15	0.19	0.26	**0.84**	−0.06
		perceived	0.67	−0.24	−0.04	−0.06	0.06	0.26	0.23	0.26	**0.35**	**0.75**	0.10
fear	200	expressed		−0.09	0.21	0.06	0.21	−0.24	−0.20	−0.23	−0.07	−0.07	0.19
		perceived	0.32	0.20	−0.14	0.22	0.28	**−0.30**	**−0.43**	−0.24	0.19	**0.45**	**0.30**
	1500	expressed		0.12	0.22	0.02	0.09	−0.03	−0.27	−0.20	0.11	−0.02	0.15
		perceived	0.02	−0.01	0.24	−0.03	0.04	−0.01	−0.15	−0.20	0.06	−0.05	−0.04
happiness	50	expressed		**0.49**	−0.21	0.28	**0.31**	−0.15	**−0.40**	−0.19	−0.03	0.21	0.24
		perceived	0.14	0.29	−0.12	0.28	−0.05	**−0.43**	**−0.47**	−0.15	0.02	0.12	0.13
	1500	expressed		**0.40**	−0.12	**0.42**	0.02	−0.06	−0.02	0.27	−0.15	0.09	**0.35**
		perceived	0.45	**0.32**	−0.15	**0.36**	0.07	0.16	0.20	**0.52**	0.25	0.10	**0.42**
interest	500	expressed		0.08	−0.20	0.01	0.05	0.20	**0.36**	**0.46**	**0.40**	0.03	**0.52**
		perceived	0.49	**0.39**	−0.22	0.08	**0.40**	0.17	0.18	**0.58**	0.26	0.25	**0.66**
	2500	expressed		−0.19	**0.38**	−0.14	0.14	**0.40**	**0.53**	**0.60**	**0.50**	−0.16	0.09
		perceived	0.50	0.02	**0.43**	−0.01	0.17	**0.30**	**0.46**	**0.66**	**0.38**	0.03	0.21
neutral	50	expressed		0.02	**0.35**	0.26	−0.17	0.07	0.20	−0.13	−0.17	−0.22	−0.26
		perceived	0.15	**−0.38**	0.22	0.00	0.06	0.27	**0.40**	−0.05	−0.06	**−0.47**	−0.22
	Full	expressed		−0.26	−0.07	−0.22	−0.10	**−0.31**	0.20	0.02	−0.17	−0.23	0.09
		perceived	0.06	−0.26	−0.03	−0.24	−0.10	−0.20	**0.35**	0.18	−0.12	−0.14	0.16
sadness	100	expressed		**−0.38**	−0.03	−0.19	0.18	0.04	0.14	−0.04	−0.20	−0.26	−0.15
		perceived	0.38	**−0.53**	**0.39**	−0.08	**−0.38**	**0.31**	**0.36**	**−0.37**	−0.18	**−0.56**	**−0.36**
	2500	expressed		−0.29	−0.05	−0.12	0.14	0.06	−0.05	−0.14	−0.15	−0.24	−0.29
		perceived	0.39	−0.28	−0.28	−0.03	−0.17	−0.17	−0.20	**−0.41**	**−0.48**	**−0.47**	**−0.48**
serenity	150	expressed		−0.22	**0.60**	−0.29	**−0.31**	0.11	0.13	−0.02	−0.28	**−0.31**	0.17
		perceived	0.51	−0.20	**0.64**	0.02	**−0.34**	0.17	**0.35**	−0.10	−0.27	**−0.64**	−0.11
	Full	expressed		−0.14	**0.31**	**0.35**	**−0.44**	−0.05	−0.27	−0.29	**−0.34**	**−0.43**	−0.16
		perceived	0.25	−0.17	0.19	−0.07	**−0.30**	**−0.34**	−0.18	**−0.31**	**−0.52**	**−0.44**	0.00
tenderness	250	expressed		−0.20	0.06	−0.08	−0.10	−0.11	−0.08	−0.05	−0.10	−0.23	−0.14
		perceived	0.12	−0.20	**0.36**	−0.15	−0.29	−0.12	−0.05	−0.17	−0.28	**−0.44**	−0.26
	2500	expressed		0.14	−0.15	0.12	0.03	**−0.43**	0.02	−0.18	−0.18	−0.09	0.21
		perceived	0.22	**0.32**	0.17	**−0.30**	0.24	−0.08	**−0.32**	**−0.46**	−0.23	−0.15	0.22

100 ms or less of the recordings were presented for both vocal and musical expressions. One difference between music and speech concerned interest, which was recognized with above chance accuracy from 50 ms gates from vocal expressions, but was much slower at 500 ms for musical expressions. Neutral expressions were overall recognized with higher accuracy from speech compared to music, but did reach better than chance accuracy at 50 ms for both modalities.

Comparisons of the EIPs revealed that their duration (in ms) was overall longer for musical ($M = 1973$, SD = 960) compared to vocal expressions ($M = 1424$, SD = 979). However, anger, happiness, fear, and sadness did receive the fastest EIPs for both modalities (although in slightly different order), see Fig. 1 (for speech) and Fig. 2 (for music). Likewise, the EIPs for serenity were slow for both modalities. One difference concerned the EIPs for neutral expressions which were faster for speech than for music.

The comparison of confusion patterns showed that confusions were more prevalent for musical expressions, which is to be expected because recognition accuracy was lower, compared to vocal expressions. Nevertheless, the main confusions were similar for anger and happiness, and also relatively similar for neutral, sadness, serenity, and tenderness (see Tables II and IV). Differences were observed for fear, which was mainly confused with sadness for speech, but was instead frequently confused with anger for music. The confusion patterns also differed for interest, which was mainly confused with neutral for vocal expressions, but was instead confused with mainly anger and happiness for musical expressions. Response biases were observed mainly for judgments of sadness for music, whereas judgments of neutral were most biased for speech.

2. Acoustic cue patterns

To assess the similarity between the acoustic cue patterns for different emotions for speech and music, the correlations (Pearson *r*) between cue patterns for speech and music were calculated. For expressed emotions, the

correlation between cue patterns for music and speech was $r = 0.40$ for cues measured at the shortest gate with above-chance accuracy and $r = 0.31$ for cues measured at the EIP. This indicates a substantial association between how acoustic cues were used by speakers and musicians to express different emotions. For perceived emotions, the correlation between speech and music cue patterns was $r = 0.56$ at the shortest above-chance accuracy gate, and $r = 0.27$ at the EIP. This in turn indicates a substantial association between how listeners use acoustic cues to infer emotions from both speech and music. All correlations were calculated across all 10 selected cues for each of the 8 emotions that were common to both music and speech, $N = 80$, using the values shown in Tables III and V.

Although cue patterns were positively correlated between speech and music, it should be noted that associations were stronger at the shortest gate with above-chance accuracy than at the EIP, for both emotion expression and perception. As previously noted, cue patterns were also less consistent across the two gate lengths ("shortest gate" and EIP) for music compared to speech. Musical expression generally took longer to reach the EIP than did vocal expression, and this may have contributed to this observation. Taken together, these findings suggest that similarities between speech and music may be more pronounced for low-level acoustic information that is present in the first few hundred milliseconds of a sound, whereas there may be more variability in acoustic cues for longer durations.

It should also be noted that some aspects of the sound may be more difficult to control in music performance compared to the voice, and vice versa. For example, whereas speakers have a high degree of control over the shape of the vocal tract and thus can manipulate formant frequencies, violinists cannot directly change the shape of the sound box but instead have to rely on indirect means such as different bowing techniques (e.g., Harajda *et al.*, 1993). Future comparisons of acoustic cue patterns in speech and music could directly compare different types of cues (i.e., frequency, intensity, spectral balance, and durations) to achieve a fuller understanding of where similarities and differences occur. Such studies could also extend the number and type of musical instruments to investigate if lack of control over some emotion-relevant acoustic cues can be compensated for by the use of other available cues instead.

IV. CONCLUDING REMARKS

To summarize, the auditory gating paradigm was employed to study the time course of emotion recognition from speech prosody (Study 1) and music performances (Study 2), using a wide variety of both positive and negative emotions. The time course of emotion recognition was investigated using two main measures: (a) the shortest gate at which recognition accuracy was above the level of chance responding, and (b) EIPs which represent the point in time at which recognition rates stabilize. Anger, happiness, neutral, and sadness were recognized with above-chance accuracy already from the shortest or second shortest gates ($\leq$100 ms), and most other emotions were recognized from gate durations of 250 ms or less, from both speech and music. Some exceptions included vocally expressed relief which was recognized much later (1250 ms). EIPs were shortest for anger, happiness, and sadness for both speech and music, while recognition of low arousal positive expressions such as tenderness and serenity took longer to stabilize. EIPs were also generally longer for music than for speech. Finally, the patterns of acoustic cues were positively correlated between speech and music. The implications and limitations of these findings are discussed below.

We speculate that the time-course of emotion recognition from speech and music may proceed in (at least) three steps. First, initial discrimination between several emotions was possible from very brief stimuli ($\leq$250 ms), presumably based on low-level acoustic cues that can be transmitted and detected very quickly. Second, recognition accuracy continued to increase with increasing gate duration, reflecting an increase in the amount of relevant acoustic information that became available (e.g., cues that develop over time or additional low-level cues). Finally, recognition rates reached a plateau and stabilized after a certain gate duration was reached, which implies that adding more acoustic information (longer gates) after this point did not produce an increase in recognition accuracy. These results are consistent with Juslin and Laukka's (2003) proposition that nonverbal communication of emotion in speech and music is based on a number of probabilistic (i.e., uncertain) and partly redundant acoustic cues (see also Juslin, 1997).

That cues are probabilistic means that each cue in isolation is not a perfect indicator of the expressed emotion because it can, for example, be used in the same way to also express a different emotion (e.g., high mean $F0$ can be associated with both happiness and fear). Therefore, the more acoustic information that becomes available, e.g., through longer gate durations, the more reliable communication becomes. That cues are partly redundant means that several cues can be associated with the same emotion, and thus provide the same information (e.g., anger can be associated with high levels of both voice intensity and high-frequency energy). This means that there may be several cue combinations that result in similar emotion inferences, leading to a robust and flexible system of communication. However, it also means that adding more acoustic cues does not lead to further increases in accuracy if the cues do not provide any new emotion-relevant information. This explains how relatively brief acoustic stimuli can be as effective in conveying emotions as longer stimuli, at least once a certain threshold of acoustic information (approximately corresponding to the EIP) has been reached.

A main finding was that some emotions were recognized faster than others, indicating that the amount of acoustic information needed for accurate recognition may vary between emotions. However, it is hard to draw any firm conclusions about which specific emotions are recognized the fastest. One possible interpretation could be that negative emotions often are recognized faster than positive emotions. This is in line with previous gating studies (e.g., Grichkovtsova *et al.*, 2007; Jiang *et al.*, 2015) and chimes in with biological and evolutionary arguments according to

which it may be desirable to react more quickly to potentially harmful, compared to potentially beneficial, stimuli (e.g., Baumeister *et al.*, 2001). However, the observation that happiness was also among the most quickly recognized emotions in the present study suggests an alternative interpretation—not in terms of the valence (positive or negative) of the stimuli, but rather in terms of basic emotions. Basic emotions have been proposed as largely automated and biologically determined reactions to situations critical to survival, and lists of basic emotions usually contain anger, fear, happiness, and sadness among select other states (e.g., Tracy and Randles, 2011). From this perspective it could be expected that basic emotions are preferentially signaled by acoustic cues that can be processed especially quickly. It should be noted that theories of basic emotions have been criticized (e.g., Barrett, 2017; Moors, 2017), and that vocal expressions are not limited to basic emotions (e.g., Laukka *et al.*, 2016; Simon-Thomas *et al.*, 2009). Nevertheless, the present results tentatively support the interpretation that basic emotions often are recognized faster than other affective states (e.g., low arousal positive states) which are not as closely related to short term survival values. It should, however, be noted that vocally expressed fear presents a possible exception with only intermediate EIPs in the present study.

Many similarities were observed between trajectories of emotion recognition for speech and music. Acoustic cue patterns also showed great similarities between the two modes of expression. Although differences were also observed, results overall support the notion that speech and music share a common acoustic code for emotional communication (e.g., Juslin and Laukka, 2003; Scherer, 1995). Such observations provide empirical support for theories about a common origin for speech and music based on an affective proto-language, for which similarities in emotional communication is a central prediction (e.g., Fitch, 2006; Thompson *et al.*, 2012). The finding that several emotions could be recognized from very brief musical stimuli was novel and demonstrates the importance of low-level acoustic cues for musical expression. Such low-level cues are not unique to music, but instead shared with speech and possibly also with other emotion-related sounds present in our environment (e.g., Weninger *et al.*, 2013). However, recognition rates were overall lower, and EIPs were longer, for music compared to speech. This suggests that music may not express emotions as directly as speech, but instead (also) operates on a more symbolic level which may require more processing effort and acoustic cues that develop over time. For example, perception of melody and tonality requires processing of the pitch-relations between subsequent notes in a musical phrase and might convey additional cues that are especially relevant for musical expression (e.g., Parncutt, 2014).

Several limitations need to be acknowledged. The present study used a fine-grained temporal resolution of gate durations which allowed a high degree of control of the amount of acoustic information that was provided to participants in each trial, and this method also counteracted any potentially confounding influences of a speed-accuracy trade-off (see Cornew *et al.*, 2009). However, one limitation with the gating paradigm is that it does not allow strong conclusions about which specific cue combinations are diagnostic for conveying which emotions, because acoustic cues are not systematically manipulated. Future research could therefore combine the gating paradigm with systematic manipulations of acoustic cues in both speech (e.g., Birkholz *et al.*, 2015; Ladd *et al.*, 1985; Yanushevskaya *et al.*, 2018) and music (e.g., Eerola *et al.*, 2013) to determine which cues contribute most to fast recognition of various emotions. Systematic cue manipulation could also be used to test hypotheses inspired by the present study, for example, if acoustic cues that develop over time are more important for certain classes of stimuli (e.g., low arousal positive states, musical expressions) than for others (e.g., basic emotions, speech prosody).

The present study included a wider than usual selection of both positive and negative emotion categories, and the creation of fine-grained gates for each emotion led to a large number of stimuli for the perceptual experiments. However, despite the large number of gated stimuli, another limitation of the present study concerns the small number of actors/musicians and the limited speech/musical material present in the full-length recordings—which may limit the generalizability of the findings. In addition, recent studies have suggested small but systematic differences between posed and spontaneous vocal expressions (Juslin *et al.*, 2018). Therefore, future gating studies could include different types of stimuli (e.g., affect bursts, single words, and phrases of varying length), collected from different real-life situations involving demographically diverse speakers. A wide and representative selection of stimuli would be required to reach a consensus about the typical trajectories of emotion recognition for various expressions.

A final limitation concerns the forced-choice methodology to assess emotion recognition, which was used for consistency with previous gating studies (e.g., Pell and Kotz, 2011). This method has been criticized for creating inflated recognition rates due to artefactual agreement (e.g., Frank and Stennet, 2001). The current study included a wide selection of emotions and response options, which makes guessing the correct answer more difficult. Analysis of response biases—by means of Wagner's (1993) unbiased hit rate and inspection of confusion patterns—also suggested that systematic misclassifications that may have inflated recognition rates occurred rarely (bias was mainly observed for judgments of neutral in speech and sadness in music). Nevertheless, research has shown that people can infer not only emotion categories, but also emotion dimensions related to valence and arousal from both speech and music (e.g., Goudbeek and Scherer, 2010; Ilie and Thompson, 2006; Laukka *et al.*, 2005). Recent research has also shown that people can consistently judge aspects of the emotion eliciting situation—described in terms of emotion appraisal dimensions such as novelty, urgency, goal conduciveness, and norm compatibility—from vocal expressions (Laukka and Elfenbein, 2012; Nordström *et al.*, 2017). Combining a gating design with such alternative ways to assess emotion recognition could yield useful data that would have bearing on the debate about whether emotion categories or emotion dimensions are primary for emotion perception (e.g., Barrett,

2017). The present study could only ask about perceivers' conscious judgments, but future studies could extend the gating paradigm to also assess implicit emotion perception; for example, using ERPs and mismatch negativity (e.g., Paulmann and Pell, 2010; Sauter and Eimer, 2010). Comparisons between explicit and implicit processing of emotional information could provide additional clues about the very early stages of emotion recognition from speech and music.

To conclude, the present results attest to the importance of the acoustic channel for emotional communication in showing that several emotions could be recognized from very brief excerpts of speech prosody and music performance. Further investigations of similarities and differences between these two modes of expression have the potential to further our understanding of the mechanisms and processes underlying human vocalizations and musical behaviors.

ACKNOWLEDGMENTS

The research was funded by the Swedish Research Council (Grant No. 2012-801) and the Marianne and Marcus Wallenberg Foundation (Grant No. 2018-0059). Data and analysis code are available at the Open Science Framework (http://osf.io/a8de7).

[1]See supplementary material at https://doi.org/10.1121/1.5108601 for recognition rates presented separately for each speech stimulus in Study 1 (see supplementary Fig. 1), correlations between perceived emotions and acoustic cues presented separately for each gate length for both speech and music (see supplementary Fig. 2), the notation of the standard melody used in Study 2 (see Fig. 3 in the (see supplementary Fig. 3) material), and recognition rates presented separately for each music stimulus in Study 2 (see Fig. 4 in the (see supplementary Fig. 4) material).

Askenfelt, A. (**1991**). "Voices and strings: Close cousins or not?," in *Music, Language, Speech and Brain*, edited by J. Sundberg, L. Nord, and R. Carlson (Palgrave, London), pp. 243–256.

Audibert, N., Aubergé, V., and Rilliard, A. (**2007**). "When is the emotional information? A gating experiment for gradient and contours cues," in *Proceedings of the 16th International Congress of Phonetic Sciences*, edited by J. Trouvain and W. J. Barry (Saarland University, Saarbrücken, Germany), pp. 2137–2140.

Balkwill, L.-L., and Thompson, W. F. (**1999**). "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues," Music Percept. **17**, 43–64.

Banse, R., and Scherer, K. R. (**1996**). "Acoustic profiles in vocal emotion expression," J. Pers. Soc. Psychol. **70**, 614–636.

Bänziger, T., Mortillaro, M., and Scherer, K. R. (**2012**). "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception," Emotion **12**, 1161–1179.

Barrett, L. F. (**2017**). "The theory of constructed emotion: An active inference account of interoception and categorization," Soc. Cogn. Affect. Neurosci. **12**, 1–23.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (**2001**). "Bad is stronger than good," Rev. Gen. Psychol. **5**, 323–370323.

Bigand, E., Filipic, S., and Lalitte, P. (**2005**). "The time course of emotional responses to music," Ann. N.Y. Acad. Sci. **1060**, 429–437.

Birkholz, P., Martin, L., Willmes, K., Kröger, B. J., and Neuschaefer-Rube, C. (**2015**). "The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study," J. Acoust. Soc. Am. **137**, 1503–1512.

Bowling, D. L., Sundararajan, J., Han, S., and Purves, D. (**2012**). "Expression of emotion in Eastern and Western music mirrors vocalization," PLoS One **7**, e31942.

Chen, X., Zhao, L., Jiang, A., and Yang, Y. (**2011**). "Event-related potential correlates of the expectancy violation effect during emotional prosody processing," Biol. Psychol. **86**, 158–167.

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., and Flynn, L. M. (**2016**). "The voice conveys emotion in ten globalized cultures and one remote village in Bhutan," Emotion **16**, 117–128.

Cornew, L., Carver, L., and Love, T. (**2010**). "There's more to emotion than meets the eye: A processing bias for neutral content in the domain of emotional prosody," Cogn. Emot. **24**, 1133–1152.

Coutinho, E., and Schuller, B. (**2017**). "Shared acoustic codes underlie emotional communication in music and speech—Evidence from deep transfer learning," PLoS One **12**, e0179289.

Cowen, A. S., Elfenbein, H. A., Laukka, P., and Keltner, D. (**2018**). "Mapping 24 emotions conveyed by brief human vocalization," Am. Psychol. (published online).

Dalla Bella, S., Peretz, I., and Aronoff, N. (**2003**). "Time course of melody recognition: A gating paradigm study," Percept. Psychophys. **65**, 1019–1028.

Dinno, A. (**2009**). "Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data," Multivar. Behav. Res. **44**, 362–388.

Eerola, T., Friberg, A., and Bresin, R. (**2013**). "Emotional expression in music: Contribution, linearity, and additivity of primary musical cues," Front. Psychol. **4**, 487.

Ekman, P. (**1992**). "An argument for basic emotions," Cogn. Emot. **6**, 169–200.

Elfenbein, H. A., and Ambady, N. (**2002**). "On the universality and cultural specificity of emotion recognition: A meta-analysis," Psychol. Bull. **128**, 203–235.

Ellsworth, P. C., and Scherer, K. R. (**2003**). "Appraisal processes in emotion," in *Handbook of Affective Sciences*, edited by R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Oxford University Press, New York), pp. 572–595.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (**2016**). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," IEEE Trans. Affect. Comp. **7**, 190–202.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (**2013**). "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, edited by A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, and R. Zimmermann (ACM Press, New York), pp. 835–838.

Filipic, S., Tillmann, B., and Bigand, E. (**2010**). "Judging familiarity and emotion from very brief musical excerpts," Psychonom. Bull. Rev. **17**, 335–341335.

Fitch, W. T. (**2006**). "The biology and evolution of music: A comparative perspective," Cognition **100**, 173–215.

Frank, M. G., and Stennett, J. (**2001**). "The forced-choice paradigm and the perception of facial expressions of emotion," J. Pers. Soc. Psychol. **80**, 75–85.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A. D., and Koelsch, S. (**2009**). "Universal recognition of three basic emotions in music," Curr. Biol. **19**, 573–576.

Ghosh, J., Li, Y., and Mitra, R. (**2015**). "On the use of Cauchy prior distributions for Bayesian logistic regression," arXiv:1507.07170 [Stat].

Gobl, C., and Ní Chasaide, A. (**2010**). "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, 2nd ed., edited by W. J. Hardcastle, J. Laver, and F. E. Gibbon (Wiley-Blackwell, Chichester, England), pp. 387–423.

Goudbeek, M., and Scherer, K. (**2010**). "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," J. Acoust. Soc. Am. **128**, 1322–1336.

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

Grosjean, F. (**1980**). "Spoken word recognition processes and the gating paradigm," Percept. Psychophys. **28**, 267–283.

Hammerschmidt, K., and Jürgens, U. (**2007**). "Acoustical correlates of affective prosody," J. Voice **21**, 531–540.

Harajda, H., Mikiel, W., Gabryelczyk, P., and Fedyniuk, P. (**1993**). "Microstructure of sound: Formants in the dynamical spectra of violin sounds," Arch. Acoust. **18**, 17–32, available at http://acoustics.ippt.gov.pl/index.php/aa/article/view/1128/963.

Ilie, G., and Thompson, W. F. (**2006**). "A comparison of acoustic cues in music and speech for three dimensions of affect," Music Percept. **23**, 319–330.

Jiang, X., Paulmann, S., Robin, J., and Pell, M. D. (**2015**). "More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures," J. Exp. Psychol. Hum. Percept. Perform. **41**, 597–612.
Juslin, P. N. (**1997**). "Emotional communication in music performance: A functionalist perspective and some data," Music Percept. **14**, 383–418.
Juslin, P. N., and Laukka, P. (**2001**). "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," Emotion **1**, 381–412.
Juslin, P. N., and Laukka, P. (**2003**). "Communication of emotions in vocal expression and music performance: Different channels, same code?," Psychol. Bull. **129**, 770–814.
Juslin, P. N., Laukka, P., and Bänziger, T. (**2018**). "The mirror to our soul? Comparisons of spontaneous and posed vocal expression of emotion," J. Nonverbal Behav. **42**, 1–40.
Keltner, D., and Haidt, J. (**1999**). "Social functions of emotions at four levels of analysis," Cogn. Emot. **13**, 505–521.
Krumhansl, C. L. (**2010**). "Plink: 'Thin slices' of music," Music Percept. **27**, 337–354.
Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (**1985**). "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," J. Acoust. Soc. Am. **78**, 435–444.
Laukka, P., Eerola, T., Thingujam, N. S., Yamasaki, T., and Beller, G. (**2013**). "Universal and culture-specific factors in the recognition and performance of musical affect expressions," Emotion **13**, 434–449.
Laukka, P., and Elfenbein, H. A. (**2012**). "Emotion appraisal dimensions can be inferred from vocal expressions," Soc. Psychol. Pers. Sci. **3**, 529–536.
Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Chui, W., Iraki, F. K., Rockstuhl, T., and Thingujam, N. S. (**2013**). "Cross-cultural decoding of positive and negative non-linguistic vocalizations," Front. Psychol. **4**, 353.
Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., and Althoff, J. (**2016**). "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," J. Pers. Soc. Psychol. **111**, 686–705.
Laukka, P., Juslin, P., and Bresin, R. (**2005**). "A dimensional approach to vocal expression of emotion," Cogn. Emot. **19**, 633–653.
Lazarus, R. S. (**1991**). *Emotion and Adaptation* (Oxford University Press, New York).
Lima, C. F., Anikin, A., Monteiro, A. C., Scott, S. K., and Castro, S. (**2018**). "Automaticity in the recognition of nonverbal emotional vocalizations," Emotion **19**(2), 219–233.
Liu, T., Pinheiro, A. P., Deng, G., Nestor, P. G., McCarley, R. W., and Niznikiewicz, M. A. (**2012**). "Electrophysiological insights into processing nonverbal emotional vocalizations," Neuroreport **23**, 108–112.
Moors, A. (**2017**). "Integration of two skeptical emotion theories: Dimensional appraisal theory and Russell's psychological construction theory," Psychol. Inq. **28**, 1–19.
Mores, R. (**2017**). "Vowel quality in violin sounds—A timbre analysis of Italian masterpieces," in *Studies in Musical Acoustics and Psychoacoustics*, edited by A. Schneider (Springer, Cham, Switzerland), pp. 223–245.
Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (**2014**). "The musicality of non-musicians: An index for assessing musical sophistication in the general population," PLoS One **9**, e89642.
Nordström, H., Laukka, P., Thingujam, N. S., Schubert, E., and Elfenbein, H. A. (**2017**). "Emotion appraisal dimensions inferred from vocal expressions are consistent across cultures: A comparison between Australia and India," R. Soc. Open Sci. **4**, 170912.
Nummenmaa, L., and Calvo, M. G. (**2015**). "Dissociation between recognition and detection advantage for facial expressions: A meta-analysis," Emotion **15**, 243–256.
Ortony, A., Clore, G. L., and Collins, A. (**1988**). *The Cognitive Structure of Emotions* (Cambridge University Press, Cambridge, UK).
Parncutt, R. (**2014**). "The emotional connotations of major versus minor tonality: One or more origins?," Music. Sci. **18**, 324–353.
Paulmann, S., and Kotz, S. A. (**2008**). "Early emotional prosody perception based on different speaker voices," Neuroreport **19**, 209–213.
Paulmann, S., and Pell, M. D. (**2010**). "Contextual influences of emotional speech prosody on face processing: How much is enough?," Cogn. Affect. Behav. Neurosci. **10**, 230–242.
Peirce, J. W. (**2007**). "PsychoPy—Psychophysics software in Python," J. Neurosci. Methods **162**, 8–13.
Pell, M. D., and Kotz, S. A. (**2011**). "On the time course of vocal emotion recognition," PLoS One **6**, e27256.
Peretz, I., Gagnon, L., and Bouchard, B. (**1998**). "Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage," Cognition **68**, 111–141.
Piironen, J., Paasiniemi, M., and Vehtari, A. (**2018**). "Projective inference in high-dimensional problems: Prediction and feature selection," arXiv:1810.02406.
Pollack, I., Rubenstein, H., and Horowitz, A. (**1960**). "Communication of verbal modes of expression," Lang. Speech **3**, 121–130.
Rigoulot, S., Wassiliwizky, E., and Pell, M. D. (**2013**). "Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition," Front. Psychol. **4**, 367.
Sauter, D. A. (**2017**). "The nonverbal communication of positive emotions: An emotion family approach," Emot. Rev. **9**, 222–234.
Sauter, D. A., and Eimer, M. (**2010**). "Rapid detection of emotion from human vocalizations," J. Cogn. Neurosci. **22**, 474–481.
Sauter, D. A., Eisner, F., Calder, A. J., and Scott, S. K. (**2010**). "Perceptual cues in non-verbal vocal expressions of emotion," Q. J. Exp. Psychol. **63**, 2251–2272.
Scherer, K. R. (**1986**). "Vocal affect expression: A review and a model for future research," Psychol. Bull. **99**, 143–165.
Scherer, K. R. (**1995**). "Expression of emotion in voice and music," J. Voice **9**, 235–248.
Scherer, K. R. (**2003**). "Vocal communication of emotion: A review of research paradigms," Speech Commun. **40**, 227–256.
Scherer, K. R. (**2019**). "Acoustic patterning of emotion vocalizations," in *The Oxford Handbook of Voice Perception*, edited by S. Frühholz and P. Belin (Oxford University Press, Oxford, England), pp. 61–91.
Scherer, K. R., Clark-Polner, E., and Mortillaro, M. (**2011**). "In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion," Int. J. Psychol. **46**, 401–435.
Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., and Eyben, F. (**2017**). "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," J. Acoust. Soc. Am. **142**, 1805–1815.
Scherer, K. R., Sundberg, J., Tamarit, L., and Salomão, G. L. (**2015**). "Comparing the acoustic expression of emotion in the speaking and the singing voice," Comput. Speech. Lang. **29**, 218–235.
Schoonderwaldt, E. (**2009**). "The violinist's sound palette: Spectral centroid, pitch flattening and anomalous low frequencies," Acta Acust. Acust. **95**, 901–914.
Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (**2011**). "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Commun. **53**, 1062–1087.
Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., and Abramson, A. (**2009**). "The voice conveys specific emotions: Evidence from vocal burst displays," Emotion **9**, 838–846.
Spreckelmeyer, K. N., Kutas, M., Urbach, T., Altenmüller, E., and Münte, T. F. (**2009**). "Neural processing of vocal emotion and identity," Brain Cogn. **69**, 121–126.
Stan Development Team. (**2018a**). RStan: The R interface to Stan. R package version 2.17.3. http://mc-stan.org (Last viewed January 25, 2019).
Stan Development Team. (**2018b**). RStanArm: Bayesian applied regression modeling via Stan. R package version 2.17.4. http://mc-stan.org (Last viewed January 25, 2019).
Tangney, J. P., and Tracy, J. L. (**2012**). "Self-conscious emotions," in *Handbook of Self and Identity*, 2nd ed., edited by M. Leary and J. P. Tangney (Guilford Press, New York), pp. 446–478.
Thompson, W. F., Marin, M. M., and Stewart, L. (**2012**). "Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis," Proc. Natl. Acad. Sci. U.S.A. **109**, 19027–19032.
Tracy, J. L., and Randles, D. (**2011**). "Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt," Emot. Rev. **3**, 397–405.
Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., and Bouchard, B. (**2008**). "Happy, sad, scary and peaceful musical excerpts for research on emotions," Cogn. Emot. **22**, 720–752.
Wagner, H. L. (**1993**). "On measuring performance in category judgment studies of nonverbal behavior," J. Nonverbal Behav. **17**, 3–38.
Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., and Scherer, K. R. (**2013**). "On the acoustics of emotion in audio: What speech, music, and sound have in common," Front. Psychol. **4**, 292.
Yanushevskaya, I., Gobl, C., and Ní Chasaide, A. (**2018**). "Cross-language differences in how voice quality and f_0 contours map to affect," J. Acoust. Soc. Am. **144**, 2730–2750.