Humanities & Social Sciences Communications



ARTICLE

https://doi.org/10.1057/s41599-020-0499-z

OPEN



1

Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters

Adi Lausen

1,2

& Kurt Hammerschmidt

3,4

Human speech expresses emotional meaning not only through semantics, but also through certain attributes of the voice, such as pitch or loudness. In investigations of vocal emotion recognition, there is considerable variability in the types of stimuli and procedures used to examine their influence on emotion recognition. In addition, accurate metacognition was argued to promote correct and confident interpretations in emotion recognition tasks. Nevertheless, such associations have rarely been studied previously. We addressed this gap by examining the impact of vocal stimulus type and prosodic speech attributes on emotion recognition and a person's confidence in a given response. We analysed a total of 1038 emotional expressions according to a baseline set of 13 prosodic acoustic parameters. Results showed that these parameters provided sufficient discrimination between expressions of emotional categories to permit accurate statistical classification. Emotion recognition and confidence judgments were found to depend on stimulus material as they could be reliably predicted by different constellations of acoustic features. Finally, results indicated that listeners' accuracy and confidence judgements were significantly higher for affect bursts than speech-embedded stimuli and that the correct classification of emotional expressions elicited increased confidence judgements. Together, these findings show that vocal stimulus type and prosodic attributes of speech strongly influence emotion recognition and listeners' confidence in these given responses.

¹ Department of Affective Neuroscience and Psychophysiology, Institute for Psychology, University of Goettingen, 37073 Goettingen, Germany. ² Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK. ³ Cognitive Ethology Laboratory, German Primate Center, University of Goettingen, Kellnerweg 4, 37077 Goettingen, Germany. ⁴ Leibniz ScienceCampus "Primate Cognition", 37077 Goettingen, Germany. ⁸email: adi.lausen@psych.unigoettingen.de

Introduction

he ability to correctly understand and appropriately respond to other persons' emotions plays an important role in everyday social interactions (Chronaki et al., 2018; Juslin and Scherer, 2005). In verbal communication, for instance, humans do not merely consider what their interlocutors are saying (i.e., semantic meaning), but also how they are conveying the spoken information (e.g., high/low pitch of their voice). An all-encompassing term for such vocal qualities of speech is prosody (i.e., tone of voice). Research has shown that prosody may support the correct interpretations of utterances independently of linguistic comprehension (Paulmann, 2016; Thompson and Balkwill, 2009; Kitayama and Ishii, 2002), with studies reporting recognition rates for emotions to be significantly higher than chance (Cowen et al., 2019a, 2019b; Lausen and Schacht, 2018; Cordaro et al., 2016; Paulmann and Uskul, 2014; Juergens et al., 2013; Pell et al., 2009; Scherer et al., 2001). In addition, metacognition, the ability to actively monitor and reflect upon one's own performance, has been argued to impact judgements of accuracy in emotion recognition tasks (Bègue et al., 2019; Kelly and Metcalfe, 2011; Dunlosky and Metcalfe, 2009). This ability reflects how faithfully our self-judgement relates to actual performance in a task and can be measured by retrospective confidence ratings (Bègue et al., 2019). To better understand the mechanisms underlying the recognition of emotions from the voice, the present study examined how different types of vocal stimuli and their acoustic attributes influence listeners' recognition of emotions and confidence ratings.

In their endeavour to assess the recognition of emotions from prosody, researchers created a wide variety of stimulus materials. Some decided to use sentences as stimuli because they have been argued to have higher ecological validity (Sauter, 2006). However, as emotions are not expressed to the same degree in each word of a sentence it has been suggested that such long-lasting stimuli might contain increased variation and noise in the signal (Sauter, 2006). Thus, other investigators choose single words as stimulus material as they do not "dilute" the characteristics of a specific emotion (Hammerschmidt and Juergens, 2007). To examine the role of semantics in prosody processing, previous research manipulated both these aspects orthogonally and then compared the conditions in which semantics and prosody are emotionally congruent or incongruent (Kotz and Paulmann, 2007). The reported results across studies indicated that specific emotions are more easily recognized if semantic information is available even if semantics are irrelevant to the given task (e.g., Ben-David et al., 2016; Paulmann and Kotz, 2008), while others reported that semantic information might facilitate or interfere with a listener's judgment about the emotional content of the stimulus when spoken in a congruent or incongruent prosody (e.g., Kotz and Paulmann, 2007; Mitchell et al., 2003; Nygaard and Lunders, 2002). For instance, Schirmer (2010) found that recognition performance is similar for words studied with emotional and neutral prosody. However, compared to words with neutral prosody, words with sad prosody were rated as more negative, while words with happy prosody were rated as more positive. Given that trade-off effects are bound to occur when semantics are congruent or incongruent with prosody (i.e., semantic information might confound prosody of the spoken stimuli), previous research used either neutral sentences/words intoned to convey different emotions (e.g., Hammerschmidt and Juergens, 2007) or presented the speech-embedded material in a pseudo-language (i.e., an artificially created language devoid of meaning). This last procedure represents a useful way to neutralize or mask the semantic content while retaining the prosodic information (Rigoulot et al., 2013; Banse and Scherer, 1996). Studies on the identification of vocal emotions from pseudo-utterances found

overall recognition rates for discrete emotions to be significantly higher than chance (e.g., Pell et al., 2009; Scherer et al., 2001) and that semantic neutral sentences and pseudo-sentences elicit similar recognition accuracy rates (e.g., Castro and Lima, 2010). However, the analysis of emotional prosody in isolation (i.e., without lexico-semantic content) might not only increase the artifice of the acted emotions but could also lead to poorer decoding accuracy (Parsons et al., 2014). Thus, it has been suggested that affective (non-speech) sounds or affect bursts (e.g., laughter, screams) are more reliable stimuli (Juergens et al., 2018) comprising the most 'natural and ancient language of emotion communication whose expressiveness no words can ever achieve' (Bostanov and Kotchoubey, 2004, p. 259).

The discussion whether the recognition of vocal emotions from speech-embedded materials has an advantage over non-speech embedded materials or vice-versa is far from settled. The extraction of acoustic cues from their created materials led. however, to an agreement among researchers that pitch, loudness, tempo and quality (stressed/breathy voice) are the most relevant paralinguistic features that speakers employ when expressing emotions (e.g., Goudbeek and Scherer, 2010). A series of statistics on these paralinguistic features has revealed that parameters related to pitch or fundamental frequency (F_0) (e.g., minimum, maximum, mean, jitter), energy/amplitude- (e.g., loudness, shimmer), temporal (e.g., duration) and quality parameters (e.g., harmonics-to-noise ratio [HNR]) are amongst the most important 'candidates' for prosodic correlates of emotion in speech (e.g., Juslin and Laukka, 2003; Johnstone and Scherer, 2000). In their seminal work on the acoustic profiles of vocal emotion expression, Banse and Scherer (1996) examined relations between 14 induced emotions and 29 acoustic parameters. Implementing linear discriminant analysis (LDA) with jackknife and crossvalidation procedures for the evaluation of classification errors the authors found that, on a 7% chance level assuming a uniform distribution, the overall patterns when categorizing emotions were similar to those of listeners' accuracy (LDA = 40% jackknife estimate of accuracy; LDA = 25% cross-validation estimate of accuracy; listeners = 48% accuracy). Subsequent studies conducted with different stimulus types (e.g., words, affect bursts, semantically emotional and neutral sentences), with less or larger acoustic parameter sets (from 3 to 40 prosodic features), with the same (i.e., LDA) or other classification methods (e.g., random forest (RF), k-nearest-neighbour classifier) obtained comparable results showing that both, classifiers and listeners perform similarly well when predicting emotion category membership based on the acoustic profiles of their utterances (e.g., Noroozi et al., 2017; Pichora-Fuller et al., 2016; Sbattella et al., 2014; Sauter et al., 2010; Castro and Lima, 2010; Paulmann et al., 2008; Hammerschmidt and Juergens, 2007; Toivanen et al., 2004; Juslin and Laukka, 2001).

Together these findings allow the conclusion that prosodic acoustic parameters (among other cues, e.g., semantics) provide listeners with appropriate cues to generally understand the intended emotion and, thus, contribute in a cumulative fashion to the communication and recognition of emotions (Thompson and Balkwill, 2009). Nevertheless, it has been argued that using different types of stimuli, different sets of acoustic parameters and implementing various classification methods causes serious difficulties when interpreting the results across studies, endangering the accumulation of empirical evidence (Eyben et al., 2016). Thus, adopting a baseline set of acoustic parameters and systematically analyzing their influence on emotion recognition accuracy across the various types of vocal stimulus material would improve methodological rigour and increase the reliability of findings (Eyben et al., 2016; Bak, 2016; Juslin and Laukka, 2003).

As emotion recognition is subject to environmental influences and dictates a great variability in the way individuals interpret emotional messages, it has been argued that metacognition (i.e., the awareness of one's own knowledge) might impact judgements of accuracy in emotion recognition tasks. Kelly and Metcalfe (2011), for instance, investigated whether individuals can accurately predict and assess their performance on two face emotion recognition tasks (i.e., Mind in the Eyes task and Ekman Emotional Expression Multimorph Task). For each emotional expression, participants were asked to predict (1) their future performance in correctly identifying the emotions, i.e., prospective judgements, and (2) the accuracy regarding their confidence in the given responses, i.e., retrospective judgements. Results from the Mind in the Eyes task showed significantly higher scores for retrospective than prospective confidence judgements, however, no significant relationship between these judgements and performance accuracy was found. Even though in the Emotional Expression Multimorph Task, the gamma correlations were slightly greater for retrospective (r = 0.43) than prospective judgements (r = 0.32), the authors found a significant relationship between both types of judgements and performance accuracy. Based on these findings, the authors concluded that individuals who perform better in emotion recognition tasks are also more accurate in their metacognitive assessments. While some studies on the perceptual-acoustic correlates of the confidence that is expressed in the voice of the speaker, indicated that for listeners both linguistic and acoustic-prosodic cues were fundamental when making retrospective judgements about speakers' mental states (e.g., Jiang and Pell, 2017, 2014; Kimble and Seidel, 1991; Scherer et al., 1973), other studies demonstrated that in tasks assessing vocal expressions of emotion, listeners' confidence increased with stimulus duration (Pell and Kotz, 2011). For instance, Rigoulot et al. (2013) investigated the time course of vocal emotion recognition employing a modified version of an auditory gating paradigm. Results showed that, independent of stimulus presentation (forward or backward), listeners' confidence in categorizing the emotions increased significantly with longer gate intervals (i.e., number of syllables). This pattern of results clearly indicates that when assessing the recognition of vocal emotions, duration, among other acoustic parameters (e.g., pitch, loudness), progressively activates emotion-specific knowledge leading to higher accuracy and confidence ratings (for an example on how the recognition of discrete vocal emotion in affect bursts evolves over time, see Schaerlaeken and Grandjean, 2018). While these findings revealed how much information might be needed for listeners to consciously reflect on and categorize vocally expressed emotions from paralinguistic attributes of speech, there is a lack of direct evidence examining the influence of vocal stimulus type and their related acoustic parameters on emotion recognition and confidence ratings. In the present investigation, we contribute to filling this gap by means of two studies.

Study 1: Performance accuracy by classification algorithms

Humans communicate emotions with the voice through prosody and vocal bursts (Cowen et al., 2019b). Research has long claimed that certain acoustic features, such as pitch, loudness, tempo or quality and their related parameters (e.g., fundamental frequency, jitter, shimmer, harmonics-to-noise ratio) drive the recognition of emotions from prosody and vocal bursts (e.g., Sauter et al., 2010; Scherer and Baenziger, 2004; Banse and Scherer, 1996). Nevertheless, a comparison of findings across different studies on emotion recognition from speech seems an invalid approach as performance accuracy has been argued to essentially depend on the stimulus material and the extracted set of acoustic parameters

(e.g., Bak, 2016; Sauter et al., 2010; Toivanen et al., 2004). Guided by these observations, we extracted a baseline set of acoustic parameters from our stimuli datasets and employed two procedures to capture the psychophysical properties of these measurements.

Method

Stimulus material and acoustic analyses. One thousand thirtyeight emotional expressions spoken in an *angry*, *disgusted*, *fearful*, *happy*, *neutral*, *sad* and *surprised* tone of voice were sampled from established speech corpora or from researchers that developed their own stimulus materials. Table 1 provides a brief summary of the datasets [for further details on stimuli selection (i.e., inclusion criteria and normalization procedures) see Lausen and Schacht, 2018].

The stimulus material was analysed for frequency-related parameters (mean fundamental frequency (F_0) , minimum F_0 , maximum F_0 , standard deviation F_0 , jitter), energy/amplitude-related parameters (shimmer, amplitude [dB], peak amplitude, mean HNR, maximum HNR, standard deviation HNR) and temporal features (duration, peak time) using GSU Praat Tools script packages developed by Owren (2008), which allows batch processing during measurement (for details on the processing of acoustic parameters see supplementary material). Following the procedures of Goudbeek and Scherer (2010), Sauter et al. (2010) and Juslin and Laukka (2001) the measurements were made over the entire utterances, across all speakers and all items of the same type of stimulus.

Statistical analysis. A LDA was then performed for each type of stimulus separately and across all stimuli in both groups (i.e., Group Words and Group Sentences; for details see "Methods" section Study 2) to determine the optimal combination of the 13 above-mentioned acoustic parameters for predicting emotion category membership. In the analysis, acoustic measurements served as independent variables while the dependent variable was the intended emotional category. As the set of acoustic parameters was not very large, no feature selection method (e.g., stepwise analysis) was used to reduce the number of parameters. LDA is optimal if the acoustic parameters have a multivariate normal distribution with different means for each emotion and identical variance matrices for all emotions. However, if the underlying multivariate structure is more complex, other classification algorithms have been suggested to yield better performance (James et al., 2013). To assess whether our LDA model shows better predictive performance than other classification techniques we implemented RF as an additional classification algorithm. This ensemble classification methodology, which combines a large number of decision trees using different sets of predictors at each node of the trees, was argued to be a more robust alternative to discriminant analysis or multinomial regression as it allows a selection of the important potential predictors among a large number of variables with complex interactions (Anikin and Lima, 2018; Breiman, 2001; see also supplementary material for more details on RF).

These two classification methods were chosen due to their easy implementation and parametrization. LDA is currently the most used method for emotion classification in vocal stimuli (e.g., Sauter et al., 2010; Hammerschmidt and Juergens, 2007; Banse and Scherer, 1996). Multivariate normality and homoscedasticity of the data are assumed for this method and, therefore, data transformations must often be carried out to reach or to come close to these requirements (Seber, 1984). A potential drawback of this method includes a tendency towards overfitting (Dixon and Brereton, 2009). In contrast, RF has previously been used for

Table 1 Summary of the spe	eech and non-speech datasets.		
Speech corpora	Description of content	Initial content	Number of selected files
Anna (Hammerschmidt and Juergens, 2007)	Name "Anna" uttered for 8 emotions [(anger, affection, contempt, despair, fear, happiness, sensual satisfaction, triumph) + neutral (baseline expression)] by 22 German drama students [10 males (M); 12 females (F)] (same for all emotions).	198 audio files	88 [(emotion category of interest + baseline expression) ×22 speakers]
Montreal Affective Voices (Belin et al., 2008)	Portrayals of non-verbal emotional sounds/affect bursts (e.g., laughing, crying) for 8 emotions [(anger, disgust, fear, happiness, pain, pleasure, sadness, surprise) + baseline expression (neutral)] by 10 francophone actors (5 M; 5 F) (same for all emotions).	90 audio files	70 [(emotion category of interest + baseline expression) ×10 Speakers]
Berlin Database of Emotional Speech (Burkhardt et al., 2005)	Portrayals of 6 emotions [(anger, boredom, disgust, fear, happiness, sadness) + baseline expression (neutral)] by 10 German untrained actors (5 M; 5 F). The database consists of 10 semantic neutral sentences (same for all emotions).	816 audio files	120 [(emotion category of interest + baseline expression) ×2 Speakers × 10 sentences]
Magdeburg Prosody Corpus (Wendt and Scheich, 2002)	Portrayals of 5 emotions [(anger, disgust, fear, happiness, sadness) + baseline expression (neutral)] by 2 German actors (1 M; 1 F). The corpus consists of 3318 nouns classified according to their positive-, negative-and neutral semantic content and of 222 pseudo-words (same for all emotions).	3318 audio files (nouns)+222 audio files (pseudo- words)	480 [(all emotions + baseline expression) ×2 Speakers × 10 nouns per semantic category (i.e., positive, negative, neutral)/10 Pseudo-words]
Paulmann Prosodic Stimuli (Paulmann and Kotz, 2008; Paulmann et al., 2008)	Portrayals of 6 emotions (anger, disgust, fear, happiness, sadness, surprise) + baseline expression (neutral) by 2 German actors (1 M; 1 F). The stimulus set consists of 210 lexical sentences and 210 pseudo-sentences (different for each emotion).	420 audio files	280 [(10 lexical sentences & 10 pseudo-sentences for each emotion + baseline expression) ×2 speakers]

discrimination among vocal emotions by only few studies (e.g., Anikin and Lima, 2018; Noroozi et al., 2017). While other machine-learning methods are available, we chose RF because it represents an adequate analysis to LDA, requires no distributional assumption of the dataset and has no possibility of overfitting (Breiman, 2001). The two classification methods were compared by the estimated classification errors using 10-fold crossvalidation.

The data was analysed using the R language and environment for statistical computing and graphics version 3.4.3 (R Core Team, 2017) and the integrated environment *R-Studio version* 1.0.153 (used packages: *lda*; *randomForest*; *ipred*).

Results

The results obtained from LDA showed that the vast majority of variance was accounted by the first two LD functions (accounted variance ranging from 37.80% to 73.96% for LD1 and from 14.86% to 32.50% for LD2). In supplementary material are presented the complete tables for the correlations between acoustic parameters and all LD functions as well as the accounted variance for each function [Table S1 ($A_{\rm I}$ to $A_{\rm XI}$)]. Figure 1 illustrates how the scores of the two LD functions separate the emotional categories for each- and across all stimuli.

Comparisons between RF and LDA revealed that the error rates were overall smaller by RF than LDA when predicting

emotion category membership across all 1038 stimuli. Specifically, error rates were reduced by 23.11% across all stimuli types, by 11.82% across all stimuli in *Group Words* and by 19.53% across all stimuli in *Group Sentences*. Table 2 displays the error rates for both classification methods and the differences between RF and LDA error rates relative to the error rates of LDA.

Discussion

Our results showed that LDA was able to correctly classify the indicated emotion category of 42% for all stimuli types unveiling specific constellations of predictors for each emotion and stimulus set with cross-validation estimates of accuracy, ranging from 41% for lexical sentences to 61% for Anna stimuli. The results of this analysis compare well with previous work reporting accuracy rates for their stimulus materials between 40% and 57% (e.g., Juergens et al., 2013; Sauter et al., 2010; Castro and Lima, 2010; Hammerschmidt and Jürgens, 2007; Banse and Scherer, 1996). By implementing RF as an additional classification method and using the same acoustic predictors, we observed that RF achieved a cross-validation classification accuracy across all stimuli, which was 32% relative higher than the accuracy of the LDA model. This result is in line with the findings reported by Noroozi et al. (2017), who investigated whether RF shows better predictive performance than deep neural networks (DNN) or more established techniques such as LDA, based on a set of 14 acoustic

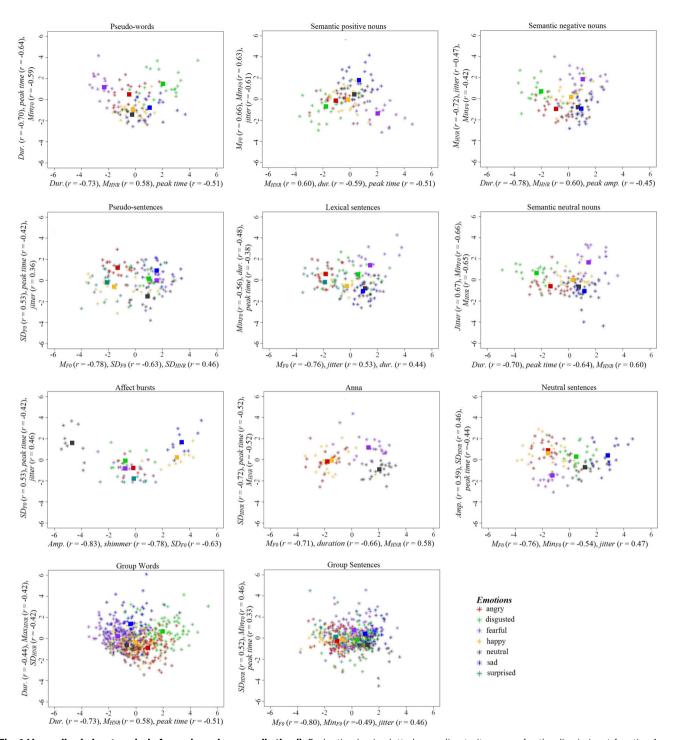


Fig. 1 Linear discriminant analysis for each- and across all stimuli. Each stimulus is plotted according to its scores for the discriminant function 1 and function 2. The squares represent the group means on each of the discriminant functions. On x- and y-axis are displayed the parameters that most strongly correlated with the first two linear discriminant functions (Min. = minimum; Max. = maximum; M = mean; SD = standard deviation; HNR = harmonics-to-noise ratio; F_0 = fundamental frequency; Dur. = duration; Amp. = amplitude).

parameters extracted from Surrey Audio-Visual Expressed Emotion database. Their results showed that on average RF recognition rate was 26% higher relative to LDA and 11% higher compared to DNN. Although the comparison between different classification methods highlights that RF outperforms techniques such as LDA (Noroozi et al., 2017), we would like to note that results from both classification analyses demonstrated that acoustic measurements alone provide sufficient information to

discriminate successfully between stimuli from different emotional categories.

Study 2: Emotion recognition and confidence ratings by stimulus type and emotion categories

Considering that many perceptual experiments (e.g., Noroozi et al., 2017; Pichora-Fuller et al., 2016; Sbattella et al., 2014; Sauter

Table 2 Linear discriminant analysis (LDA) and random forest (RF) 10-fold cross-validation classification error rates for predicting vocal stimuli emotional category membership.

Stimulus types	Error rates	i	
	LDA	RF	Δ (%) ^a
Anna	0.3903	0.3778	3.20
Pseudo-words	0.4597	0.4225	8.09
Semantic positive nouns	0.4112	0.3428	16.63
Semantic negative nouns	0.4614	0.4019	12.90
Semantic neutral nouns	0.4402	0.4438	-0.82
Group Words [across all	0.4257	0.3754	11.82
stimuli (<i>N</i> = 568)]			
Affect bursts	0.4347	0.4701	-8.14
Pseudo-sentences	0.5211	0.4259	18.27
Lexical sentences	0.5866	0.4448	24.17
Neutral sentences	0.4777	0.4255	10.93
Group Sentences [across all	0.5981	0.4813	19.53
stimuli (<i>N</i> = 470)]			
Overall [across all stimuli	0.5802	0.4461	23.11
types ($N = 1038$)]			

^aThe relative difference between RF error rates and LDA error rates were calculated as follows: (1–(RF error rates/LDA error rates))*100. As it can be observed, the error rates were smaller by RF than LDA, except for semantic neutral nouns and affect bursts. The accuracy rates for both classification methods can be obtained as follows: (1–error rates)*100.

et al., 2010), including our first study, have demonstrated that automatic classification of simulated vocal emotions is possible with a reasonably high accuracy rate, the general finding across studies has so far been that the automatic classification of emotions is considerably less successful than human classification performance (for a review see Toivanen et al., 2004). As outlined in the introduction, research on metacognition has related this skill to vital aspects of socio-emotional processes and has reliably shown that confidence judgements are higher when given after rather than before a response to an emotion recognition task. In this emerging science, however, it is not well-understood how different types of vocal stimuli and their acoustic attributes influence listeners' recognition of emotions and confidence ratings. Such data would allow a more differentiated assessment of the factors assumed to impact both emotion recognition accuracy and a person's confidence in a given response. We therefore sought in our second study to analyse: (1) whether the automatic classification of vocal emotions is less successful than human classification performance (by taking as a benchmark the overall 'best' performing classification method from our first study, i.e., RF); (2) which vocal stimulus types are best predicted by the acoustic parameters and which acoustic parameters are best suited for which stimulus types; (3) whether listeners' performance accuracy and confidence ratings vary across certain types of stimuli and specific emotion categories; (4) if correct recognition of emotions is related to increased confidence judgements.

Method

Participants. Two-hundred ninety participants (143 females, 147 males; age range = 18–36 years) completed the study after responding to advertisements posted on social media (e.g., Facebook) or to flyers distributed across the university campus. Participants averaged 23.83 years in age (SD = 3.73) with 62% having completed a general qualification for university entrance, 25% a bachelor degree, 12% a master degree and 1% a general certificate of secondary education. To reduce the length of the experiment, participants were allocated to two groups of equal size. One group listened to words and pseudo-words (Group Words: n = 145, $M_{\rm age} = 24.00$, ${\rm SD}_{\rm age} = 3.67$), while the other group listened to

affect bursts, sentences and pseudo-sentences (Group Sentences: n=145, $M_{\rm age}=23.66$, ${\rm SD}_{\rm age}=3.80$). No significant age difference between the two groups was observed ($t_{(288)}=0.786$; p=0.432; ${\rm CI}_{95\%}=[-0.52;\ 1.21]$). All participants were native speakers of German and reported no hearing difficulties.

Procedure and experimental task. Up to four participants were invited to each experimental session, which lasted ~60 min. At arrival, the experimenter informed the participants about the aim of the study, i.e., to validate a set of auditory stimuli with emotional content. Prior to formal testing, participants signed a consent form and completed a short demographic questionnaire concerning age, gender and education level. Participants were informed that all stimuli would be presented only once, the number of presented emotions might vary from the number of categories given as possible choices, and some of the stimuli were not supposed to carry any semantic meaning and might sound 'foreign'. After these instructions and completion of ten practice trials, participants started the main experiment, presented via Presentation software (Version 14.1, Neurobehavioral Systems Inc., Albany, CA). Stimuli were presented to the participants binaurally with Bayerdynamic DT 770 PRO headphones pluggedin the tower box of a Dell OptiPlex 780 SFF Desktop PC Computer. To ensure equal physical volume of stimulus presentation across the four PCs, we measured the sound level meters of the ten practice stimuli with a professional sound level meter (Norsonic Nor140, 2017). No significant difference in volume intensity was observed ($F_{(3,27)} < 1$). Following each stimulus presentation listeners rendered two judgements. First, they classified which emotion was being expressed by the speaker from a list of seven categories presented on the computer screen. To assess metacognition, this rating was followed by a 7-point rating scale on the screen to estimate their confidence in the preceding response (1 = not at all confident; 7 = extremely confident). Each trial began with a white fixation-cross presented on a grey screen, which was shown until participants' response had been recorded. The presentation of the stimuli was initiated by pressing the Enter-key. The auditory stimulus was then presented alongside the fixation cross. The responses were made using the marked computer keyboard (Z to M for the emotion judgements, which were labelled corresponding to the emotion categories, and 1-7 for confidence). There was no time limit for emotion judgements or confidence ratings. At the end of each block a visual message in the centre of the screen instructed participants to take a break if they wished to or to press the Spacebar to proceed with the next block. The set of stimuli in Group Words was split into three blocks (i.e., Anna, Pseudo-words, and Nouns), while in Group Sentences the set of stimuli was divided into four blocks (i.e., Pseudo-sentences, Lexical Sentences, Neutral sentences, and Affect bursts). The order of blocks and of the stimuli within each block were randomized. Blocks were separated by a break of selfdetermined duration. The reimbursement of participants consisted of 8€ or course credit.

The study has been approved by the Ethical Committee of the Georg-Elias- Mueller-Institute of Psychology, University of Goettingen, Germany (*number 149*) and conducted in accordance with the ethical principles formulated in the *Declaration of Helsinki* (2013).

Study design and power analysis. To assess listeners' judgements of emotions and confidence in their judgements a *within-subjects design* was fitted for *Group Words* and *Group Sentences*. The design was balanced for emotion categories in each stimulus type. Independent within-subject factors were *stimuli types*, *acoustic parameters* and *emotion categories*. Dependent variables were

emotion recognition and confidence ratings. To assess whether we had enough power to answer our research questions, an approximate correlation power analysis was calculated and, Bonferroni corrected for the 13 acoustic parameters. A sample size of 145 participants per group with a minimum set of stimuli per participant (i.e., 70) allowed us to detect correlations of r = 0.037 with a type I error rate of 5% and power 80%. To describe the power to detect differences between emotion categories and stimulus types an approximate Tukey's multiple pairwise comparisons power analysis was computed. Assuming a minimum set of 10 stimuli for each emotion category and a sample size of 145 participants per group allowed us to detect a difference of 0.044 for recognition probability at 0.80 with a type I error rate of 5% and power 80% (for further details on statistical power calculations see supplementary material).

Statistical analysis. The data was analysed by *generalized linear models* (quasi-binomial logistic regression) for the binary response variable emotion recognition and by *linear models* for the response variable confidence ratings. To find a reduced model that best explains the data on the 13 acoustic parameters for the two dependent variables a *backward stepwise variable selection* (R function *step*) was conducted in a generalized linear model (binomial logistic regression) for the binary response variable emotion recognition. The dispersion parameter of the quasi-binomial model and the nominal variable participants accounted for dependencies caused by repeated measurements within the participants.

In the global models, stimulus types, acoustic parameters and emotions were included as predictor variables. Participants, emotions, and stimulus types were fitted as nominal variables and acoustic parameters as quantitative variables. The order of the acoustic parameters in the models was determined by importance in a backward stepwise variable selection, that is in descending order starting with the acoustic parameter that explained most of the deviance. Conditional models were fitted for each stimulus type to account for interactions between stimulus types, emotions and acoustic parameters, since differences between fitted parameters of the models can be interpreted in terms of interactions (Hothorn et al., 2008). The relation between confidence ratings and emotion recognition was analysed by a linear model with the response variable confidence ratings and the predictor variables stimulus types and emotion recognition. Chi-square tests of the deviance analysis were used to analyse effects of predictor variables. In the quasi-binomial logistic regression, odds ratios (ORs) were used to compare emotion categories as well as stimulus types. Confidence ratings of the linear model were compared by calculating the differences of the means. Tukey's method of multiple pairwise comparisons was used to compute simultaneous 95% confidence intervals for both, OR and mean differences.

For the descriptive analysis of the data the following calculations were carried out: relative frequencies, confusion matrices, classification errors by RF and listeners' judgements of emotion categories, confidence intervals by binomial test and Wagner's (1993) unbiased hit rate ($H_{\rm u}$), which is the rate of correctly identified stimuli multiplied by the rate of correct judgements of the stimuli. The data was analysed using the R language and environment for statistical computing and graphics version 3.4.3 (R Core Team, 2017) and the integrated environment R-Studio version 1.0.153 (used packages: pwr; MASS; multcomp; mvtnorm; ggplot2; stats {binom.test}).

Results

Human classification performance vs. RF. The accuracy percentage of correctly identified emotions, Wagner's unbiased hit

rate, the comparisons in classification errors between listeners' judgements of emotion categories and RF algorithm, as well as, the 95%CI of the exact binomial test are displayed in Table 3. Figure S1A, B in supplementary material displays listeners' recognition accuracy by stimulus type and emotions. As it can be observed from these analyses, listeners most frequently misclassified happy for surprise (Anna, nouns, affect bursts, neutral sentences, pseudo-sentences), and fear for sadness (Anna, pseudowords, semantic negative nouns, semantic positive nouns, lexical sentences, pseudo-sentences). In lexical- and pseudo-sentences, however, participants often mistook surprise for happy, whereas the sad tone of voice was frequently misclassified as neutral (pseudo-words, nouns, lexical sentences). Although generally wellrecognized, utterances spoken in an angry tone of voice were often mistaken for surprise (affect bursts, lexical- and pseudosentences, pseudo-words) and disgust (nouns), while for neutral and disgusted prosody no clear error pattern emerged [e.g., some utterances spoken in a neutral tone of voice were either mistaken for angry (Anna, nouns) or sad (lexical-, neutral sentences), while utterances spoken in a disgusted tone of voice were misclassified as angry (nouns, affect bursts) or neutral (lexical-, pseudosentences)]. Comparing the proportion of classification errors between listeners' judgements of emotions and RF, one could assume that globally, i.e., across all emotions, humans were significantly better at predicting emotion category membership relative to RF, except for Anna stimuli, where no significant difference was observed. Looking at specific emotion categories, results indicated that in some stimulus sets the RF algorithm significantly outperformed listeners when classifying disgust and sad (pseudo-words and nouns), fear (pseudo-words and neutral sentences), happy (Anna) and surprise (i.e., pseudo-, lexical sentences).

Vocal stimulus types predicted by acoustic parameters. The proportion of variance explained by the acoustic measures for emotion recognition and confidence ratings is displayed in Table 4.

An inspection of Table 4 indicates that the proportion of explained variance for emotion recognition ranged from 0.06 for neutral sentences to 0.24 for non-speech sounds (affect bursts), while for confidence ratings from 0.02 for semantic positive nouns to 0.07 for affect bursts. In general, the results from the full models indicate that non-speech sounds are best predicted by the acoustic parameters. Listeners' emotion recognition and confidence ratings were significantly driven by many acoustic parameters and the specific constellation of predictors was unique for each stimulus type. In non-speech sounds, for instance, listeners' recognition of emotions was predicted by amplitude parameters (peak amplitude, mean and standard deviation HNR), while their confidence in the given answer was explained by temporal parameters (duration). For Anna, semantic neutral nouns and neutral sentences, the recognition of emotions was also best predicted by amplitude-, as well as, by frequency-(minimum F_0 for Anna and standard deviation F_0 for neutral sentences) and temporal related parameters (duration for Anna and semantic neutral nouns). In contrast, for lexical sentences, semantic positive- and negative nouns, the recognition of emotions was best predicted by temporal- (duration) and frequency related parameters (duration, standard deviation F_0 and jitter accounted for more than 1.5% of explained variance). Although in pseudo-sentences emotion recognition was best predicted by temporal features (duration), in pseudo-words this perceptual property had no significant effect. Even though the proportion of explained variance was higher for emotion recognition than for confidence ratings, one could observe that

mial test fo	r listeners classi	fication erro	ors.							its of classification effors by failuoin forest (NF) affur fisteriers judgments of effortoil categories affur 55 70CF of the	5
Stimulus type E	Emotions portrayed	Emotion judgements	gements						Í	Classification errors	
		Anger	Fear	Happiness	Neutral	Sadness	Disgust	Surprise	Hu	Listeners [Cl _{95%}]	RF
Anna	Anger Fear	8 4.83 1.41	4.20 55.27	0.75	2.92	0.72	2.79	3.79	0.688	15.17% [13.94%; 16.46%] 44.73% [43.00%; 46.48%]	45.45%
. 4.	Happiness	15.36	10.25	28.81	3.98	6.02	5.08	30.50	0.257	71.19% [69.59%; 72.76%]	50.00%
× 8	Neutral Sadness	2.95	1.44	1.41	80.81 -	11.70	- 1.25	0.44	0.688	19.19% [17.84%; 20.60%] -	- 18.18% -
7	Disgust	ı	1	1	1	ı	ı	ı	1	ı	1
~ı	Surprise	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1		37 78%
Pseudo-words A	Anger	90.28	1.62	1.72	0.72	0.49	2.24	2.93	0.693	9.72% [8.70%; 10.86%]	52.63%
- I	Fear Hanniness	0.59	59.00 ○ 34	3.21 56 14	3.59 24.10	18.27	1.27	14.07	0.503	41.00% [39.20%; 42.82%] 43.86% [42.05%; 45.69%]	26.32%
. < '	Neutral	14.49	0.76	5.17	74.34	0.45	1.31	3.48	0.410	25.66% [24.08%; 27.30%]	40.00%
ν, Δ	Sadness Disgust	3.00 7.31	4.00 3.45	9.31 6.55	26.94 5.00	52.40 7.56	1.28 60.71	3.07 9.42	0.345 0.547	47.60% [45.77%; 49.44%] 39.29% [37.51%; 41.09%]	35.00%
SF	Surprise	ı	1	1	ı	ı	ı	ı		, , , , , ,	, ,
 Semantic negative A	I otal Anger	84.45	1.59	0.83	0.79	0.38	9.41	2.55	0.639	34.52% [33.82%; 35.24%] 15.55% [14.25%; 16.92%]	42.25% 35.00%
	Fear	0.66	75.23	2.52	0.52	12.07	1.00	8.00	0.604	[23.20%;	25.00%
	nappiness Neutral	9.00 9.00	6.14 0.62	3.59	81.07	1.00	1.59 2.07	17.38 2.65	0.530	35.79% [34.05%; 37.57%] 18.93% [17.52%; 20.41%]	80.00% 50.00%
S L	Sadness	1.14	5.24	2.24	28.39	60.26	0.83	1.90	0.460	39.74% [37.95%; 41.55%]	10.00%
	Disgust Surprise	13:21	00'+	20.0	0.10	67.6	66.60); -	0.430		0.00.00
Cemantic neutral	Total Anger	- 88 44	- 0.65	- 0.52	176	- 0.21	- 06.9	157	- 0 709	29.91% [29.23%; 30.59%]	40.19%
. —	Fear	0.59	73.71	4.38	1.41	9.04	0.83	10.04	0.650	26.28% [24.69%; 27.93%]	30.00%
. <	Happiness Neutral	0.66 7.03	1.69 0.17	67.38 2.14	11.48 88.11	0.62 0.45	0.55 0.86	17.62 1.24	0.536 0.513	32.62% [30.92%; 34.36%] 11.90% [10.74%; 13.13%]	65.00% 75.00%
S C	Sadness	0.62	3.03	3.35	41.70	49.09 3.86	0.52 54 50	1.69	0.381	50.91% [49.08%; 52.75%] 45.41% [43.59%; 47.25%]	30.00%
3 0 1	Surprise	20.5	t 7	0. 1	0 -)) 	ì,		t 0 1 0	,0/ 00.04	2000
 Semantic positive A	l otal Anger	82.41	0.97	1.10	1.45	0.21	10.72	3.14	0.643	29.78% [29.10%; 30.47%] 17.59% [16.22%; 19.02%]	44.38% 35.00%
nouns	Fear Lannings	0.49	73.72	3.93	1.10	10.10	0.69	9.97	0.673	26.28% [24.68%; 27.92%]	35.00%
	nappiness Neutral	8.38	0.35	4.62	80.41	1.04	1.17	4.03	0.503	19.59% [18.16%; 21.08%]	%00.09 60.00%
~, U	Sadness Disgust	0.90	2.83 2.24	7.14 10.07	27.63 6.00	58.81 3.90	0.65 57.86	2.04 9.03	0.460 0.466	41.19% [39.39%; 43.00%] 42.14% [40.33%; 43.96%]	15.00% 35.00%
ο F	Surprise Total	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	- 	34.78%
Affect bursts A	Anger	64.83	7.86	4.35	3.59	0.48	1.10	17.79	0.508		90.00%
	reur Happiness	0.62	0.07	96.76	2.62 0.48	0.69	0.00	1.38	0.869	3.24% [23.13%, 33.36%] 3.24% [2.39%; 4.29%]	%00:09 60:00%
< 0	Neutral Sadness	1.38	0.34	0.76 3.03	95.45	0.21 95.77	0.48	1.38	0.826	4.55% [3.54%; 5.75%] 4.28% [3.30%; 5.45%]	10.00%
, 7	Disgust	5.25	0.62	0.69	4.55	0.83	83.09	4.97	0.719	16.91% [15.01%; 18.94%]	20.00%
√;	Surprise Total	4.00	24.90	1.52 -	3.51	0.41	7.59	58.07	0.333	41.93% [39.38%; 44.52%] 19.66% [18.89%; 20.45%]	50.00%
Pseudo-sentences A	Anger	83.45	0.42	5.55	1.24	0.17	0.55	8.62	0.717	16.55% [15.22%; 17.96%]	30.00%
	-eur Happiness	4.41	0.76	60.14	3.76	0.17	0.41	30.35	0.275	43.34% [41.35%, 43.17%] 39.86% [38.07%; 41.67%]	55.00%
< 5	Neutral Sadness	0.24 0.41	0.31 12.39	3.17 1.55	91.73 3.38	0.79 77.99	0.24 3.11	3.52 1.17	0.670	8.28% [7.30%; 9.34%] 22.02% [20.52%; 23.57%]	30.00%
7	Disgust	3.14	10.07	5.59	15.62	12.97	47.27	5.34	0.388	52.72% [50.89%; 54.55%]	55.00%
·,	Surprise Total	2.00	co: -	14:41	0.82	0.20	1.0		0.1.0	90.41% [30.01%, 92.20%] 34.74% [34.09%; 35.40%]	42.59%

Lexical sentences	Emotions portraved	Emotion ju	Emotion judgements							Classification errors	
Lexical sentences	•	Anger	Fear	Happiness	Neutral	Sadness	Disgust	Surprise	H _u	Listeners [Cl _{95%}]	RF
	Anger	96.07	0.10	0.83	0.42	0.10	0.48	2.00	0.835	3.93% [3.25%; 4.70%]	20.00%
	Fear	0.52	77.72	0.97	2.93	13.93	0.41	3.52	0.731	22.28% [20.77%; 23.84%]	40.00%
	Happiness	0.56	0.03	75.17	12.59	0.10	0.17	11.38	0.397	24.83% [23.26%; 26.44%]	%00.06
	Neutral	1.03	0.73	1.38	91.69	3.14	0.62	1.41	0.684	8.31% [7.33%; 9.38%]	65.00%
	Sadness	0.45	3.24	0.83	3.97	90.23	0.62	99.0	0.752	9.76% [8.71%; 10.90%]	25.00%
	Disgust	8.52	69:0	5.11	10.28	0.72	67.51	7.17	0.652	32.49% [30.79%; 34.23%]	35.00%
	Surprise	3.41	0.17	58.14	0.97	0.07	0.10	37.14	0.218	62.86% [61.07%; 64.62%]	40.00%
	Tota/	1	1	1	1	1	1	1	1	23.50% [22.91%; 24.09%]	44.48%
Neutral sentences	Anger	97.28	0.10	1.17	0.79	00:0	0.17	0.49	0.851	2.72% [2.16%; 3.38%]	%00.09
	Fear	3.86	56.97	8.62	5.35	0.86	1.03	23.31	0.425	43.03% [41.22%; 44.86%]	35.00%
	Happiness	2.76	1.00	76.00	2.59	0.83	0.55	16.27	999:0	24.00% [22.46%; 25.60%]	75.00%
	Neutral	4.86	0.76	0.38	86.17	6:29	0.62	0.62	0.647	13.83% [12.60%; 15.14%]	15.00%
	Sadness	0.28	12.69	0.03	9.31	77.31	0.17	0.21	0.605	22.69% [21.18%; 24.26%]	30.00%
	Disgust	2.10	4.86	0.48	10.55	13.28	67.52	1.21	0.651	32.48% [30.78%; 34.22%]	65.00%
	Surprise	1	1	ı	1	1	1	1	ı		ı
	Total	1	1	1	1	1	1	1	1	23.13% [22.50%; 23.76%]	42.55%
Group Words	Anger	96.06	1.85	0.98	1.55	0.41	6.34	2.81	0.674	13.94% [13.39%; 14.51%]	41.58%
	Fear	0.76	67.15	3.03	2.86	14.01	1.03	11.16	0.567	32.85% [32.09%; 33.61%]	29.70%
	Happiness	4.56	3.94	55.77	11.76	2.00	1.77	20.20	0.422	44.24% [43.44%; 45.04%]	26.86%
	Neutral	8.26	69:0	3.35	80.94	3.09	1.33	2.34	0.520	19.06% [18.43; 19.70%]	35.29%
	Sadness	1.41	3.78	5.51	31.17	55.14	0.82	2.17	0.358	44.86% [43.95%; 45.77%]	26.25%
	Disgust	11.60	3.70	8.04	5.75	4.78	57.13	9.00	0.458	42.87% [41.97%; 43.78%]	38.75%
	Surprise	1	1	1	1	1		1	1		ı
	Tota/	1	1	1	1	1	1	1	ı	32.13% [31.81%; 32.45%]	37.54%
Group Sentences	Anger	88.35	1.30	2.78	1.21	0.15	0.50	5.71	0.758	11.66% [11.04%; 12.30%]	37.14%
	Fear	2.72	64.46	3.21	5.38	10.51	2.51	11.21	0.497	35.54% [34.61%; 36.48%]	48.57%
	Happiness	2.29	0.52	74.20	5.48	0.41	0.33	16.77	0.465	25.80% [24.95%; 26.67%]	51.43%
	Neutral	1.95	0.56	1.52	29.06	3.04	0.48	1.78	0.688	9.33% [8.77%; 9.92%]	42.86%
	Sadness	0.34	8.20	1.12	4.77	83.83	1.13	0.61	0.663	16.17% [15.46%; 16.90%]	42.86%
	Disgust	4.68	4.56	3.29	11.07	7.82	63.95	4.63	0.583	36.05% [35.11%; 36.99%]	67.14%
	Surprise	3.63	5.46	45.32	1.34	0.23	1.72	42.30	0.180	57.70% [56.55%; 58.84%]	52.00%
	Tota/	ı	ı	1	ı	ı	ı	1	ı	26.18% [25.85%; 26.51%]	48.13%

Note: The accuracy percentage of correctly judged portrayals are given on the main diagonal in boldface type. H_u = the rate of correctly identified stimuli. Classification errors of RF outside the 95%CI of listeners indicate a significant difference between the observed classification errors of listeners and RF. To calculate row, column and overall percentages see Table S2 in the supplementary material, which display the raw number of stimulus trials.

9

Table 4 Proportion of variance explained by the acoustic mea	ι of var	iance expl	ained by t	he acousti	c measure	s for emoti	isures for emotion recognition and confidence ratings by stimulus type.	tion and co	nfidence ra	atings by st	timulus typ	je.			
Stimulus types	⋖	Acoustic parameters	eters												Full model ^a
	, 	Frequency					Energy/amplitude	nde					Temporal		
	. <	MFO	Min _{FO}	Махғо	SDro	Jitter	Amp. (dB)	Peak amp.	Shimmer	MHNR	Maxhnr	SD _{HNR}	Duration	Peak time	
Emotion recognition AB An Ann Confidence LS SP PS SP P	e Z Z Z E E Z		0.0008 0.00065 0.00005 0.00006 0.00009 0.00009 0.000040 0.000074 0.00067 0.0016 0.0016 0.0016	0.0187*** 0.0051*** 0.0007 0.0004 0.0005 0.0008 0.0001 0.0005 0.0005 0.0005 0.0005 0.0005 0.0005	0.0024" 0.0024" 0.0014" 0.00330" 0.00330" 0.00018" 0.00018" 0.00018" 0.00014" 0.00014" 0.00027"	0.0012 0.0187*** 0.0187*** 0.0005 0.0005 0.0013*** 0.0013*** 0.00047*** 0.00045*** 0.00045*** 0.0006*** 0.00016***	0.0018* 0.0291*** 0.0269*** 0.0053*** 0.0037*** 0.0024*** 0.0017*** 0.0017*** 0.0017*** 0.0013*** 0.0013***	0.0660 0.0327 0.0327 0.0105 0.0015 0.0017 0.0019 0.0019 0.0002 0.0002 0.0002	0.0019*** 0.0139*** 0.0032*** 0.00039*** 0.00100*** 0.00100*** 0.0012*** 0.00012*** 0.00005*** 0.00005*** 0.00005***	0.0229**** 0.0126*** 0.0107*** 0.0003 0.0005** 0.0005*** 0.0005*** 0.0003**** 0.0003**** 0.0003**** 0.0001	0.0019"" 0.0038" 0.0038" 0.0011" 0.00025" 0.00011" 0.00057" 0.00057" 0.00067" 0.00067" 0.00067"	0.03577777777777777777777777777777777777	0.0006 0.0006 0.0005 0.0003 0.0013 0.0014 0.0018 0.0004 0.0004 0.0004 0.0007	0.0010 0.0131 0.0013 0.0003 0.0035 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003	0.2380"" 0.1484"" 0.1484"" 0.0927" 0.0027" 0.00597" 0.00597" 0.00597" 0.00597" 0.0372" 0.0372" 0.0372" 0.0372"
SPN 0.0008 0.0011 0.0007 0.0002 0.0038 0.0007 0.000	alues represe	0.0008° sent $pseudo-R^2$ for adjusted R^2 for $\frac{1}{10000000000000000000000000000000000$	0.0011	Ssion (see Cox ann. All p-values w	0.0002 and Snell, 1989 favere Bonferroni c	o.0038 or details) and w	0.0038 0.00037 0.0007 0.0020 0.0003 - 0.0045 0.0057 0.0057 0.0012 0.0217 0.0217 0.0038 0.00045 0.00057 0.00012 0.0217 0.0217 0.0012 0.0	0.0007° s follow: $1-e^{2^{\circ}(\text{dev})}$ itic parameters (*	0.0020*** (0) - dev (model))/c ***p < 0.001; **p *	0.0003 If, where dev(0) < 0.01; *p < 0.05)	is the deviance α . $AB = affect but a size AB = affect$	0.0045 of the null model rsts, SNeN = ser	0.0057 and dev (model mantic neutral m	0.0012" is the deviance ouns; $PW = pseu$	of the full model.

for shorter stimuli (semantic neutral-, negative and positive nouns) listeners confidence in the given answer was, in general, best predicted by temporal features (duration), while in longer stimuli (lexical and pseudo-sentences) by frequency related parameters (mean F_0 , minimum F_0 and jitter).

Stimuli- and emotion comparisons/confidence predicted by correct emotion recognition. The quasi-binomial and linear models revealed that stimuli types (p < 0.001), acoustic parameters (most p-values < 0.001) and emotions (p < 0.001) significantly influenced listeners' performance accuracy of recognizing emotions and their confidence judgements. Moreover, results showed that listeners' confidence judgements were significantly affected by the correct identification of emotions (p < 0.001). The corresponding test statistics from the global- (across all stimuli) and conditional models (for each stimulus type separately) are reported in supplementary material [see Tables S3 (A, A_I , A_{II}) for Group Words, A (A, A (A), A (A) for Group Words, A (A), A (A) for Group Sentences and A (A), A (A) for A (A) for each stimulus type].

Figure 2 illustrates the comparisons between stimuli types in both, Group Words and Group Sentences. As indicated by the OR estimates, listeners were significantly more accurate at recognizing emotions in stimuli with a semantic connotation than in those spoken in a language devoid of meaning (pseudo-speech) or expressing a person's name (Anna), with the latter being worst recognized. No significant differences in recognition accuracy were found in comparisons across the different stimuli types carrying semantic content (Fig. 2a). As indicated by the multiple comparisons of the estimated means, the pattern of the differences in confidence judgements occurred similar to the pattern of recognition accuracies. Listeners, however, were less confident when identifying emotions in semantically meaningful nouns spoken in neutral or positive compared to negative prosody (Fig. 2a_I). After adjusting for emotion recognition in the linear model, confidence ratings were significantly lower for pseudo-words than for Anna stimuli (Fig. 2a_{II}). The odds of correctly identifying emotions as well as listeners' confidence judgements were also significantly higher for affect bursts and lexical sentences than for neutral sentences and, lower for lexical sentences when compared to affect bursts (Fig. 2b, b_I). Recognition accuracy and confidence ratings were significantly lower for pseudo-sentences than for affect bursts, lexical- and neutral sentences (Fig. 2b, b_I). This pattern remained similar even after adjusting for emotion recognition in the linear model (Fig. 2b_{II}).

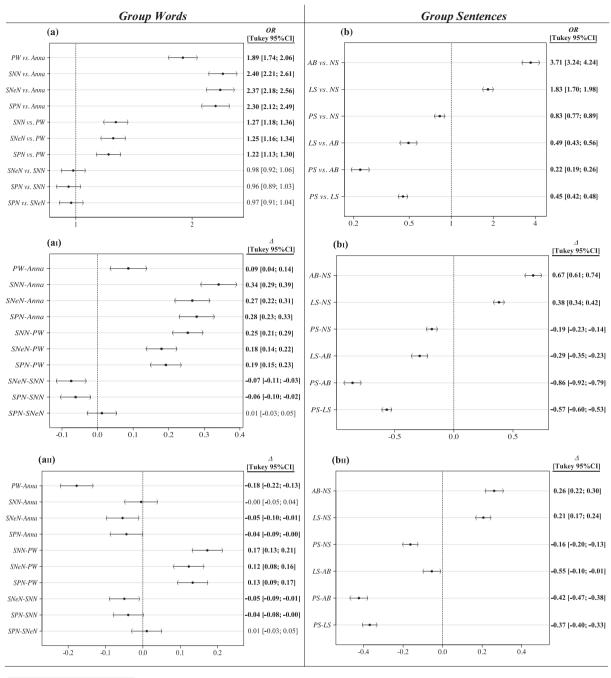
The comparison of performance accuracy across the emotion categories showed that in Group Words listeners were significantly less accurate and rated themselves as less confident when identifying emotional expressions spoken in a disgusted, neutral, fearful, happy, sad and surprised tone of voice than when spoken in an angry prosody. Although in Group Sentences, listeners were significantly more accurate at categorizing utterances spoken in a neutral- than in an angry tone of voice, the difference in confidence ratings was significantly lower for neutral than for angry expressions. In addition, results showed that both, recognition accuracy and confidence ratings, were significantly higher when comparing neutral to other emotional prosodies. Figure 3 illustrates the comparisons between emotion categories in Group Words and Group Sentences. The pattern of results obtained for the comparisons between emotion categories for each type of stimulus are presented in supplementary material (see Fig. S2A, A_I and A_{II} and Table S4 for a schematic overview).

Discussion

ratio.

proportion of variance, in the full model, explained by the acoustic measures in each stimulus type is displayed in decreasing order. No values were calculated for the parameters that were dropped by the backward variable selection

Human classification performance vs. RF. Using RF as a benchmark of listeners' performance accuracy, we observed that



Stimuli types

PW = Pseudo-words

SNN = Semantic negative nouns

SNeN = Semantic neutral nouns

SPN = Semantic positive nouns

AB = Affect bursts

LS = Lexical sentences

NS = Neutral sentences

PS = Pseudo-sentences

Fig. 2 Comparisons between stimuli types. Emotion recognition *odds ratio estimates* (OR) for the comparisons between stimuli types are illustrated in panel **a** and panel **b**. The *linear contrasts* (Δ) for confidence ratings are illustrated in panel **a**_{II} and **b**_{II}, while the *confidence ratings after adjusting for emotion recognition* are displayed in panel **a**_{II} and **b**_{II}. Odds ratio of stimulus 1 (e.g., AB) vs. stimulus 2 (e.g., NS) less than 1 indicate that the recognition probability of stimulus 2 (e.g., NS) is higher than of stimulus 1 (e.g., AB), whereas values greater than 1 vice-versa. If the odds ratio of 1 is covered in the confidence interval, the difference in the recognition probabilities is not significant. Negative differences of confidence ratings of stimulus 1 (e.g., AB) vs. stimulus 2 (e.g., NS) indicate that the confidence ratings of stimulus 2 (e.g., NS) is higher than of stimulus 1 (e.g., AB), whereas positive differences vice-versa. If the difference of zero is covered in the 95%CI, the difference in the confidence ratings is not significant. The significant combinations are highlighted in bold.

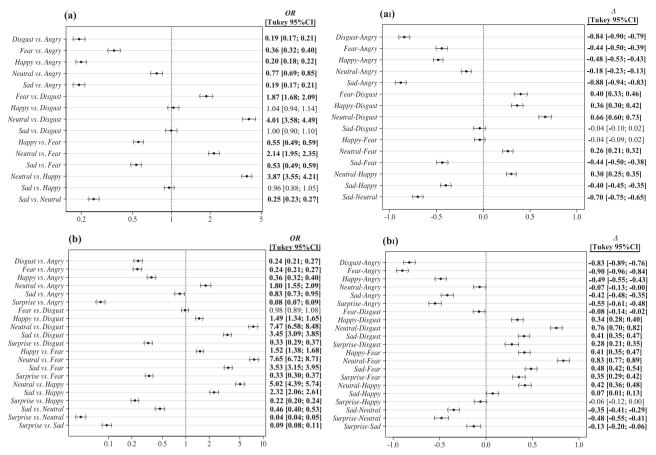


Fig. 3 Comparisons between emotion categories in Group Words (GW) and Group Sentences (GS). The *odds ratio* estimates (OR) for the comparisons between emotion categories are illustrated in panel **a** for GW and panel **b** for GS, while the *linear contrasts* (Δ) for confidence ratings are illustrated in panel **a**₁ for GW and **b**₁ for GS. Odds ratio of emotion 1 (e.g., *disgust*) vs. emotion 2 (e.g., *angry*) less than 1 indicate that the recognition probability of emotion 2 (e.g., *angry*) is higher than of emotion 1 (e.g., *disgust*), whereas values greater than 1 vice-versa. If the odds ratio of 1 is covered in the confidence interval, the difference in the recognition probabilities is not significant. Negative differences of confidence ratings of emotion 1 (e.g., *disgust*) vs. emotion 2 (e.g., *angry*) indicate that the confidence ratings of emotion 2 (e.g., *angry*) is higher than of emotion 1 (e.g., *disgust*), whereas positive differences vice-versa. If the difference of zero is covered in the 95%CI, the difference in the confidence ratings is not significant. The significant combinations are highlighted in bold.

despite the fact that RF error patterns were significantly lower for certain emotions than those of listeners, across all emotions the automatic classification of emotions was considerably less successful than listeners' classification performance. Listeners' superior performance may be due to the fact that they could draw on a greater number of emotional markers [e.g., intonation patterns, emblems of distinct emotions (i.e., laughter, sighs), valence, arousal] inaccessible to the statistical algorithm, whose emotion category predictions were based on a fully automatic set of acoustic parameters extracted from relatively small learning datasets. Although classification algorithms do not seem to replicate the inference processes of human decoders, they appear to have lower error rates than listeners' when classifying certain emotions (e.g., disgust) solely based on their acoustic profiles. This has been indicated by previous work (Banse and Scherer, 1996) as well as our current dataset.

Vocal stimulus types predicted by acoustic parameters. In line with previous findings (e.g., Lima et al., 2013; Sauter, 2006), our results clearly indicated that the recognition of emotions from vocal bursts is much more dependent on the preservation of acoustic information (amplitude, pitch) than speech-embedded stimuli. This result corresponds to the idea that vocal bursts are more primitive and salient signals of emotion than speech-

embedded vocalizations in an evolutionary sense (Krumhuber and Scherer, 2011). Since they do not require the dynamic spectral shaping caused by the rapid movements of the articulators (i.e., the tongue, jaw, lips and soft palate which shape the sound produced at the larynx), it has been suggested that vocal bursts resemble animal vocalizations more than they do spoken language (Krumhuber and Scherer, 2011; Scott et al., 2010). For instance, laughter has been described as more akin to modified breathing, involving inhalation as well as exhalation, than to speaking (Kohler, 2008). In addition, we found that neutral words (i.e., Anna; semantic neutral nouns) are better predicted by acoustic parameters than other speech-embedded stimuli. This fits well with the idea that words, similar to affect bursts, are arbitrary combinations of phonemes that derive their valence from what they symbolize (Schirmer, 2010).

Although amplitude-, frequency-, and temporal related parameters are important in the emotional inflection of spoken language (Baenziger and Scherer, 2005; Banse and Scherer, 1996), as shown by our data the proportion of variance explained by these measures is slightly lower for sentences or stimuli with a semantic negative- or positive connotation. While these differences may be due to the quality of our speech-embedded stimuli, the difference between emotional speech and vocal bursts may also reflect the fact that emotion in speech is overlaid on the speech signal. This would mean speech is somewhat more

constrained in its emotional expression than are non-verbal vocalizations (Scott et al., 2010). Thus, there could be conflicts between the prosodic cues in sentence-level speech, which denote the emotional information, and those that cue linguistic information. Indeed, research has suggested that in speechembedded stimuli emotional information derived from semantics is somehow stronger than that conveyed by prosodic cues (e.g., Pell et al., 2011; Paulmann and Kotz, 2008). Similar to previous studies on perceptual-acoustic correlates of the confidence that is expressed in the voice of the speaker (e.g., Jiang and Pell, 2017; Rigoulout et al., 2013; Pell and Kotz, 2011), our findings highlight that listeners confidence judgements are predicted by distinct sets of acoustic parameters. Of note, in shorter stimuli (i.e., affect bursts, nouns) duration was a significant predictor of listeners' confidence judgements, while for longer stimuli (pseudo-, lexicaland neutral sentences) the fundamental frequency parameters explained the largest amount of variance in confidence ratings.

Stimuli- and emotion comparisons/confidence predicted by correct emotion recognition. Results from both logistic and linear models, implemented in our study, revealed that most of the acoustic predictors and vocal stimuli types had a significant influence on listeners' recognition of emotions and their confidence judgements. Moreover, our findings showed that correct judgements of emotions elicited confident interpretations. These findings will be discussed in detail in the following.

The pattern of results revealed in our study clearly indicated that listeners were significantly more accurate and confident at judging emotions from non-speech sounds (i.e., affect bursts) than speech-embedded stimuli (i.e., sentences/pseudo-sentences). This finding adds to previous research which demonstrated that affect bursts are decoded more accurately than speech-embedded prosody (Hawk et al., 2009). Further evidence comes from neurophysiological studies, showing that non-speech sounds facilitate early stages of perceptual processing in the form of decreased N1 amplitudes and enhanced P2 amplitudes in eventrelated brain potentials (Pell et al., 2015; Liu et al., 2012). In other words, affect bursts, as evolutionary fundamental signals, seem to evoke a more rapid capture of attention than speech-embedded stimuli thought to involve more effortful cognitive processes and acoustic analysis (Pell et al., 2015) for successful recognition. Similar to recent findings (Schaerlaeken and Grandjean, 2018), our data indicated that the accurate decoding of affect bursts led to greater confidence judgements. In light of the abovementioned studies, one could argue that these types of stimuli carry more ecologically relevant information and, thus are given precedence by the neurocognitive system, allowing individuals to be more accurate and confident in their judgements of desirable/ undesirable events in their environment. In addition, our results demonstrated that for stimuli with lexico-semantic content (i.e., nouns, lexical and neutral sentences) the accuracy of performance and confidence ratings is significantly higher than for stimuli devoid of meaning (i.e., pseudo-words; pseudo-sentences). Support for these findings comes from validation studies reporting greater accuracy for semantic compared to pseudospeech utterances (e.g., Castro and Lima, 2010) and from studies showing that congruent combinations of semantics and prosody, yield not only greater accuracy (e.g., Ben-David et al., 2016) but, as shown by the patterns in our study, also higher confidence ratings compared to neutral sentences. When comparing stimuli with a semantic positive, negative and neutral content, no significant differences in recognition accuracy were observed, yet, listeners felt significantly more confident at detecting emotions in semantic negative than semantic positive or neutral single words (nouns). One explanation that has been put forth as to why such a

negativity bias occurs in social judgements is that people may generally consider negative information to be more diagnostic than positive information in forming an overall impression (e.g., Hamilton and Huffman, 1971). This is supported by studies showing that people consider negative information to be more important to impression formation and, when such information is available to them, they are subsequently more confident (e.g., Baumeister et al., 2001; Hamilton and Zanna, 1972). Another possible explanation is that negative stimuli exert a stronger influence on people's evaluations because these stimuli are more complex than positive ones and, thus, require greater attention and cognitive processing (e.g., Ito et al., 1998; Peeters and Czapinski, 1990; Abele, 1985). Correct recognition seemed to determine an increase of confidence judgements, however, when comparing Anna stimuli to other stimuli types, we observed that, when correctly categorizing the emotions for this stimulus type, listeners' felt more confident regarding the correctness of their answer. As listeners' decisions had to be made on the same item repeatedly presented (i.e., the name Anna), one could argue that this might have led to a familiarity effect with the stimulus. Studies on metacognition have shown that item familiarity leads to higher confidence ratings, because participants rest on the belief that more knowledge about the item means they are more accurate, although this has been shown to be an irrelevant factor (e.g., Koriat, 2008; Metcalfe et al., 1993).

Another important finding of the current study was that, for a vast majority of the stimuli, listeners' performance accuracy and confidence ratings were significantly higher when spoken in an angry and neutral tone of voice than in any other emotional prosody. In contrast, recognition accuracy and confidence ratings were lowest for disgust, excepting affect bursts. One interesting pattern that emerged from our data was that listeners felt more confident at categorizing surprise than disgust and fear (i.e., pseudo- and lexical sentences) or happy (i.e., lexical sentences), despite the fact that their performance accuracy was significantly lower for the latter emotion. Moreover, we found that for nouns, fearful prosody led to higher accuracy rates and confidence ratings than sadness, although for other types of stimuli the exact opposite pattern was observed. Happiness yielded higher accuracy scores and confidence ratings in comparison to disgust, while when compared to other emotions this largely depended on the type of stimulus (for an overview, see Table S4 in supplementary material). Previous studies on prosody revealed conflicting evidence regarding the differences between emotional and neutral prosody. They reported stimuli spoken in a neutral tone to be identified more accurately and with higher confidence ratings compared to other emotional prosodies, regardless of semantics (Cornew et al., 2009; Schirmer and Kotz, 2003). Our findings converge with previous work on vocal emotion recognition, together indicating a general advantage for the recognition of neutral prosody, but also of angry expressions compared to other vocal expressions (e.g., Cowen et al., 2019b; Chronaki et al., 2018; Paulmann and Uskul, 2014; Pell et al., 2009; Scherer et al., 2001). This evidence is also compatible with evolutionary theories arguing that humans as well as other primates are biologically prepared to respond rapidly to vocal cues associated with threat or anger (e.g., Oehman, 1993). It has been argued that particularly non-speech sounds (e.g. growls of anger, the laughter of happiness or cries of sadness) convey emotions clearer and faster than words (Pell et al., 2015). Our results, however, showed that when identifying emotions from affect bursts, listeners were less accurate and felt less confident at detecting anger when compared to other emotional prosodies. Previous findings report similar accuracy patterns when decoding anger from affect bursts (Pell et al., 2015; Belin et al., 2008). However, it remains unclear why this effect emerged. A possible explanation relates to the acoustics

of acting anger sounds which might differ from natural ones. Anikin and Lima (2018), for instance, reported that authentic vocalizations (e.g., anger, fear) differed from actor portrayals in a number of acoustic characteristics by showing a higher pitch and lower harmonicity, as well as, a less variable spectral slope and amplitude (but also see, Juergens et al., 2015, 2011, for a discussion on authentic vs. play acted expressions). Thus, it might be plausible that these acoustic characteristics of authenticity are hard-to-fake markers of a speaker's emotional state and, thus, signal a distinction between honest communication and a bluff (Anikin and Lima, 2018). The patterns obtained for disgust are also consistent with those of previous studies showing that, after surprise, it is the most difficult emotional expression to be recognized from speech-embedded stimuli (e.g., Paulmann and Uskul, 2014) but not from non-speech sounds, presumably, because this emotion is merely expressed in affect bursts or short interjections (e.g., yuck) rather than in sentential context (Johnstone and Scherer, 2000; Banse and Scherer, 1996). In contrast, surprise yielded higher confidence scores than disgust or fear, which could be due to the fact that humans are more prone to notice and focus on surprising events and, therefore are more likely to attend to them (Wilson and Gilbert, 2008). Following these arguments, one could speculate that similar to anger, surprise might also serve a functional and adaptive purpose as people might devote their energy to judging whether what is unfolding before them is a threat, a joke or a harmless event, thus, eliciting more confident evaluations. A similar argument may apply to our results regarding the comparisons between fear and sadness. Fear, as an expression that signals threat, might require less auditory input to be decoded accurately (i.e., shorter stimuli -in our case nouns), while identifying sadness from speech might activate additional social meanings that take more time to analyse and more careful post-message processing.

To summarize, this set of results extends previous findings from the facial domain (Kelly and Metcalfe, 2011) by showing that listeners who were better at recognizing vocal expressions of emotion were also more confident in their judgements. Although slight variations between emotion recognition accuracy and confidence ratings were observed for some stimuli types or emotion categories, overall our results support that the correct recognition of emotions increases confident judgements. Our findings suggest that individuals can predict and assess their performance for the recognition of emotional prosody.

General discussion

In the present investigation, we aimed to examine the influence of specific types of vocal stimuli and their acoustic parameters on emotional prosody recognition and retrospective confidence judgements. Previous research has suggested that when investigating the effects of prosody on emotion recognition it is critical to study a wide variety of stimuli to determine the full dimensionality of that space (e.g., Cowen et al., 2019a; Sauter et al., 2010). To cover the spectrum of materials used in emotional prosody research, we studied a vast array of vocal stimuli (i.e., for speech: words, lexical and neutral sentences; pseudo-speech: pseudo-words/sentences; for non-speech: vocalizations) from which we extracted a standard set of acoustic parameters considering previous findings that emphasized the importance of paralinguistic emotional content in verbal communication (e.g., Eyben et al., 2016).

In two studies, we examined (1) the extent to which listeners and classifiers use acoustic parameters as perceptual cues for identifying portrayed emotions and (2) whether listeners' performance accuracy and confidence ratings vary across certain types of stimuli and specific emotion categories. To determine

whether the set of extracted parameters provided sufficient information to successfully discriminate between stimuli from different emotional categories, discriminant analysis and RF were implemented (study 1). By employing a backward stepwise logistic regression analysis, we determined which of the acoustic predictors explained most of the deviance in listeners' recognition rates (study 2). Results showed high cross-validation estimates of accuracy for both classification methods (study 1), indicating that the stimuli contained detectable acoustic contrasts which helped listeners to differentiate the portraved emotions and that most, if not all, parameters explained a significant amount of variance in listeners' recognition rates (study 2). This set of results corresponds to previous findings in the vocal emotion literature (e.g., Cowen et al., 2019a; Hammerschmidt and Juergens, 2007; Banse and Scherer, 1996) and, in analogy, they parallel research on visual signals of emotions which reported that statistical classification methods can successfully discriminate facial expressions of different emotions on the basis of their pixel intensities (e.g., Calder et al., 2001). Thus, for both vocal and facial modalities, it is possible to classify emotional expressions on the basis of basic perceptual features in a manner that models human performance.

With respect to the comparisons between the types of stimuli and emotion categories, we uncovered that affect bursts are richer and more nuanced than typically thought (study 2). Although our data showed that emotion and confidence ratings could be reliably predicted from speech-embedded stimuli acoustic attributes, we found that in affect bursts the acoustic parameters explained the largest proportion of variance. Moreover, our study revealed that vocal bursts proved to be highly effective means of expressing specific emotions, such as disgust, happiness and sadness, in comparison to speech-embedded stimuli, with recognition accuracies above 80% (except for surprise). Results from crosscultural studies corroborate these findings by reporting strong to moderate evidence for the universal recognizability of these emotions from vocal bursts (e.g., Cowen et al., 2019a; Cordaro et al., 2016). Thus, one could ask whether there is a reason why these specific emotions are better recognized in this type of vocal stimuli? An argument that has been put forth is that vocal bursts are unique to some emotions (Goddard, 2014). For instance, laughter could be interpreted as a signal of happiness, crying as a signal of sadness, while interjections such as 'argh', 'eek' are typically indicative of disgust. Moreover, it has been suggested that people quite rarely vocalize disgust or surprise in the form of sentences (Schaerlaeken and Grandjean, 2018; Banse and Scherer, 1996). Since vocal bursts bear a heavy functional load in social interactions, as they are "so highly overlearned" and clearly attached to certain emotions (Goddard, 2014; Scherer, 1994), their accurate recognition might occur instantaneously and without conscious effort. Similar to previous interpretations (Cowen et al., 2019a), our results regarding the efficiency of vocal bursts in emotion recognition may be explained by the innate psychological basicness of these signals (which might be universal to all humans).

Limitations and future research. Although the main groups of paralinguistic features and their acoustic parameters (Eyben et al., 2016; Juslin and Laukka, 2003), were covered in our study, not all relevant properties of vocal emotional expression have been considered. Thus, future research would profit by implementing, for instance, spectral parameters or prosodic contours, as they might index physiological changes in voice and are sensitive to emotional expressions (e.g., Eyben et al., 2016; Mozziconacci, 2002). A limitation to this issue is the fact that parameters were extracted from the entire utterances. Although this is a common approach (Pichora-Fuller et al., 2016; Castro and Lima, 2010;

Paulmann et al., 2008), it has been supposed that it disregards the phonetic identity of speech segments in emotional expression (Goudbeek and Scherer, 2010). There is also the proviso that gender of the speaker may have had an effect on the discrimination accuracy (Lausen and Schacht, 2018). The present study, however, focused on the patterns of voice cues used to portray specific emotions, rather than on gender differences. By keeping in line with previous research in this area we extracted the acoustic parameters across both genders (e.g., Sauter et al., 2010; Paulmann et al., 2008; Juslin and Laukka, 2001). This does not rule out that gender might have had an effect, however, it should be noted that for the majority of the stimulus types the emotions were expressed by two speakers of different sexes (1 male, 1 female). Future work with a greater number of speakers would not only be able to establish the degree to which the acoustic factors in this study can be generalized but, would also help to explain the variation in these factors alongside speakers' gender characteristics. A further limitation regards the absence of some emotional categories within our stimuli datasets. In comparison to the classification algorithms that categorized emotions based on the existing number of emotion categories, listeners were supposed to choose from a fixed set of given alternatives. This might explain why for certain emotion categories, listeners had higher error rates than the RF algorithm. Moreover, machinelearning efforts to decode emotion from speech stimuli will need to expand in focus beyond a small set of discrete categories. Finally, the sample in the current study was limited to a university-educated population and included predominantly young adults, which may limit the generalizability of the findings to the wider population.

Conclusion

The present study provided a systematic investigation of the influence prosodic parameters and stimulus types exert on vocal emotion recognition and confidence ratings. Addressing a timely issue with the abundance of vocal emotion corpora, our findings are essential both empirically and conceptually. First, they replicated earlier research by establishing that humans can infer emotion from vocal expression, based on differential acoustic patterning. Second, our results extend previous findings by indicating that emotional expressions are more accurately recognized and confidently judged from non-speech sounds than from semantically inflected. Finally, and most importantly, our findings yield data-driven hypotheses into important questions within emotion science (e.g., Sauter et al., 2010) by showing that this pattern is not constant across all emotional categories and that listeners do not rely on the same acoustic cues when decoding emotions from speech and nonspeech embedded sounds. While the current findings demonstrate that correct recognition of emotions promotes confident interpretations, more research is needed to uncover the underlying mechanisms of how individuals use this metacognitive knowledge.

Data availability

The datasets generated during and/or analysed during the current study are available in the Open Science Framework (OSF) repository: https://doi.org/10.17605/OSF.IO/J695T

Received: 10 October 2019; Accepted: 19 May 2020; Published online: 17 June 2020

References

Abele A (1985) Thinking about thinking: causal, evaluative and finalistic cognitions about social situations. Eur J Soc Psychol 15:315–332. https://doi.org/10.1002/ejsp.2420150306

- Anikin A, Lima CF (2018) Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. Q J Exp Psychol 71:622–641. https://doi.org/10.1080/17470218.2016.1270976
- Bak HK (2016) The state of emotional prosody research—a meta-analysis. In: Bak HK (ed) Emotional prosody processing for non-native English speakers, 1st edn. Springer International Publishing, pp. 79–112
- Banse R, Scherer KR (1996) Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 70:614–636. https://doi.org/10.1037/0022-3514.70.3.614
- Baumeister RF, Bratslavsky E, Finkenauer C et al. (2001) Bad is stronger than good. Rev Gen Psychol 5:323–370. https://doi.org/10.1037//1089-2680.5.4.323
- Baenziger T, Scherer KR (2005) The role of intonation in emotional expressions. Speech Commun 46:252–267. https://doi.org/10.1016/j.specom.2005.02.016
- Bègue I, Vaessen M, Hofmeister J et al. (2019) Confidence of emotion expression recognition recruits brain regions outside the face perception network. Soc Cogn Affect Neurosci 4:81–95. https://doi.org/10.1093/scan/nsy102
- Belin P, Fillion-Bilodeau S, Gosselin F (2008) The Montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. Behav Res Methods 40:531–539. https://doi.org/10.3758/BRM.40.2.531
- Ben-David BM, Multani N, Shakuf V et al. (2016) Prosody and semantics are separate but not separable channels in the perception of emotional speech: test for rating of emotions in speech. J Speech Lang Hear Res 59:1–18. https://doi.org/10.1044/2015_jslhr-h-14-0323
- Bostanov V, Kotchoubey B (2004) Recognition of affective prosody: continuous wavelet measures of event-related brain potentials to emotional exclamations. Psychophysiology 41:259–268. https://doi.org/10.1111/j.1469-8986.2003.00142.x
 Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1033/
- Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/ A:1010933404324
- Burkhardt F, Paeschke A, Rolfes M et al. (2005) A database of German emotional speech. In: European conference on speech and language processing, Lisbon, Portugal, pp. 1517–1520. https://www.researchgate.net/publication/221491017_A_database_of_German_emotional_speech. Accessed 10 Nov 2015
- Calder AJ, Burton AM, Miller P et al. (2001) A principal component analysis of facial expressions. Vis Res 41:1179–1208. https://doi.org/10.1016/S0042-6989 (01)00002-5
- Castro SL, Lima CF (2010) Recognizing emotions in spoken language: a validated set of Portuguese sentences and pseudosentences for research on emotional prosody. Behav Res Methods 42:74–81. https://doi.org/10.3758/BRM.42.1.74
- Chronaki G, Wigelsworth M, Pell MD et al. (2018) The development of crosscultural recognition of vocal emotions during childhood and adolescence. Sci Rep. 8:8659. https://doi.org/10.1038/s41598-018-26889-1
- Cordaro DT, Keltner D, Tshering S et al. (2016) The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. Emotion 16:117–128. https://doi.org/10.1037/emo0000100
- Cornew L, Carver L, Love T (2009) There's more to emotion than meets the eye: a processing bias for neutral content in the domain of emotional prosody. Cogn Emot 24:1133–1152. https://doi.org/10.1080/02699930903247492
- Cowen AS, Laukka P, Elfenbein HA et al. (2019a) The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. Nat Hum Behav 3:369–382. https://doi.org/10.1038/s41562-019-0533-6
- Cowen AS, Elfenbein HA, Laukka P et al. (2019b) Mapping 24 emotions conveyed by brief human vocalization. Am Psychol 74:698–712. https://doi.org/10.1037/amp0000399
- Cox DR, Snell EJ (1989) Analysis of binary data, 2nd edn. Chapman & Hall Dixon SJ, Brereton RG (2009) Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure. Chemom Intell Lab Syst 95:1–17. https://doi.org/10.1016/j.chemolab.2008.07.010
- Dunlosky J, Metcalfe J (2009) Confidence judgements. In: Dunlosky J, Metcalfe J (eds) Metacognition, 1st edn. Sage Publications, Washington, pp. 118–139
- Eyben F, Scherer KR, Schuller BW et al. (2016) The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7:190–202. https://doi.org/10.1109/TAFFC.2015.2457417
- Goddard C (2014) Interjections and emotion (with special reference to "surprise" and "disgust"). Emot Rev 6:53–63. https://doi.org/10.1177/1754073913491843
- Goudbeek M, Scherer KR (2010) Beyond arousal: valence and potency/control cues in the vocal expression of emotion. J Acoust Soc Am 128:1322–1336. https:// doi.org/10.1121/1.3466853
- Hamilton DL, Huffman LJ (1971) Generality of impression-formation processes for evaluative and nonevaluative judgments. J Pers Soc Psychol 20:200–207. https://doi.org/10.1037/h0031698
- Hamilton DL, Zanna MP (1972) Differential weighting of favorable and unfavorable attributes in impressions of personality. J Exp Res Pers 6:204–212
- Hammerschmidt K, Juergens U (2007) Acoustical correlates of affective prosody. J Voice 21:531–540. https://doi.org/10.1016/j.jvoice.2006.03.002
- Hawk ST, van Kleef GA, Fischer AH et al. (2009) "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. Emotion 9:293–305. https://doi.org/10.1037/a0015178

- Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biom J 50:346–363. https://doi.org/10.1002/bimj.200810425
- Ito TA, Larsen JT, Smith NK et al. (1998) Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. J Pers Soc Psychol 75:887–900. https://doi.org/10.1037/0022-3514.75.4.887
- James G, Witten D, Hastie T et al. (2013) An introduction to statistical learning with applications in R. In: Cassella G, Fienberg S, Olkin I (eds) Springer texts in statistics. Springer, New York, pp. 303–332
- Jiang X, Pell MD (2017) The sound of confidence and doubt. Speech Commun 88:106–126. https://doi.org/10.1016/j.specom.2017.01.011
- Jiang X, Pell DM (2014) Encoding and decoding confidence information in speech. In: Proceedings of the 7th international conference in speech prosody (social and linguistic speech prosody). pp. 573–576. http://fastnet.netsoc.ie/sp7/sp7book.pdf. Accessed 30 Nov 2018
- Johnstone T, Scherer KR (2000) Vocal communication of emotion. In: Lewis M, Haviland J (eds) The handbook of emotion, 2nd edn. Guildford, New York, pp. 220–235
- Juslin PN, Scherer KR (2005) Vocal expression of affect. In: Harrigan JA, Rosenthal R, Scherer KR (eds) The new handbook of methods in nonverbal behavior research, 1st edn. Oxford University Press, Oxford, pp. 65–135
- Juslin PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: different channels, same code? Psychol Bull 129:770–814. https://doi.org/10.1037/0033-2909.129.5.770
- Juslin PN, Laukka P (2001) Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. Emotion 1:381–412. https://doi.org/10.1037//1528-3542.1.4.381
- Juergens R, Fischer J, Schacht A (2018) Hot speech and exploding bombs: autonomic arousal during emotion classification of prosodic utterances and affective sounds. Front Psychol 9:228. https://doi.org/10.3389/fpsyg.2018.00228
- Juergens R, Grass A, Drolet M et al. (2015) Effect of acting experience on emotion expression and recognition in voice: non-actors provide better stimuli than expected. J Nonverbal Behav 39:195–214. https://doi.org/10.1007/s10919-015-0209-5
- Juergens R, Drolet M, Pirow R et al. (2013) Encoding conditions affect recognition of vocally expressed emotions across cultures. Front Psychol 4:111. https:// doi.org/10.3389/fpsyg.2013.00111
- Juergens R, Hammerschmidt K, Fischer J (2011) Authentic and play-acted vocal emotion expressions reveal acoustic differences. Front Psychol 2:180. https:// doi.org/10.3389/fpsyg.2011.00180
- Kelly KJ, Metcalfe J (2011) Metacognition of emotional face recognition. Emotion 11:896–906. https://doi.org/10.1037/a0023746
- Kimble C, Seidel S (1991) Vocal signs of confidence. J Nonverbal Behav 15:99–105. https://doi.org/10.1007/BF00998265
- Kitayama S, Ishii K (2002) Word and voice: spontaneous attention to emotional speech in two cultures. Cogn Emot 16:29–59. https://doi.org/10.1080/ 0269993943000121
- Kohler KJ (2008) 'Speech-smile', 'speech-laugh', 'laughter' and their sequencing in dialogic interaction. Phonetica 65:1–18. https://doi.org/10.1159/000130013
- Koriat A (2008) When confidence in a choice is independent of which choice is made. Psychon Bull Rev 15:997–1001. https://doi.org/10.3758/PBR.15.5.997
- Kotz SA, Paulmann S (2007) When emotional prosody and semantics dance cheek to cheek: ERP evidence. Brain Res 1151:107–118. https://doi.org/10.1016/j. brainres.2007.03.015
- Krumhuber EG, Scherer KR (2011) Affect bursts: dynamic patterns of facial expression. Emotion 11:825–841. https://doi.org/10.1037/a0023856
- Lausen A, Schacht A (2018) Gender differences in the recognition of vocal emotions. Front Psychol 9:882. https://doi.org/10.3389/fpsyg.2018.00882
- Lima CF, Castro SL, Scott SK (2013) When voices get emotional: a corpus of nonverbal vocalizations for research on emotion processing. Behav Res Methods 45:1234–1245. https://doi.org/10.3758/s13428-013-0324-3
- Liu T, Pinheiro AP, Deng G et al. (2012) Electrophysiological insights into processing nonverbal emotional vocalizations. Neuroreport 23:108–112. https://doi.org/10.1097/WNR.0b013e32834ea757
- Metcalfe J, Schwartz BL, Joaquim SG (1993) The cue-familiarity heuristic in metacognition. J Exp Psychol Learn Mem Cogn 19:851–861. https://doi.org/ 10.1037//0278-7393.19.4.851
- Mitchell RLC, Elliott R, Barry M et al. (2003) The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. Neuropsychologia 41:1410–1421. https://doi.org/10.1016/S0028-3932(03)00017-4
- Mozziconacci S (2002) Prosody and emotions. In: Proceedings of speech prosody, pp. 1–9. https://www.isca-speech.org/archive/sp2002/. Accessed 30 Nov 2018
- Noroozi F, Sapiński T, Kamińska D et al. (2017) Vocal-based emotion recognition using random forests and decision tree. Int J Speech Technol 20:239–246. https://doi.org/10.1007/s10772-017-9396-2
- Norsonic Nor140 (2017) Instruction manual. Lierskogen, Norway. https://www.campbell-associates.co.uk/norsonic-140-sound-level-meter

- Nygaard L, Lunders E (2002) Resolution of lexical ambiguity by emotional tone of voice. Mem Cogn 30:583–593. https://doi.org/10.3758/BF03194959
- Owren MJ (2008) GSU Praat tools: scripts for modifying and analyzing sounds using Praat acoustics software. Behav Res Methods 40:822–829. https://doi. org/10.3758/BRM.40.3.822
- Oehman A (1993) Fear and anxiety as emotional phenomenon: clinical phenomenology, evolutionary perspectives, and information-processing mechanisms. In: Lewis M, Haviland JM (eds) Handbook of emotions. Guildford Press, New York, pp. 511–536
- Parsons CE, Young KS, Craske MG et al. (2014) Introducing the Oxford Vocal (OxVoc) Sounds database: a validated set of non-acted affective sounds from human infants, adults, and domestic animals. Front Psychol 5:562. https:// doi.org/10.3389/fpsyg.2014.00562
- Paulmann S (2016) The neurocognition of prosody. In: Hickok G, Small S (eds) Neurobiology of language. Elsevier, San Diego, pp. 1109–1120
- Paulmann S, Uskul AK (2014) Cross-cultural emotional prosody recognition: evidence from Chinese and British listeners. Cogn Emot 28:230–244. https://doi.org/10.1080/02699931.2013.812033
- Paulmann S, Kotz SA (2008) An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical sentence context. Brain Lang 105:59–69. https://doi.org/10.1016/j.bandl.2007.11.005
- Paulmann S, Pell MD, Kotz SA (2008) How aging affects the recognition of emotional speech. Brain Lang 104:262–269. https://doi.org/10.1016/j. bandl.2007.03.002
- Pell MD, Rothermich K, Liu P et al. (2015) Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. Biol Psychol 111:14–25. https://doi.org/10.1016/j.biopsycho.2015.08.008
- Pell MD, Kotz SA (2011) On the time course of vocal emotion recognition. PLoS ONE 6(11):e27256. https://doi.org/10.1371/journal.pone.0027256
- Pell MD, Jaywant A, Monetta L et al. (2011) Emotional speech processing: disentangling the effects of prosody and semantic cues. Cogn Emot 25:834–853. https://doi.org/10.1080/02699931.2010.516915
- Pell MD, Monetta L, Paulmann S et al. (2009) Recognizing emotions in a foreign language. J Nonverbal Behav 33:107–120. https://doi.org/10.1007/s10919-008-0065-7
- Pell MD, Paulmann S, Dara C et al. (2009) Factors in the recognition of vocally expressed emotions: a comparison of four languages. J Phon 37:417–435. https://doi.org/10.1016/j.wocn.2009.07.005
- Peeters G, Czapinski J (1990) Positive–negative asymmetry in evaluations: the distinction between affective and informational negativity effects. In: Stroebe W, Hewstone M (eds) European review of social psychology, vol. 1. Wiley, Chichester, pp. 33–60
- Pichora-Fuller MK, Dupuis K, Van Lieshout P (2016) Importance of F0 for predicting vocal emotion categorization. J Acoust Soc Am 140:3401–3401. https://doi.org/10.1121/1.4970917
- R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Rigoulot S, Wassiliwizky E, Pell MD (2013) Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition. Front Psychol 4:367. https://doi.org/10.3389/fpsyg.2013.00367
- Sauter DA, Eisner F, Calder AJ et al. (2010) Perceptual cues in nonverbal vocal expressions of emotion. Q J Exp Psychol 63:2251–2272. https://doi.org/ 10.1080/17470211003721642
- Sauter DA (2006) An investigation into vocal expressions of emotions: the roles of valence, culture, and acoustic factors. Unpublished Ph.D. thesis, University College London
- Sbattella L, Colombo L, Rinaldi C et al. (2014) Extracting emotions and communication styles from prosody. In: da Silva H, Holzinger A, Fairclough S, Majoe D (eds) Physiological computing systems, vol. 8908. Springer, Heidelberg, pp. 21–42
- Schaerlaeken S, Grandjean D (2018) Unfolding and dynamics of affect bursts decoding in humans. PLoS ONE 13:e0206215. https://doi.org/10.1371/ journal.pone.0206216
- Scherer KR, Baenziger T (2004) Emotional expression in prosody: a review and an agenda for future research. In: Bel B, Marlien I (eds) Speech prosody, Nara, Japan, pp. 359–366
- Scherer KR, Banse R, Wallbott H (2001) Emotion inferences from vocal expression correlate across languages and cultures. J Cross Cult Psychol 32:76–92. https://doi.org/10.1177/0022022101032001009
- Scherer KR (1994) Affect bursts. In: van Goozen SHM, van de Poll NE, Sergeant JA (eds) Emotions: essays on emotion theory. Erlbaum, Hillsdale, pp. 161–193
- Scherer KR, London H, Wolf J (1973) The voice of confidence: Paralinguistic cues and audience evaluation. J Res Pers 7:31-44. https://doi.org/10.1016/0092-6566(73)90030-5
- Schirmer A (2010) Mark my words: tone of voice changes affective word representations in memory. PLoS ONE 5(2):e9080. https://doi.org/10.1371/journal.pone.0009080

Schirmer A, Kotz SA (2003) ERP evidence for a sex-specific Stroop effect in emotional speech. J Cogn Neurosci 15:1135–1148. https://doi.org/10.1162/089892903322598102

Scott SK, Sauter D, McGettigan C (2010) Brain mechanisms for processing perceived emotional vocalizations in humans. In: Brudzynski SM (ed), Handbook of behavioral neuroscience, Elsevier, pp. 187–197

Seber GAF (1984) Multivariate observations. John Wiley & Sons

Thompson WF, Balkwill LL (2009) Cross-cultural similarities and differences. In: Juslin PN, Sloboda JA (eds) Handbook of music and emotion: theory, research, applications, 1st edn. Oxford University Press, New York, pp. 755–791. https://doi.org/10.1093/acprof:oso/9780199230143.003.0027

Toivanen J, Väyrynen E, Sepännen T (2004) Automatic discrimination of emotion from spoken Finnish. Lang Speech 47:383–412. https://doi.org/10.1177/00238309040470040301

Wagner HL (1993) On measuring performance in category judgement studies of nonverbal behaviour. J Nonverbal Behav 17:3–28. https://doi.org/10.1007/ BE00987006

Wendt B, Scheich H (2002) The "Magdeburger Prosodie Korpus"—a spoken language corpus for fMRI-Studies. In: Bel B, Marlien I (eds) Speech prosody. Aix-en-Provence, SproSIG, pp. 699–701

Wilson TD, Gilbert DT (2008) Explaining away: a model of affective adaptation.

Perspect Psychol Sci 3:370–386. https://doi.org/10.1111/j.1745-6924.2008.00085.x

World Medical Association (2013) World Medical Association Declaration of Helsinki: ethical principles form medical research involving human subjects. JAMA 310:2191–2194. https://doi.org/10.1001/jama.2013.281053

Acknowledgements

This research was conducted through a project funded by German Research Foundation (Deutsche Forschungsgemeinschaft, DFG)—Project number 254142454/GRK 2070. The funder had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript. We would like to thank Silke Paulmann for generously providing us with her stimuli sets, Julia Fischer for helpful suggestions

on the manuscript, Carlotta Dove, Isabel Noethen and Christina Broering for help with data acquisition and all individuals who participated in the research presented here.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1057/s41599-020-0499-z.

Correspondence and requests for materials should be addressed to A.L.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing,

adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020