Acoustic Features Distinguishing Emotions in Swedish Speech

*M. Ekberg, *G. Stavrinos, *J. Andin, †S. Stenfelt, and *Ö. Dahlström, *†Linköping, Sweden

Summary: Few studies have examined which acoustic features of speech can be used to distinguish between different emotions, and how combinations of acoustic parameters contribute to identification of emotions. The aim of the present study was to investigate which acoustic parameters in Swedish speech are most important for differentiation between, and identification of, the emotions anger, fear, happiness, sadness, and surprise in Swedish sentences. One-way ANOVAs were used to compare acoustic parameters between the emotions and both simple and multiple logistic regression models were used to examine the contribution of different acoustic parameters to differentiation between emotions. Results showed differences between emotions for several acoustic parameters in Swedish speech: surprise was the most distinct emotion, with significant differences compared to the other emotions across a range of acoustic parameters, while anger and happiness did not differ from each other on any parameter. The logistic regression models showed that fear was the best-predicted emotion while happiness was most difficult to predict. Frequency- and spectral-balance-related parameters were best at predicting fear. Amplitude- and temporal-related parameters were most important for surprise, while a combination of frequency-, amplitude- and spectral balance-related parameters are important for sadness. Assuming that there are similarities between acoustic models and how listeners infer emotions in speech, results suggest that individuals with hearing loss, who lack abilities of frequency detection, may compared to normal hearing individuals have difficulties in identifying fear in Swedish speech. Since happiness and fear relied primarily on amplitude- and spectral-balance-related parameters, detection of them are probably facilitated more by hearing aid use.

Key Words: Acoustic features−Emotions−Speech.

INTRODUCTION

Expression of emotions is a fundamental aspect of human communication, and the ability to recognize and interpret emotions in speech is crucial for successful social interactions. Mechanisms underlying emotional prosody are complex, involving acoustic features, higher-level perceptual integration, and cognitive processes guided by cultural norms. There is a general agreement that distinct acoustic features of different emotions and acoustic differences between emotions are related to listeners' ability to infer them.[1] While previous studies have identified distinct acoustic features of different emotions in speech,[2−5] they have mainly used descriptive methods. In this study, the aim is to address this gap by applying inferential statistical analyses to explore the acoustic features that characterize and distinguish between five different emotions in Swedish speech. Findings may have implications for understanding the effect of voice pathologies on emotion expression and vocal emotion recognition difficulties in individuals with hearing loss.

Emotions and emotional prosody

Emotions are difficult to define, and hence there is no consensus regarding their exact definition. There is, however, general agreement that emotions are constituted of different components, such as appraisal of stimuli, action preparation, physiological responses, expressions, and subjective feelings.[6] There is also substantial agreement that emotions are of great importance for motivating and guiding individual actions, as well as social interactions between individuals.[7] Deficits in the ability to accurately perceive emotions in others may have adverse effects on relationships and function in social as well as work environments.[8]

Emotions are expressed in speech through prosody and supra-segmental modulations of features such as pitch, intensity/loudness, duration, and the rhythm of speech.[9,10] The term *emotional prosody* is commonly used to refer to prosody expressing emotions in speech. Different emotions expressed through speech prosody have mutually distinct acoustic profiles (distinct combinations of acoustic features) which are related to the ability of listeners to accurately perceive the emotion expressed by speakers.[1]

Acoustic parameters of emotions in speech

Acoustic parameters are used to describe different acoustical variations in sounds, such as fundamental frequency (F0)/pitch, intensity/amplitude/loudness, and Hammarberg index. The acoustic parameters can be divided into four features; 1) frequency, including e.g., the fundamental frequency and the frequency of the formants, 2) amplitude, including e.g., loudness, shimmer, and Harmonics-to-Noise Ratio (HNR), 3) spectral balance, including parameters which characterize the relative energy in different frequency bands, such as the

Accepted for publication March 10, 2023.
Declarations of interest: none.
From the *Department of Behavioural Sciences and Learning, Linköping University, Linköping, Östergötland, Sweden; and the †Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Östergötland, Sweden.
Address correspondence and reprint requests to M. Ekberg, Department of Behavioural Sciences and Learning, Linköping University, Olaus Magnus vag 37, 583 30, Linköping, Östergötland, Sweden. E-mail: mattias.ekberg@liu.se
Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2023.03.010

Hammarberg Index, and 4) temporal, including parameters which are related to changes in time, such as length of voiced and unvoiced segments, amplitude, spectral-balance, and temporal.[11] Relatively small numbers of acoustic parameters (related to frequency, intensity, spectral balance, and temporal) can predict several emotion categories expressed through speech prosody.[12] Findings related to acoustic parameters are relatively consistent across emotions. Anger, happiness, and fear are described by relatively high mean amplitude, and relatively high pitch,[2–4,13,14] while sadness has been described as having a comparatively low pitch, and low amplitude.[2,13] Surprise has been described as having high pitch,[14,15] as well as pitch variation.[13] Other parameters such as the Harmonics-to-noise ratio (HNR),[2,3] spectral balance-related parameters,[3,16] and temporal (time-related) parameters,[2,3,17] show less consistent patterns. In the context of Swedish speech, specifically Nordström[3] has shown that correctly identified expressions of anger, happiness, fear, and sadness are associated with acoustic profiles in which multiple parameters in the frequency, amplitude, spectral balance, and temporal domains strongly diverge from a neutral voice, showing partly similar patterns across emotions, but also differences between emotions. For example, loudness is highest for anger and lowest for sadness. Another example is that HNR is higher for fear compared to anger, happiness, and sadness.[3]

While many studies have characterized different emotions descriptively, based on their acoustic parameters, there are relatively few studies focusing on statistical comparisons of acoustic parameters between emotions. In addition, there is a lack of studies comparing a variety of emotions using a more comprehensive set of acoustic parameters representing different features Although the human mind differs from acoustical statistical models, the acoustic parameters which distinguish between emotions statistically might also be used by listeners to infer emotions. Therefore, knowledge of such parameters could inform hypotheses regarding human performance. Knowledge of which acoustic parameters characterize different emotions may be of significance, for example in understanding vocal emotion recognition difficulties associated with hearing loss.

Purpose

The overall aim of the present study is to understand the relation between acoustic parameters and emotions expressed in speech. We will therefore extract acoustic parameters from speech and examine how acoustic parameters differ between emotions (anger, fear, happiness, sadness, and surprise) and how they can be used to predict emotions to answer the following research questions: 1) Which acoustic parameters differ between different emotions expressed in Swedish speech? 2) How well do the overarching acoustic features (frequency, amplitude, spectral-balance, temporal) explain emotions in Swedish speech? 3) How well do specific acoustic parameters explain emotions in Swedish speech?

METHOD

Material

The procedure for obtaining the audio recordings used in the present study included selection of sentences and speakers, recordings, and validation of the emotional prosody as perceived by listeners.

Fourteen sentences, emotionally neutral in terms of semantic content, were selected from the Swedish version of the Hearing in Noise test (HINT).[18] The sentences are listed in Appendix A. Four different actors read the sentences with the emotions of anger, happiness, sadness, fear, surprise, and interest (interest was not included in the following analyses, see below). In addition, the actors also expressed emotionally neutral versions of the sentences. The actors were a 69-year-old female, a 73-year-old male, a 19-year-old female, and a 29-year-old male.

The sentences were recorded with the aid of a sound technician in a sound-attenuated booth at Linköping University Hospital, Sweden. Recordings were made using AudacityTM version 3.2,[19] a Pearl CC30 microphone (ser. nr. 3573), a Behringer U-Phoria UMC202HD soundcard, and with a 24-bit resolution, and a 44.1 kHz sampling rate. The clearest version out of several recordings of each sentence and emotion for each of the four actors was selected by three of the authors (M.E., Ö.D, and J.A.) With few exceptions, the sentences are 2−3 seconds long.

To validate the sentences, the chosen recordings, one for each combination of speaker and emotion, were used in an online emotion recognition task. The experiment was constructed in PsychoPy version 3.0.3 (see [20] for a description of version 2.0) and administered online via Pavlovia (https://pavlovia.org/) which is based on PSYCHOJS version 2021.1.4. The recordings were divided into four separate lists each of which consisted of four blocks, one for each speaker. Each block consisted of the fourteen sentences spoken with emotional prosody. Before each block, the fourteen sentences spoken by the target speaker in neutral emotion were played. Thereafter, the task was to choose, between seven options, which emotion was perceived as being expressed in that sentence. The seven options were the six emotions and, to reduce bias and to avoid forcing participants to choose an answer they were not comfortable with, the option “none of the listed emotions”. Neutral was not included as an option in the task. The emotions were listed with numbers on the computer screen and responses were given by pressing the number on the keyboard corresponding to the number on the screen. Participants were recruited through social media and posters distributed at Linköping University. Participants between the ages of 18 and 70 performed the task online using their own computers. They were instructed to use headphones and to set the volume at a comfortable level. In total, the four lists were responded to by 79 unique respondents (17, 25, 20, and 17 for each of the lists, 66.5% female). Mean age of the responders was 43 years (SD = 15). Several individuals have most likely responded to more than one list.

Sentences that were classified as the intended emotion by more than 50% of listeners were classified as well-recognized and were subsequently included in the present study. In total 162 recordings (33 for anger, 33 for fear, 29 for happiness, 31 for sadness, and 31 for surprise) were classified as well-recognized and were consequently used for analyses (interest was excluded due to few (seven) well-recognized sentences). Only the sentences that were well-recognized are included in the following analyses.

Ethical approval for the study has been obtained from the Swedish Ethical Review Authority, 2020-03674.

Acoustic analyses

Acoustic parameters from the Geneva Minimalistic Parameter Set (GeMAPS) were extracted using openSmile version 2.3,[21] in Python 3.9,[22] for the recorded sentences.[11] GeMAPS has previously been used to extract acoustic parameters from running speech/sentences.[11] All acoustic parameters are described in Table 1.

Standardized z-scores were calculated for each acoustic parameter using the mean score and standard deviation of the neutral recordings for each speaker. Following Scherer,[1] acoustic parameters are referred to as high or low, compared to a neutral voice, when the z-scores are significantly different from 0 (based on 95% Confidence Intervals) and medium if they do not differ from that of a neutral voice.

Statistical analyses

To investigate differences in acoustic parameters between emotions, we performed one-way ANOVAs separately for each acoustic parameter using emotion as the independent variable (anger, happiness, fear, sadness, surprise) and the z-scores for each recording as the dependent variable. Significant results were followed up by post hoc-tests using the Bonferroni correction for multiple comparisons. The significance level was set to 5%.

Further, to investigate how much the individual parameters and the overarching acoustic features can contribute to the classification of the emotions, we performed logistic regression analyses with each emotion vs. the remaining emotions as the outcome. This was performed in three steps, first by simple regression models with each parameter as predictor (22 analyses per emotion), then by multiple regression models using the parameters within each acoustic feature as predictors (four analyses per emotion), and finally by multiple models based on all acoustic parameters across all features as predictors (one analysis per emotion). The purpose of using the multiple regression models was to assess how much different acoustic parameters could

TABLE 1.
Acoustic Parameters Analyzed in the Present Study Based on Definitions in Eyben et al.,[11] Divided by Acoustic Features (Frequency Related, Amplitude Related, Spectral-Balance Related, and Temporal Related)

Acoustic features (parameters)	Definition
Frequency related	
Pitch	Mean logarithmic fundamental frequency (F0) on a semitone scale starting at 27.5 Hz
Jitter	Mean deviations in individual consecutive F0 period lengths.
Frequency-formant 1	Mean of the center frequency of the first formant
Frequency-formant 2	Mean of the center frequency of the second formant
Frequency-formant 3	Mean of the center frequency of the third formant
Pitch percentile range	Range of the 20th to 80th percentile of the logarithmic fundamental frequency (F0) on a semitone scale starting at 27.5 Hz.
Formant 1 bandwidth	Mean bandwidth of the first formant
Amplitude related	
Shimmer	Mean difference of the peak amplitudes of consecutive F0 periods
Loudness	Estimate of the mean perceived signal intensity from an auditory spectrum
Harmonics-to-noise ratio (HNR)	Mean ratio of energy in harmonic components to energy in noise-like components.
Spectral-balance related	
Alpha ratio	Mean ratio of the summed energy from 50−1000 and 1−5 kHz
Hammarberg index	Mean ratio of the strongest energy peak in the 0−2 kHz region to the strongest energy peak in the 2−5 kHZ region
Spectral Slope V 0-500 Hz	Mean of linear regression slope of the logarithmic power spectrum within the 0-500 Hz spectral band for voiced regions
Spectral slope V 500-1500 Hz	Mean of linear regression slope of the logarithmic power spectrum within the 500-1500 Hz spectral band for voiced regions
Formant 1 relative energy	Mean of the relative energy of the first formant and the ratio of the energy of the spectral harmonic peak at the first formant's center frequency to the energy of the spectral peak at the fundamental frequency
Formant 2 Relative energy	Mean of the relative energy of the second formant and the ratio of the energy of the spectral harmonic peak at the second formant's center frequency to the energy of the spectral peak at the fundamental frequency
Formant 3 Relative energy	Mean of the relative energy of the third formant and the ratio of the energy of the spectral harmonic peak at the third formant's center frequency to the energy of the spectral peak at the fundamental frequency
Harmonic difference H1-H2	Mean ratio of energy of the first F0 harmonic to the energy of the second F0 harmonic
Harmonic difference H1-A3	Mean ratio of the energy of the first F0 harmonic to the highest harmonic in the third formant range
Temporal related	
Rate of loudness peaks	Mean number of loudness peaks per second
Length of continuously voiced regions	Mean length of continuously voiced regions
The length of unvoiced regions	Mean length of unvoiced regions
Pseudo syllable rate	Mean number of continuous voiced regions per second

contribute to the characterization of the different emotions, and therefore an iterative procedure, testing all possible multiple models to identify those with the highest explanatory power (highest Nagelkerke's R^2), was performed. ANOVAs were performed using IBM SPSS v.28,[23] while the binary logistic regressions were done using R 4.2.2,[24] with the foreign,[25] the combinat,[26] and the fmsb[27] packages.

RESULTS

Comparisons of acoustic parameters

An overview of the acoustic parameter differences patterns is presented in Figure 1, and a more comprehensive presentation of z-scores and significant differences is presented in Table 2.

For the *frequency-related parameters*, the one-way ANOVA showed significant differences between the emotions in all parameters except F1 Bandwidth (Table 2 and Figure 1). Happiness had the highest values of all emotions for pitch and the F1, F2, and F3 frequencies, differing significantly from surprise (pitch) and sadness (F1, F2, and F3 frequency). For the F2 frequency, fear had also significantly higher values than sadness. For jitter, surprise had significantly higher values compared to sadness, fear, and anger.

All *amplitude-related parameters* showed significant differences between emotions. Shimmer was significantly higher for surprise compared to all other emotions. Loudness, was significantly higher for anger, and happiness compared to sadness and surprise, and significantly higher for fear compared to surprise. The HNR parameter was significantly higher for fear compared to anger, sadness, and surprise, and happiness was significantly higher than surprise.

The *spectral balance-related parameters* showed significant differences between emotions in all parameters except F1 amplitude and H1A3. Alpha ratio, v500v1500, F2 amplitude, and F3 amplitude were significantly higher in anger and happiness compared to surprise. Hammarberg was significantly higher in surprise compared to anger. For v0v500, fear was significantly higher than all other emotions and for H1H2, happiness was significantly higher than sadness.

The *temporal-related parameters* showed significant differences between emotions in loudness peak as well as in pseudo-syllable rate. Surprise had significantly higher values compared to anger and happiness, for both loudness peak and pseudo-syllable rate, as well as significantly higher than sadness in loudness peak.

In sum, across all four acoustic features, surprise was the emotion that differed the most from the other emotions. Anger and happiness did not differ significantly from each other in any parameter.

Predictions of emotional prosody

The results from the multiple logistic regression analysis, which examined how well the acoustic features explain each

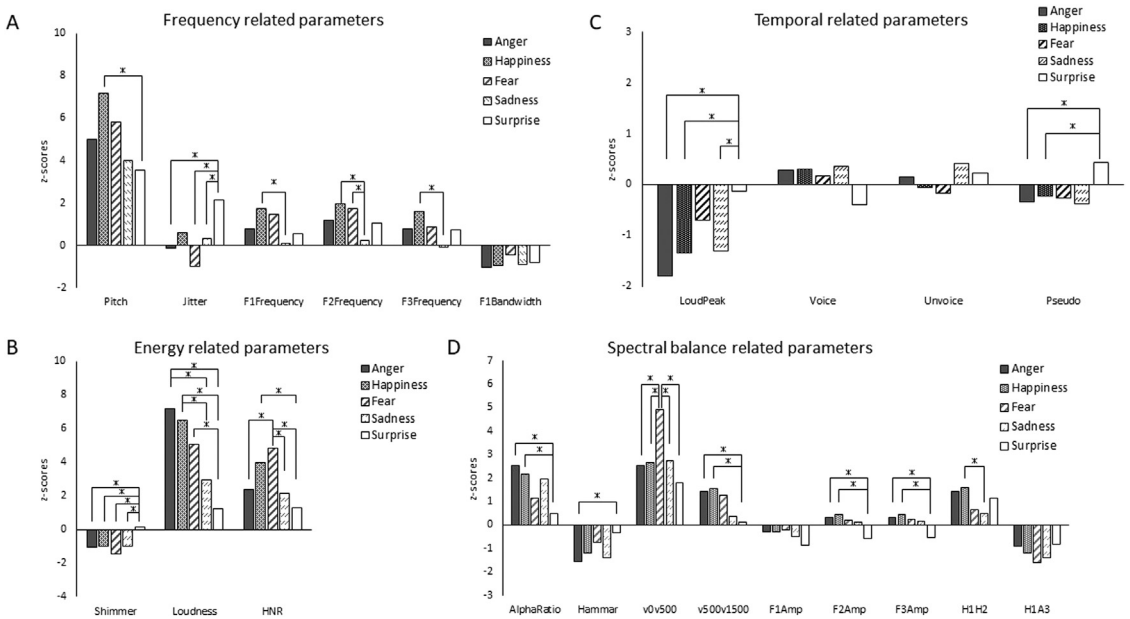

FIGURE 1. Differences in acoustic parameters between emotions. Parameters are divided into four categories; A) frequency-related parameters, B) amplitude-related parameters, C) spectral-balance-related parameters, and D) temporal-related parameters. F1Frequency = Frequency- formant 1, F2Frequency = Frequency-formant 2, F3Frequency = Frequency-formant 2, F1Bandwidth = Formant 1 bandwidth, HNR = Harmonics-to Noise ratio, AlphaRatio = Alpha ratio, Hammar = Hammarberg index, v0v500 = Spectral Slope V 0-500 Hz, v500v1500 = Spectral slope V 500-1500 Hz, F1Amp = Formant 1 relative energy, F2Amp = Formant 2 relative energy, F3Amp = Formant 3 relative energy, H1H2 = Harmonic difference H1-H2, H1A3 = Harmonic difference H1-A3, LoudPeak = Rate of loudness peaks, Voice = Length of continuously voiced regions, Unvoice = The length of unvoiced regions, Pseudo = Pseudo syllable rate. Significant differences, $P < .05$ are marked by *. Note that the range of the y-axes differs between figures.

TABLE 2.
Comparisons of Acoustic Parameters Between Emotions

Acoustic features (parameters)	Anger	Happiness	Fear	Sadness	Surprise				Comparisons
	M (*SD*)	M (*SD*)	M (*SD*)	M (*SD*)	M (*SD*)	*F*	*P*	pEta2	Post hoc-tests (Bonferroni adjusted)
Frequency-related:									
pitch	5.00 (5.39)	7.18 (6.25)	5.81 (2.31)	3.99 (5.36)	3.56 (4.14)	2.77	0.029	.068	Happiness > Surprise ($P = 0.039$)
jitter	-0.13 (0.38)	0.58 (0.38)	-0.98 (0.41)	0.32 (0.39)	2.14 (0.39)	8.24	<0.001	.178	Surprise > Anger ($P < 0.001$) Surprise > Fear ($P < 0.001$) Surprise > Sadness ($P = 0.014$)
F1Frequency	0.78 (0.34)	1.75 (0.34)	1.47 (0.37)	0.12 (0.35)	0.57 (0.35)	3.60	0.008	.086	Happiness > Sadness ($P = 0.012$)
F2Frequency	1.20 (0.35)	1.94 (0.35)	1.75 (0.37)	0.23 (0.36)	1.03 (0.36)	3.53	0.009	.085	Happiness > Sadness ($P = 0.008$) Fear > Sadness ($P = 0.038$)
F3Frequency	0.80 (0.34)	1.59 (0.34)	0.88 (0.37)	-0.10 (0.35)	0.72 (0.35)	2.95	0.022	.072	Happiness > Sadness ($P = 0.008$)
F1Bandwidth	-1.05 (1.29)	-0.96 (0.95)	-0.44 (0.94)	-0.88 (1.35)	-0.82 (0.88)	1.38	0.244		
Amplitude-related:									
shimmer	-1.03 (0.21)	-1.02 (0.21)	-1.43 (0.23)	-1.02 (0.22)	0.13 (0.22)	7.12	<0.001	.158	Surprise > Anger ($P = 0.002$), Surprise > Fear ($P < 0.001$) Surprise > Happiness ($P = 0.002$) Surprise > Sadness ($P = 0.003$)
loudness	7.16 (0.66)	6.49 (0.66)	5.09 (0.71)	2.96 (0.68)	1.24 (0.68)	13.36	<0.001	.260	Anger > Sadness ($P < 0.001$) Anger > Surprise ($P < 0.001$) Fear > Surprise ($P = 0.001$) Happiness > Sadness ($P = 0.003$) Happiness > Surprise ($P < 0.001$)
HNR	2.36 (0.52)	3.99 (0.52)	4.83 (0.55)	2.16 (0.54)	1.31 (0.54)	7.09	<0.001	.157	Fear > Anger ($P = 0.014$) Fear > Sadness ($P = 0.007$) Fear > Surprise ($P < 0.001$) Happiness > Surprise ($P = 0.004$)
alphaRatio	2.52 (0.40)	2.15 (0.40)	1.14 (0.43)	1.95 (0.41)	0.48 (0.41)	4.05	0.004	.096	Anger > Surprise ($P = 0.005$) Happiness > Surprise ($P = 0.043$)
Hammarberg	-1.57 (0.28)	-1.19 (0.28)	-0.74 (0.30)	-1.4 (0.29)	-0.35 (0.29)	2.97	0.022	.072	Surprise > Anger ($P = 0.032$)
slopeV0V500	2.53 (0.43)	2.68 (0.43)	4.90 (0.46)	2.76 (0.44)	1.81 (0.44)	6.54	<0.001	.147	Fear > Anger ($P = 0.002$) Fear > Happiness ($P = 0.006$) Fear > Sadness ($P = 0.010$) Fear > Surprise ($P < 0.001$)
slopev500V1500	1.45 (0.32)	1.57 (0.32)	1.28 (0.34)	0.35 (0.33)	0.12 (0.33)	4.21	0.003	.100	Anger > Surprise ($P = 0.042$) Happiness > Surprise ($P = 0.019$)
F1Amplitude	-0.3 (0.22)	-0.31 (0.22)	-0.19 (0.24)	-0.49 (0.23)	-0.85 (0.23)	1.30	0.274		
F2Amplitude	0.32 (0.20)	0.43 (0.20)	0.21 (0.21)	0.10 (0.20)	-0.56 (0.20)	3.68	0.007	.088	Anger > Surprise ($P = 0.024$) Happiness > Surprise ($P = 0.007$)
F3Amplitude	0.34 (0.20)	0.46 (0.20)	0.24 (0.21)	0.14 (0.21)	-0.52 (0.21)	3.54	0.009	.085	Surprise < Anger ($P = 0.030$) Happiness > Surprise ($P = 0.008$)
H1H2	1.44 (0.24)	1.61 (0.24)	0.66 (0.25)	0.48 (0.25)	1.13 (0.25)	4.00	0.004	.095	Happiness > Sadness ($P = 0.012$)
H1A3	-0.91 (0.29)	-1.19 (0.29)	-1.61 (0.30)	-1.40 (0.29)	-0.83 (0.29)	1.23	0.301		
Temporal-related:									
loudnesspeaksRate	-1.79 (0.27)	-1.35 (0.27)	-0.71 (0.28)	-1.30 (0.27)	-0.13 (0.27)	5.65	<0.001	.129	Surprise > Anger ($P < 0.001$) Surprise > Happiness ($P = 0.016$) Surprise > Sadness (P=.029)
voicedLength	0.28 (0.19)	0.31 (0.19)	0.17 (0.20)	0.35 (0.19)	-0.40 (0.19)	2.68	0.034	.066	
unvoicedLength	0.15 (0.27)	-0.05 (0.27)	-0.17 (0.28)	0.42 (0.28)	0.22 (0.28)	0.69	0.598		
pseudoyllableRate	-0.34 (0.19)	-0.22 (0.19)	-0.26 (0.20)	-0.38 (0.19)	0.44 (0.19)	3.00	0.020	.073	Surprise > Anger ($P = 0.046$) Surprise > Sadness ($P = 0.034$)

Note: F1Frequency = Frequency- formant 1, F2Frequency = Frequency-formant 2, F3Frequency = Frequency-formant 2, F1Bandwidth = Formant 1 bandwidth, HNR = Harmonics-to Noise ratio, AlphaRatio = Alpha ratio, Hammar = Hammarberg index, v0v500 = Spectral Slope V 0-500 Hz, v500v1500 = Spectral slope V 500-1500 Hz, F1Amp = Formant 1 relative energy, F2Amp = Formant 2 relative energy, F3Amp = Formant 3 relative energy, H1H2 = Harmonic difference H1-H2, H1A3 = Harmonic difference H1-A3, LoudPeak = Rate of loudness peaks, Voice = Length of continuously voiced regions, Unvoice = The length of unvoiced regions, Pseudo = Pseudo syllable rate.

emotion, are presented in Table 3. In Table B1, the univariate predictors (not analyzed in conjunction with other parameters) are also presented. For the models including frequency-related parameters, the model predicting fear showed the best performance (Nagelkerke's R^2 = .30) followed by surprise (.18) and sadness (.15). Fear was characterized by lower jitter (in opposite to surprise), by higher F2Frequency, by higher F1Bandwidth (in contrast to sadness) and by lower F3Frequency.

Amplitude-related acoustic features were primarily useful for predicting surprise (Nagelkerke's R^2 = .34). Surprise was characterized by higher shimmer (similar to happiness but opposite to sadness) and by lower loudness (similar to sadness but in contrast to happiness and anger).

Spectral balance-related acoustic features were useful for predicting fear (Nagelkerke's R^2 = .32), surprise (.22), and sadness (.20). Fear was primarily characterized by lower alphaRatio (like surprise but contrary to anger), by lower

TABLE 3.
Multiple Logistic Regression Models by Acoustic Features With Emotions as Outcome and all Other Emotions as Reference, Presented by Odds Ratios (95% Confidence Intervals). In the First Step, All Possible Models Using Parameters Within Respective Acoustic Feature Were Examined. In the Final Step, New Models Were Produced in the Same Way But With All Variables as Independent Variables (Only Presented by Nagelkerke's R^2)

Acoustic features (parameters)	Anger	Happiness	Fear	Sadness	Surprise
Frequency-related:					
pitch		1.10 (1.02-1.18)*[b]			[e]
jitter	[a]		0.69 (0.52-0.91)* [c]		1.41 (1.19-1.68)*[e]
F1Frequency	[a]			* [d]	
F2Frequency	[a]		2.53 (1.38-4.64)[c]	0.60 (0.45-0.80)*[d]	[e]
F3Frequency			0.54 (0.31-0.95)	*	
F1Bandwidth	[a]		2.23 (1.27-3.91)*	0.59 (0.37-0.96)[d]	
Nagelkerke R^2		**0.06**	**0.30**	**0.15**	**0.18**
Amplitude-related:					
shimmer		1.61 (1.02-2.56)[b]	*	0.46 (0.26-0.81)[d]	1.78 (1.20-2.62)*
Loudness	1.23 (1.11-1.36)*[a]	1.16 (1.04-1.29)*[b]		0.87 (0.78-0.97)*[d]	0.77 (0.66-0.89)*
HNR	0.84 (0.71-0.99)[a]	1.25 (1.05-1.49)*	1.23 (1.09-1.39)*[c]	0.77 (0.62-0.94)[d]	*
Nagelkerke R^2	**0.18**	**0.13**	**0.12**	**0.16**	**0.34**
Spectral-balance-related:					
alphaRatio	1.61 (1.23-2.10)*[a]		0.26 (0.11-0.61)	[d]	0.69 (0.55-0.87)*
Hammarberg	*	[b]	0.27 (0.11-0.70)		* [e]
slopeV0V500		[b]	1.69 (1.34-2.13)*[c]		*
slopeV500V1500	[a]	1.23 (1.00-1.52)	1.44 (1.03-2.02)	0.71 (0.55-0.91)*[d]	[e]
F1Amplitude		*		0.26 (0.10-0.65)[d]	* [e]
F2Amplitude		*		4.47 (1.62-12.36)	0.53 (0.37-0.77)* [e]
F3Amplitude		*	[c]	[d]	*
H1H2	[a]	1.38 (1.05-1.81)*	[c]	0.48 (0.32-0.71)*[d]	[e]
H1A3	1.83 (1.25-2.69)				[e]
Nagelkerke R^2	**0.16**	**0.09**	**0.32**	**0.20**	**0.22**
Temporal-related:					
LoudnessPeaks	0.65 (0.48-0.88)*[a]		[c]		1.61 (1.19-2.17)*
voicedLength					0.40 (0.21-0.76)*
unvoicedLength				[d]	
pseudosyllableRate					* [e]
Nagelkerke R^2	**0.09**				**0.23**
Nagelkerke R^2, final models	**0.53**	**0.25**	**0.74**	**0.61**	**0.57**

* Indicates variables that were univariately significantly associated with the outcomes (also indicated by shaded cells), see Appendix B for these models. Variables included in final models based on all 22 parameters for [a] Anger, [b] Happiness, [c] Fear, [d] Sadness, [e] Surprise, see Appendix B for these models.

Hammarberg, by higher slopeV0V500 (only significant for fear), and by higher slopeV500V1500 (contrary to sadness but similar to happiness). The relative energy of the formants was only useful for predicting sadness (lower F1Amplitude and higher F2Amplitude) and surprise (lower F2Amplitude).

Temporal acoustic features were primarily useful for predicting surprise (Nagelkerke's R^2 = .23), and to some extent anger (.09). Fear was characterized by higher loudnessPeaks (contrary to anger) and by lower voicedLength.

Overall, when using all acoustic features, happiness was the most difficult emotion to predict (Nagelkerke's R^2 = .25) compared to the other emotions (all R^2:s $\geq$ .53). The predictive ability of separate parameters did not always correspond with parameters included in multiple models (feature by feature, or over all parameters), since some univariate significant variables were not included in multiple models, and vice versa, indicating a complex relationship between independents and emotions.

DISCUSSION

In the present study, we investigated emotional prosody to understand which acoustic parameters distinguish between emotions using a set of parameters from several different acoustic domains. Examination of the acoustic features and their ability to correctly predict emotions differed across emotions, while examination of the predictive ability of the different parameters showed a complex association between acoustic parameters and emotions.

We found differences primarily between surprise and other emotions. Surprisingly, anger and happiness did not differ on any parameter, which most probably will disagree with most listeners' experiences of separating between sentences expressing anger or happiness. Yildirim et al.[28] found little acoustic difference between anger and happiness in speech noting that the two emotions have poor separability. Preti et al.[29] found that anger was characterized by smaller pitch variability and lower speech rate compared to happiness. They argue that the acoustic profile of anger in their

study mostly correspond to a variant of anger labelled "cold anger".[29] This suggests that there may be variants of anger which are more or less acoustically similar to happiness.

When analyzing which acoustic features that predict emotion detection, we found that frequency-related parameters contributed the most to distinguish fear from other emotions (Nagelkerke's $R^2 = 0.30$). Further, sadness was also distinguished from other emotions by frequency-related parameters. Contrary, anger was not predicted by frequency parameters at all, except in conjunction with parameters of other features. This may indicate that frequency features are less important for characterizing anger acoustically which is in line with findings from Polzehl et al.[30] Consistent with previous descriptions,[2–4,13,14] we find that happiness, anger, and fear are characterized by a high pitch compared to a neutral voice. Additionally, the finding that surprise is characterized by high jitter is consistent with a previous description of surprise as having high pitch variability.[13]

Different combinations of amplitude parameters contributed to distinguish all emotions from one another, but most strongly to distinguish surprise from the rest (Nagelkerke's $R^2 = 0.34$). Importantly, HNR contributed to distinguish between several emotions. Sadness was characterized by low loudness compared to happiness, anger, and fear which is consistent with previous descriptions of sadness as having comparatively low amplitude.[1,2] However loudness was also low for surprise compared to happiness, anger, and fear which is inconsistent with some previous descriptions of surprise in speech.[13]

For the spectral balance feature, measures of relative energy in low- to mid-range frequencies contributed to the differentiation of several emotions, and in some cases, such as for sadness and surprise, relative energy of formant frequencies also contributed. This further supports that formant characteristics are relevant for distinguishing emotions in speech acoustically. In the present study, corroborating Nordström[3], we found that anger, fear, happiness, and sadness are characterized by a low Hammarberg index compared to a neutral voice. Contrary to Guzman et al.[16] we did not find evidence for sadness being characterized by a higher ratio of low frequency energy compared to anger and fear.

The temporal feature only contributed to distinguish surprise, and to some degree anger, from the other emotions. Our finding that surprise was characterized by a high rate of loudness peaks and high pseudo-syllable rate is consistent with surprise being characterized by a fast tempo in Scherer and Oschinsky.[15]

In sum, fear and sadness were most strongly predicted by a combination of different acoustic parameters, while happiness was the least well-predicted emotion. However, the relationship between acoustic features and recognition of emotions warrants further examination through experimental studies systematically manipulating the acoustic parameters.[31]

All of the emotions of anger, happiness, fear, and sadness differ from a neutral voice in multiple frequency-related, amplitude related, and spectral balance-related parameters. The findings regarding how anger, happiness, fear, and sadness differ from a neutral voice show a similar pattern to the acoustic parameters reported for the frequency, amplitude, and spectral balance features in Nordström.[3] However, mutual differences between individual emotions vary across studies,[1,3] which may depend on for example differences in inter-speaker characteristics.

The present results have some implications regarding recognition of emotions in speech. Sensorineural hearing loss will probably affect performance of emotion recognition in general, and in relation to all acoustic features, but the performance will in individuals vary depending on the specific type of hearing loss characteristics. The use of hearing aids (relying on linear amplification and compression) cannot fully restore frequency selectivity and pitch perception,[32] meaning that such performance depending on frequency-related parameters will not be ameliorated to the same extent as recognition of emotions relying more on amplitude- or spectral-balance-related parameters. Neither is temporal-related acoustic parameters expected to be ameliorated by linear amplification to the same extent. Based on the results in this study, it can therefore be hypothesized that emotion recognition of anger and happiness, both relying most on amplitude- and spectral-balance related features, will benefit more from linear amplification compared with emotion recognition of fear (who also relies heavily on frequency-related features). In general, it can still be expected that fear will be the easiest emotion to recognize (largest Nagelkerke's R^2 in results). Additionally, pathological changes to the voice affecting the pitch, and aspects and extensions thereof such as jitter and formant frequencies, may make the expression of different emotions less clear. Changes affecting the strength of the voice and harmonicity of sound (HNR) may also negatively impact the clarity of emotional expressions.

One limitation of the study is that there may be variations in how speakers produce emotional prosody that is not captured by the present study. Compared to previous studies, there were variations among speakers regarding age and sex, which should capture some of such potential variability. Future studies may add analyses of inter-individual variation to extend the external validity of our findings. Another limitation is that the set of sentences is not completely balanced over speakers and emotions, but the decision to use stimuli with clear emotional prosody was considered more important in terms of internal validity. A third limitation is that broader emotion categories such as anger, fear, and sadness can be divided into different subcategories which differ acoustically. This potential source of within-emotion variation may be one reason why happiness did not result in a model performing at the same level as models for the other emotions, see for example Scherer,[1] who treated this by including several sub-categories of happiness in the study design. It may be the case that the four speakers used this freedom to produce happiness in different ways by shifting between sub-categories when producing the sentences. Finally, it should be emphasized that this study focused

solely on acoustic parameters, eliminating other factors such as semantic context and facial expressions, that may influence emotion recognition in a natural setting.

Regarding methods, there is no consensus about which set of acoustic parameters are optimal for classifying emotions in speech.[33] GeMAPS[11] is developed to analyze affect and emotions in speech. It has been validated for analyses of sustained phonations as well as whole sentences.[11] Previous research has demonstrated the relevance of the GeMAPS parameters, which gives a proposed standard for parameter extraction for easier comparisons across studies.[11] Furthermore, GeMAPS was previously used in the most extensive acoustic analysis of emotions in Swedish speech[3], including parameters that are contested whether they can be reliably extracted from running speech such as HNR.

Finally, some of the inconsistencies between the present results and the findings in other studies may be related to differences between single-predictor and multi-predictor analyses. In the present study, this is shown by differences between results from ANOVA simple logistic regression models, and results of multiple logistic regression models.

Conclusions

Findings provide insights into acoustic properties of emotional speech and highlight a complex relationship between acoustic parameters and emotions. There were significant differences between the emotions of anger, happiness, fear, sadness, and surprise for several acoustic parameters extracted from speech. Surprise differed the most from the other emotions while anger and happiness did not differ from each other any parameter. The overall model used to predict how fear is distinguished from all other emotions showed the best performance while the model predicting happiness showed the lowest performance. For the models applied to the acoustic features, frequency- and spectral balance-related parameters performed best when predicting fear, while amplitude- and temporal-related parameters performed best when predicting surprise. Assuming that there are similarities between statistical acoustic models and listener inference of different emotions in speech, it may be hypothesized that linear amplification to restore audibility for individuals with sensorineural hearing loss may be effective for identification of anger and happiness. In sum, acoustic parameters and their ability to correctly predict emotions differed across emotions, and the predictive ability of the different parameters showed a complex association between acoustic parameters and emotions.

APPENDIX A SENTENCES USED IN THIS STUDY. ENGLISH TRANSLATIONS ARE PRESENTED WITHIN BRACKETS

1. Anden simmar i dammen (The duck swims in the pond)
2. Bollen studsar ut på vägen (The ball bounces out onto the road)
3. Ungdomarna köper varsin glass (The young people each buy an ice cream)
4. Flickan har kort rött hår (The girl has short red hair)
5. Farfar lagar mat åt barnen (Grandpa cooks for the children)
6. Äggen ska kokas sju minuter (The eggs are to be boiled for seven minutes)
7. Flickan handlade ost och korv. (The girl bought cheese and sausage)
8. Morfar provade för stora skor. (Grandpa tries on too large shoes)
9. Två svarta skjortor hängde på tork (Two black shirts hung to dry)
10. Jackan hängde i garderoben. (The jacket hung in the closet)
11. Båda tröjorna var svarta. (Both of the shirts were black)
12. Kossan betar grönt gräs i hagen (The cow grazes green grass in the pasture)
13. Hundarna rullade runt i snön. (The dogs rolled around in the snow)
14. Katten ska få ungar (The cat will have kittens)

APPENDIX B

TABLE B1.

TABLE B1.
Simple Logistic Regression Models by Acoustic Features with Emotions as Outcome and all Other Emotions as Reference, Presented by Odds Ratios (95% Confidence intervals).

Acoustic features (parameters)	Anger	Happiness	Fear	Sadness	Surprise
Frequency related:					
pitch	0.99 (0.92-1.07)	**1.10 (1.02-1.18)**	1.03 (0.96-1.11)	0.94 (0.85-1.03)	0.91 (0.82-1.00)
jitter	0.87 (0.72-1.05)	1.04 (0.89-1.22)	**0.61 (0.45-0.81)**	0.98 (0.83-1.16)	**1.41 (1.19-1.68)**
F1Frequency	0.95 (0.78-1.16)	**1.26 (1.05-1.51)**	1.15 (0.96-1.39)	**0.72 (0.55-0.93)**	0.88 (0.71-1.10)
F2Frequency	0.99 (0.82-1.19)	**1.22 (1.02-1.47)**	1.15 (0.96-1.39)	**0.69 (0.54-0.88)**	0.94 (0.77-1.15)
F3Frequency	1.00 (0.83-1.21)	**1.27 (1.05-1.54)**	1.03 (0.84-1.25)	**0.72 (0.57-0.92)**	0.98 (0.80-1.19)
F1Bandwidth	0.81 (0.57-1.14)	0.88 (0.63-1.24)	**1.55 (1.04-2.32)**	0.96 (0.67-1.36)	1.02 (0.71-1.46)
Amplitude related:					
shimmer	0.88 (0.65-1.21)	0.89 (0.65-1.22)	**0.60 (0.40-0.89)**	0.89 (0.65-1.23)	**2.15 (1.48-3.11)**
Loudness	**1.19 (1.08-1.31)**	**1.13 (1.03-1.24)**	1.03 (0.94-1.13)	**0.89 (0.80-0.98)**	**0.74 (0.64-0.84)**
HNR	0.93 (0.82-1.06)	**1.13 (1.01-1.27)**	**1.23 (1.09-1.39)**	0.90 (0.78-1.04)	**0.76 (0.63-0.91)**
Spectral-balance related:					
alphaRatio	**1.20 (1.02-1.41)**	1.11 (0.95-1.30)	0.89 (0.74-1.06)	1.06 (0.90-1.25)	**0.73 (0.60-0.90)**
Hammarberg	0.79 (0.63-1.00)	0.94 (0.75-1.18)	1.16 (0.90-1.49)	0.86 (0.68-1.09)	**1.42 (1.09-1.85)**
slopeV0V500	0.93 (0.80-1.09)	0.96 (0.82-1.12)	**1.42 (1.19-1.70)**	0.97 (0.84-1.14)	**0.78 (0.65-0.94)**
slopeV500V1500	1.19 (0.97-1.45)	**1.23 (1.01-1.51)**	1.11 (0.90-1.37)	0.80 (0.64-1.00)	**0.72 (0.57-0.92)**
F1Amplitude	1.11 (0.82-1.52)	1.10 (0.81-1.50)	1.21 (0.87-1.68)	0.95 (0.70-1.30)	**0.73 (0.53-0.99)**
F2Amplitude	1.24 (0.87-1.76)	1.40 (0.97-2.02)	1.10 (0.77-1.58)	0.99 (0.71-1.39)	**0.56 (0.39-0.79)**
F3Amplitude	1.22 (0.86-1.74)	1.40 (0.96-2.03)	1.09 (0.77-1.56)	1.00 (0.72-1.40)	**0.57 (0.40-0.80)**
H1H2	1.25 (0.96-1.64)	**1.39 (1.06-1.83)**	0.76 (0.56-1.03)	**0.66 (0.48-0.90)**	1.03 (0.78-1.36)
H1A3	1.14 (0.90-1.45)	1.00 (0.79-1.26)	0.82 (0.64-1.05)	0.90 (0.71-1.15)	1.18 (0.92-1.52)
Temporal related:					
LoudnessPeaks	**0.65 (0.48-0.88)**	0.86 (0.67-1.12)	1.18 (0.93-1.50)	0.89 (0.69-1.15)	**1.60 (1.21-2.11)**
voicedLength	1.15 (0.82-1.61)	1.18 (0.85-1.65)	1.02 (0.71-1.48)	1.22 (0.87-1.71)	**0.39 (0.21-0.72)**
unvoicedLength	1.02 (0.79-1.31)	0.91 (0.69-1.19)	0.84 (0.62-1.14)	1.16 (0.91-1.47)	1.06 (0.82-1.36)
pseudosyllableRate	0.82 (0.57-1.18)	0.93 (0.66-1.32)	0.90 (0.62-1.30)	0.79 (0.54-1.14)	**1.84 (1.26-2.68)**

Note: Significant OR are presented in bold.

REFERENCES

1. Scherer KR. Acoustic patterning of emotion vocalizations. In: Frühholz S, Belin P, eds. *The Oxford Handbook of Voice Perception*. Oxford University Press; 2018:60–92. https://doi.org/10.1093/oxfordhb/9780198743187.013.4.
2. Laukka P, Thingujam NS, Iraki FK, et al. The expression and recognition of emotions in the voice across five nations: a lens model analysis based on acoustic features. *J Pers Soc Psychol*. 2016;111:686–705. https://doi.org/10.1037/pspi0000066.
3. Nordström H. Emotional Communication in the human voice. 2019. https://www.diva-portal.org/smash/get/diva2:1304804/FULLTEXT01.pdf. Last accessed 26 Feb. 2023.
4. Özseven T. Investigation of the relation between emotional state and acoustic parameters in the context of language. *Eur J Sci Technol*. 2018;14:241–244. https://doi.org/10.31590/ejosat.448095.
5. Liu P, Pell MD. Recognizing vocal emotions in Mandarin Chinese: a validated database of Chinese vocal emotional stimuli. *Behav Res Methods*. 2012;44:1042–1051. https://doi.org/10.3758/s13428-012-0203-3.
6. Scherer KR, Moors A. The emotion process: event appraisal and component differentiation. *Annu Rev Psychol*. 2019;70:719–745. https://doi.org/10.1146/annurev-psych-122216-011854.
7. Izard CE. The many meanings/aspects of emotion: definitions, functions, activation, and regulation. *Emot Rev*. 2010;2:363–370. https://doi.org/10.1177/1754073910374661.
8. Picou EM, Singh G, Goy H, et al. Hearing, emotion, amplification, research, and training workshop: current understanding of hearing loss and emotion perception and priorities for future research. *Trends Hear*. 2018;22:1–24. https://doi.org/10.1177/2331216518803215.
9. Liebenthal E, Silbersweig DA, Stern E. The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. *Front Neurosci*. 2016;10:1–13. https://doi.org/10.3389/fnins.2016.00506.
10. Meyer M, Keller M, Giroud N. Suprasegmental speech prosody and the human brain. In: Frühholz S, Belin P, eds. *The Oxford Handbook of Voice Perception*. Oxford University Press; 2018:142–166. https://doi.org/10.1093/oxfordhb/9780198743187.013.7.
11. Eyben F, Scherer KR, Schuller BW, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans Affect Comput*. 2016. https://doi.org/10.1109/TAFFC.2015.2457417.
12. Juslin PN, Laukka P, Harmat L, et al. Spontaneous vocal expressions from everyday life convey discrete emotions to listeners. *Emotion*. 2021;21:1281–1301. https://doi.org/10.1037/emo0000762.
13. Abelin Å, Allwood J Cross linguistic interpretation of emotional prosody. Int Tutor Res Work Speech Emot. 2000;110−113. http://www.isca-speech.org/archive_open/speech_emotion/spem_110.html%5Cnpapers3://publication/uuid/52D5FD30-CC3A-4D9F-BB29-526984098F23. Last accessed 26 Feb. 2023.
14. Kamiloğlu RG, Fischer AH, Sauter DA. Good vibrations: a review of vocal expressions of positive emotions. *Psychon Bull Rev*. 2020;27:237–265. https://doi.org/10.3758/s13423-019-01701-x.
15. Scherer KR, Oshinsky JS. Cue utilization in emotion attribution from auditory stimuli. *Motiv Emot*. 1977;1:331–346. https://doi.org/10.1007/BF00992539.

16. Guzman M, Correa S, Muñoz D, et al. Influence on spectral energy distribution of emotional expression. *J Voice*. 2013;27:129.e1–129.e10. https://doi.org/10.1016/j.jvoice.2012.08.008.
17. Pell MD. Influence of emotion and focus location on prosody in matched statements and questions. *J Acoust Soc Am*. 2001;109:1668–1680. https://doi.org/10.1121/1.1352088.
18. Hällgren M, Larsby B, Arlinger S. A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *Int J Audiol*. 2006;45:227–237. https://doi.org/10.1080/14992020500429583.
19. Audacity Team. Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.2. 2022.
20. Peirce J, Gray JR, Simpson S, et al. PsychoPy2: experiments in behavior made easy. *Behav Res Methods*. 2019;51:195–203. https://doi.org/10.3758/s13428-018-01193-y.
21. Eyben F, Wöllmer M, Schuller B. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. MM'10 - *Proc ACM Multimed 2010 Int Conf*. 20101459–1462. https://doi.org/10.1145/1873951.1874246.
22. Python Software Foundation. Python Language Reference. Version 3.9. 2020. https://www.python.org/. Last accessed 18 Dec. 2022.
23. IBM Corp. IBM SPSS Statistics for Windows, Version 28.0. 2021.
24. R Core Team. R: A language and environment for statistical computing. 2021. https://www.r-project.org/. Last accessed 18 Dec. 2022.
25. Bivand R, Carey VJ, DebRoy S, et al. foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... R package version 0.8-65. 2022. http://cran.r-project.org/package=foreign. Last acccessed 18 Dec. 2022.
26. Chasalow S. Combinat: combinatorics utilities. Version 0.0-8. 2012. https://cran.r-project.org/web/packages/combinat. Last accessed 18 Dec. 2022.
27. Nakazawa M. fmsb: Functions for Medical Statistics Book with some Demographic Data. Version 0.7.4. 2022. https://cran.r-project.org/web/packages/fmsb. Last accessed 18 Dec. 2022.
28. Yildirim S, Bulut M, Lee CM, et al. An acoustic study of emotions expressed in speech. *Interspeech 2004*. ISCA: ISCA; 2004:2193–2196. https://doi.org/10.21437/Interspeech.2004-242.
29. Preti E, Suttora C, Richetin J. Can you hear what I feel? A validated prosodic set of angry, happy, and neutral Italian pseudowords. *Behav Res Methods*. 2016;48:259–271. https://doi.org/10.3758/s13428-015-0570-7.
30. Polzehl T, Schmitt A, Metze F, et al. Anger recognition in speech using acoustic and linguistic cues. *Speech Commun*. 2011;53:1198–1209. https://doi.org/10.1016/j.specom.2011.05.002.
31. Arias P, Rachman L, Liuni M, et al. Beyond correlation: acoustic transformation methods for the experimental study of emotional voice and speech. *Emot Rev*. 2021;13:12–24. https://doi.org/10.1177/1754073920934544.
32. Oxenham AJ. How we hear: the perception and neural coding of sound. *Annu Rev Psychol*. 2018;69:27–50. https://doi.org/10.1146/annurev-psych-122216-011635.
33. Doğdu C, Kessler T, Schneider D, et al. A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech. *Sensors*. 2022;22:7561. https://doi.org/10.3390/s22197561.