Comparing the acoustic expression of emotion in the speakingand the singing voice

Scherer, Klaus R.; Sundberg, Nancy Johanne Ruth; Tamarit, Lucas; Salomão, Gláucia L.

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 29 (2015) 218–235

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Comparing the acoustic expression of emotion in the speaking and the singing voice☆

Klaus R. Scherer[a,*], Johan Sundberg[b,2], Lucas Tamarit[a,1], Gláucia L. Salomão[b,2]

[a] *Swiss Center of Affective Sciences, University of Geneva, 7, rue des Battoirs, CH-1205 Geneva, Switzerland*
[b] *Department of Speech, Music and Hearing, School of Computer Science and Communication, Royal Institute of Technology, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden*

Received 26 April 2013; received in revised form 5 October 2013; accepted 14 October 2013
Available online 24 October 2013

Abstract

We examine the similarities and differences in the expression of emotion in the singing and the speaking voice. Three internationally renowned opera singers produced "vocalises" (using a schwa vowel) and short nonsense phrases in different interpretations for 10 emotions. Acoustic analyses of emotional expression in the singing samples show significant differences between the emotions. In addition to the obvious effects of loudness and tempo, spectral balance and perturbation make significant contributions (high effect sizes) to this differentiation. A comparison of the emotion-specific patterns produced by the singers in this study with published data for professional actors portraying different emotions in speech generally show a very high degree of similarity. However, singers tend to rely more than actors on the use of voice perturbation, specifically vibrato, in particular in the case of high arousal emotions. It is suggested that this may be due to by the restrictions and constraints imposed by the musical structure.

Keywords: Vocal expression; Emotional interpretation in singing; Comparison between emotion expression in speech and singing; Acoustic analyses of emotion

1. Introduction

Affect bursts (short, spontaneous, nonverbal expressions of emotion in voice, face, and body) can be considered as "living fossils" of early human affect expression that may have served as precursors of parallel evolution of speech, song, music and dance. Affect expression is likely to have played a special role because (1) innate mechanisms for spontaneous expression in face, voice, and body may have been the earliest communicative mechanisms in place, rapidly followed for control structures for learned behaviors, (2) both affect bursts and controlled vocalizations are widely shared across many species, and (3) the production mechanisms, at least for spontaneous expressions, are located in the subcortical regions of the mammalian brain (see Scherer, 2013a,b). This assumption is compatible with

☆ This paper has been recommended for acceptance by 'Dr. S. Narayanan'.

* Corresponding author. Tel.: +41 22 379 9211; fax: +41 22 379 9844.
E-mail addresses: klaus.scherer@unige.ch (K.R. Scherer), jsu@csc.kth.se (J. Sundberg), Lucas.Tamarit@unige.ch (L. Tamarit), gsalomao@kth.se (G.L. Salomão).
[1] Tel.: +41 22 379 9801; fax: +41 22 379 9844.
[2] Tel.: +46 8 790 7561; fax: +46 8 790 7854.

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.10.002

other suggestions concerning the origin of speech and music (such as the role of gestures or phonetic symbolism) and does not address the issue of their respective evolutionary priority (in fact, the common production substrate does not disambiguate between different evolutionary scenarios). The parallelism of human emotion expression in speech and music has been demonstrated by a comprehensive review of empirical studies on patterns of acoustic parameters in these two forms of human affect communication (Juslin and Laukka, 2003). The assumption of powerful "affect primitives" in speech and language is also supported by research on the recognition of emotion in speech (Bryant and Barrett, 2008; Laukka et al., 2013b; Pell et al., 2009; Sauter et al., 2010; Scherer et al., 2001) and music (Laukka et al., 2013a). This research has generally shown the existence of both a fairly high degree of universality of the underlying expression and recognition mechanisms and of sizeable differences between cultures, especially for self-reflective, social, and moral emotions.

The emotional power of the speaking voice, taken for granted ever since the ancient works of rhetoric (Cicero, Quintilian), has been frequently studied in empirical research (Scherer, 1986, 2003). The emotional power of the singing voice, although frequently acknowledged (Scherer, 1995; Sundberg, 1989), has only rarely been studied in an experimental fashion. Yet, in performing vocal music in Western music traditions (liturgical works, opera, and different kinds of song), professional singers have to be able to produce an extraordinary range of emotional meanings. A wide variety of means help to achieve such interpretation, such as gestures and facial expression, but the central instrument to evoke the subtle shadings of emotion is the human voice. In fact, singers are often judged in terms of their ability to produce the emotional modulation of their vocal performance that is considered appropriate to nature of the emotion to be portrayed (as implied by the text of a song or the libretto of an opera). In addition, like actors in spoken theater, they are expected to express the required emotions in a credible, authentic fashion, giving the impression of "inhabiting" the emotional feelings of the character they are performing (Scherer, 2013a).

Previous studies that have been devoted to the understanding of the emotional power of the singing voice tried to identify the acoustic cues used by listeners to recognize the emotional meaning in singing voice. A standard method for this purpose has been to correlate acoustic profiles of different emotions portrayed by professional singers with the listeners' judgments of emotions perceived (Kotlyar and Morozov, 1976) as well as with the level of expressiveness (Sundberg et al., 1995) or strength of the perceived emotion (Jansens et al., 1997). Results have agreed in that performances characterized by higher arousal levels (as joy and anger) tend to show higher average sound pressure levels and fast tempi than performances characterized by lower arousal levels (as sadness). Also a clear association between anger and the presence of vibrato, and sadness and the absence of vibrato has been found (Jansens et al., 1997). Another method used to investigate how the emotional meaning is conveyed in singing voice has been to compare recordings of a same song performed by various singers and analyze listeners' judgments of emotions perceived for each particular performer. Siegwart and Scherer (1995) and Howes et al. (2004) found that listeners' preferences and emotion judgments were indeed associated with specific acoustics characteristics. Correlations between different acoustic parameters and listeners' perception of emotions in singing voice (as well as in music in general) can also be studied by investigating the listeners' emotion judgments of sounds that had each of different parameters systematically and independently manipulated (Scherer and Oshinsky, 1977; Kotlyar and Morozov, 1976). Procedures of synthesis and resynthesizes have also been used to systematically manipulate acoustics parameters, in order to investigate the effects and relevance of each of the parameters for listeners emotion judgment (e.g., Goto et al., 2012; Fonseca, 2011; Kenmochi and Ohshita, 2007; Risset, 1991; Sundberg, 1978). A comparison between acoustic patterns that characterizes both expressive speech and expressive singing suggests a striking parallel between the expression of emotions in the speaking and the singing voice between. For instance, both in speech and singing anger is associated with high F0 variability (assuming that F0 variability in speech is translated into vibrato extent in singing) and with high vocal intensity, while sadness is associated with slow speech rate (and low tempo) as well as with low vocal intensity.

In consequence, we expect that emotion expression is similar in speaking and singing voice – because of the evolutionary origin of the expression mechanisms and the need for authenticity (Maynard Smith and Harper, 2003; Mortillaro et al., 2013). Obviously, there may well be important differences across languages and cultures due in large part to language characteristics such as phonemic structure or intonation rules. Yet, given the stability of findings across music and speech (Juslin and Laukka, 2003), one can expect similarities across studies in different languages and cultures between the expression of emotion in speech and singing.

Unfortunately, this issue not well researched. Compared to the study of facial expression of emotion, research on vocal expression is relatively rare, particularly with respect to the comparison between languages and cultures (see a recent review by Scherer et al., 2011). To the best of our knowledge, there have been no systematic, empirical attempts

to compare the acoustic patterns of vocal emotion expression in speech and singing. This article is one of the first attempts to throw some light on the hypothesis, based on evolutionary considerations (see above), that the expression of emotion in speech and singing have evolved in parallel and thus share many features. The approach chosen here is to compare the results of an acoustic analysis of emotion portrayals in neutral phrases sung by three professional opera singers with evidence from recent studies on the vocal expression of emotion (Goudbeek and Scherer, 2010; Patel et al., 2011; Sundberg et al., 2011) in order to examine the plausibility of the hypothesis of isomorphism.

When studying vocal expression through acoustic parameters it is important to distinguish pure vocalization such as affect bursts or interjections and regular speech-like utterances, because of formant structure of vowels, intonation and other factors. This is similar for "vocalises" and text based singing. In consequence, we examine nonsense text and /a/ vocalizations in both speech utterances and sung phrases. To examine the isomorphism hypothesis, we will report the results of the acoustic analysis of the phrases and vocalises sung by the three opera singers using ANOVAs to test for emotion and vocal production differences as well as post hoc comparisons of emotion groups and compare the results systematically to the published results from the Goudbeek and Scherer (2010) and Patel et al. (2011), Sundberg et al. (2011) studies using profile correlations (see also Goudbeek and Scherer, 2010, Table 2, for comparison between two studies using this approach).

2. Method

2.1. Vocal emotion portrayals in speech

The details of the methods used for the recording and analysis of the spoken stimuli the results for which are used for the comparison with the study of expressions in singing presented in this paper are reported in two separate studies (Goudbeek and Scherer, 2010; Patel et al., 2011). Only a brief overview is provided here for ease of reference.

The stimulus material was selected from the Geneva Multimodal Emotional Portrayal (GEMEP) corpus, a multimodal database in which 10 (5 male and 5 female) professional French-speaking actors (mean age of 37.1 years; age range: 25–57 years) portrayed a variety of emotions (Bänziger and Scherer, 2010). The actors were given written scenarios prior to the date of recording to help induce the emotions during an interaction with a professional stage director. The actors were requested to enact (using Stanislavski or method acting techniques) a given affective state during improvised interactions with the director (see Bänziger and Scherer, 2010, pp. 274–276, for further details). Two standard sentences (consisting of meaningless syllable combinations reflecting major vowel types and transitions) were used to realize the enactments: (a) ne kal ibam soud molen(! – indicating a statement intonation contour), (b) koun se mina lod belam(? – indicating a question contour). The sentences include only a limited number of phonemes with similar realizations in most European languages. Both sentences were constructed to include the same phonemes. In addition, the actors expressed a nonverbal affect burst (the sustained vowel /a/). The GEMEP data base contains 18 different emotions that were theoretically selected to represent the affective space spanned by the dimensions of valence, arousal, and power. In order to assess the degree to which the GEMEP stimuli can be considered valid representatives of the different emotions, extensive judgment studies were performed in different modalities and for both the sentences and the /a/s. The data on recognition accuracy for the 18 emotions in all modalities and both production types are reported in Table 6.2.4 in Bänziger and Scherer (2010, p. 281). The mean accuracy percentage over all 18 emotions amounted to .44 for the sentences and .43 for the /a/s (expected by chance 1/18 = .056).

In the study by Goudbeek and Scherer (2010) standard sentence stimuli for 12 selected emotions (covering the affective space) were acoustically analyzed (elated joy, hot anger/rage, amusement, panic fear, pride, despair, pleasure, cold anger/irritation, relief, anxiety/worry, interest, sadness/depression). Patel et al. (2011) analyzed the /a/ affect burst stimuli for a subset of five emotions (relief, sadness/depression, elated joy, panic fear, and hot anger). A large set of acoustic features were extracted and analyzed in both studies (see the respective papers for greater detail).

2.2. Vocal emotion portrayals in singing

As part of the Music & Emotion Research Focus of the Swiss Center for Affective Sciences we were able to interest three professional opera singers who are regularly interpreting major roles in leading international opera houses to

collaborate and provide experimental portrayals of emotional interpretations of musical phrases. At the first stage of the project we obtained recordings from a soprano, a mezzo-soprano, and a tenor.

2.2.1. Design and vocal production material

With respect to the emotions to be interpreted, we chose a subset of the categories that had been used in the spoken expression studies that seemed appropriate to be interpreted in the context of an opera libretto, namely anxiety, anger, despair, fear, joy, pride, sadness. We added the emotions of admiration and tenderness as we thought that these would be particularly appropriate for a lyric repertoire. We used one of the standard sentences employed in GEMEP to realize the spoken enactment "ne kal ibam soud molen" and the sustained vowel /a/. The singers were asked to sing both of these types of material on a normal vocalise (upwards and downwards), imagining that they would want to project that emotional tonality in their interpretation of a lyrical work.

2.2.2. Recording of the singing samples

The tenor was recorded at the Conservatoire de Musique de Genève. A Sennheiser HSP-4EW head worn microphone was used (cardioid pick-up pattern, condenser capsule), placed at the corner of his mouth. The audio signal was recorded by a professional sound engineer on dedicated hardware, at 96 kHz, 24-bit, on a single mono channel.

The soprano and the mezzo-soprano were recorded under the same conditions, at the Brain & Behavior Laboratory, Geneva. A Neumann BCM-104 broadcast microphone was used (cardioid pick-up pattern, condenser capsule), placed at approximately 1.5 m in front of the singer, at the height of the mouth. Before digitization, the audio signal passed through a Soundcraft RW5674 LX716-II audio console, where it was split in two separate channels. One of the two signals was attenuated by 20 dB, while the other was left untouched. This 20 dB separation increases the dynamic range covered during quantization. Later on, when analyzing recordings containing low intensity vocal performances, one typically uses the high gain channel as a larger portion of the dynamic range is covered, whereas for high intensity vocal performances, where the high gain channel is likely to saturate, one can fall back on the low gain channel. These two signals were then digitized and recorded on a PC equipped with a Digigram VX882HR sound card, at 44.1 kHz, 16-bit.

2.2.3. Segmentation of the recordings

Each session was recorded as a whole, resulting in large audio files containing all combinations of vocal content and emotion for each singer, in a randomized order. During the recording, singers were allowed to comment their own performance and to repeat stimuli which did not satisfy them. In consequence, the raw recordings then had to be manually segmented, selecting and rejecting stimuli in accordance to the singer's comments. Finally, the individual audio files were manually labeled and trimmed, removing any unwanted sounds before or after the actual stimuli (throat clearing, breathing, etc.).

As in the case of the speech stimuli, we ran a judgment study with $N = 71$ adult raters from different countries (via web experimentation) to obtain an estimation of the recognition accuracy for the productions of the three singers studied here in order to validate the representativeness of the rendering of the different emotions. The following results were obtained for the sung phrases (in parentheses we provide the comparable values for the sentences in the GEMEP study, as shown in Table 6.2.4 in Bänziger and Scherer (2010): Anger 0.64 (0.67), Fear 0.59 (0.66), Joy 0.43 (0.35), Pride 0.46 (0.24), Sadness 0.45 (0.45), Tenderness 0.36 (0.22) Chance expectancy for the sung stimuli is $1/6 = 0.17$). The results show that the sung emotion samples are reasonably well recognized (in comparing with the speech accuracy one needs to keep in mind that 18 categories were provided in the latter judgment tasks increasing the likelihood of confusions within emotion families).

2.2.4. Acoustic analysis

Acoustic parameters were extracted from long-term-average spectra of the complete utterances of the nonsense text as well as of the vowel /a/. Measures for each of the three singers were therefore organized in two sets of data, with and without voiceless elements. The analysis parameters were divided into two main groups: measures relating to the energy distribution of high and low frequencies in the spectrum, and measures reflecting the variability of the signal, both in frequency and in intensity.

The spectral energy distribution was analyzed in terms of the long-term-average spectrum (LTAS) obtained from the Line Spectrum subroutine of the Soundswell Signal Workstation software (http://savtech.se/medical/index.php/logopedi/swell). The analysis, based on FFT, calculates the absolute value of the voltage within each of a number of

frequency bins, each 400 Hz wide in the 0–5000 Hz frequency range. This analysis bandwidth was chosen since it suppressed the influence of long, high-pitched tones on the LTAS curve also in female voices.

The variability of the signal was analyzed using the PRAAT software (http://www.fon.hum.uva.nl/praat/). Since vocal loudness has a major influence on most voice parameters, measures of the sound intensity level (SIL) were also collected, again using program Soundswell Signal Workstation. SIL was computed in terms of the equivalent sound level rather than as the SPL average because the latter is systematically affected by voiceless elements and pauses and thus SIL seemed to be a more reliable choice. The equivalent sound level is a time average of the linear sound pressure, expressed in dB, and was set equal to the SPL of the (constant) calibration tone in the recording; in other words if the calibration tone had an SPL of 86 dB, equivalent sound level and hence also its SIL was 86 dB.

Measures of the following parameters were all obtained from the LTAS:

- Proportion between energy below 500 Hz and total energy (up to 5000 Hz).
- Proportion between energy below 1000 Hz and total energy (up to 5000 Hz).
- Alpha ratio, in dB, calculated as the ratio between the summed sound energy in the spectrum above and below 1000 Hz.
- $H1–H2_{LTAS}$, in dB, calculated as the difference between two averages of LTAS level, one calculated across the two or three filter bands that covered the F0 range of the utterance and the other covering the frequency bands one octave higher. For example, if F0 range of the sample was 200–350 Hz, the LTAS levels in the frequency bands corresponding to this frequency range were averaged and the same was done for the frequency bands one octave higher. Then the $H1–H2_{LTAS}$ was calculated as the difference between the former and the latter averages. As the vocalise example extended to an octave plus two semitones, the second partial of the lowest notes appeared at the F0 of the top tones. To avoid this overlap, top tones of the vocalise were excluded. For this analysis a frequency range from 0 to 1400 Hz and a 30 Hz analysis bandwidth were used.
- Hammarberg index, in dB, calculated as the difference between energy maxima in the 0–2000 Hz and 2000–5000 Hz ranges.
- Spectral slope, calculated as the slope of the linear regression line of the LTAS between 1000 and 5000 Hz.
- Spectral flatness, calculated as the logarithm of the ratio of the geometric and the arithmetic mean of the energy in the frequency bands of the LTAS.
- Spectral centroid, in Hz, calculated as the weighted mean along the frequency continuum of the energy in the frequency bands of the LTAS.

Periodicity was analyzed over the entire phrase or vocalise using the voice report option of the PRAAT program and using its standard values (frame duration 0.01 s). The algorithm for periodicity detection is based on the autocorrelation method (for further details see Boersma & Weenink, 2010). The following PRAAT parameters were collected:

- Jitter (local), defined as the average absolute difference in period between consecutive periods, divided by the average period and given as a percentage;
- Shimmer (local), defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude and given as a percentage;
- The harmonic-to-noise ratio (HNR);
- Mean autocorrelation.

We also measured F0 parameters but did not analyze these data as the F0 values were almost completely determined by the pitches of the major vocalise used to realize the sung expressions (although the tenor showed some variation of F0 values depending on the emotion to be interpreted).

To assess tempo, given the fixed nature of the vocalise, we computed the average length of tones, i.e., the duration of the sung utterance divided by the number of tones (15).

Due to the differences in F0 of the voice registers as well as idiosyncratic voice characteristics of the singers it is difficult to compare the raw values of the extracted parameters across singers. In consequence, we computed z-scores separately for each singer for each of the parameters. All of the following analyses are performed on the z-scored values.

Table 1
ANOVA results for Emotion × Materials using *z*-scores of acoustic parameters.

	Emotion		Phrases/vocalises		Interaction	
Tempo	8.1***	.65	.22	.01	.19	.04
SIL	9.3***	.68	.02	.00	.41	.09
Prop. Energy < .5 k	5.5***	.55	24.4***	.41	.61	.12
Prop. Energy < 1 k	9.7***	.69	.06	.00	.69	.14
Hammarberg index	4.7**	.52	15.6***	.31	.47	.1
Slope	3.6*	.45	4.9	.12	1.6	.27
Spectral flatness	12.9***	.75	58.3***	.63	.94	.18
Alpha	7.5***	.63	.01	.00	.67	.13
$H1H2_{LTAS}$	3.7*	.46	14.3*	.29	.30	.06
Spectral centroid	11.1***	.72	3.0	.08	1.0	.19
HNR	39.5***	.90	1.8	.05	.64	.13
Autocorrelation	59.5***	.93	43.0***	.55	6.1***	.58
Jitter	9.9***	.69	22.5***	.39	2.1	.33
Shimmer	7.5***	.63	.20	.01	.32	.07

Note: Cell entries show *F* values with significance levels *<.05, **<.01, ***<.001, and effect sizes (eta squared).

3. Results

3.1. The effects of emotions and sung material

Table A in the Supplementary Material shows the mean *z*-scores for the extracted acoustic parameters for each singer and each emotion for both the vocalises and phrases. To determine the relative importance of the effects of the major factors, *Emotion expressed* (9 different emotions) and *Materials sung* (phrase/vocalise), as well as their interaction, on the acoustic parameters, we ran a series of separate univariate ANOVAs (we did not analyze speaker effects given the small *N*). The results, including effect size estimates, are shown in Table 1. The results show significant, strong effects of emotion on all of the parameters measured, indicating that the chosen parameter set efficiently differentiates emotions. A significant effect of phrases vs. vocalises is found for six parameters. In only one case, autocorrelation, we find an interaction between the two factors.

Fig. 1 shows a typical example for separate (and highly significant) main effects of emotion and material (Fig. 1a; Proportion of energy below 500 Hz) and the only significant interaction effect (Fig. 1b; Autocorrelation). As to the effects of Material, vocalises have lower values on proportion of energy under 500 Hz, which is probably due to the constant first formant of approx. 600 Hz of the vowel /a/ used for singing the vocalise. However, this systematic difference does not affect the capacity of this parameter to differentiate emotions.

In addition, phrases generally show a lower Hammarberg index. The reason for this is probably that the phrase produced a prominent peak at 1550 and 1800 Hz in the female voices while /a/ produced a much lower peak just above 1000 Hz, presumably reflecting the second formant. Thus, the level difference between the highest peak below 1000 Hz and the highest peak above 1000 Hz was much smaller in the phrase. Finally, phrases show a lower level of autocorrelation and a higher extent of jitter. For autocorrelation we find a highly significant interaction effect (see Table 1) while for jitter the interaction is only marginally significant ($p = .059$; although the effect size is very similar). The reason for the interaction is most likely that the phrase/vocalise differences are particularly pronounced for anger and panic fear. These effects are likely to be due to rapid vowel–consonant transitions in highly aroused material.

The main effects for type of sung material reported above concern the differential *levels* of the acoustic parameters measured, *not the relative profile* over the emotions expressed. Except in the case of interaction effects (present for only one parameter – autocorrelation), the *profile* of emotion difference can be the same even though the *level* is different. To examine the profile similarity between phrases and vocalises in our data in a direct fashion, we used profile correlations between the two types of productions to examine the degree of convergence. Concretely, we correlated, separately for each acoustic parameter, the respective mean *z*-score values of the productions across the list (profile) of equivalent emotion portrayals. The results are shown in the column "a" of Table 2. The extraordinarily high correlations (effect size, >.50, Cohen, 1992) demonstrate that as far as the patterning of parameter vectors over the emotions are concerned,

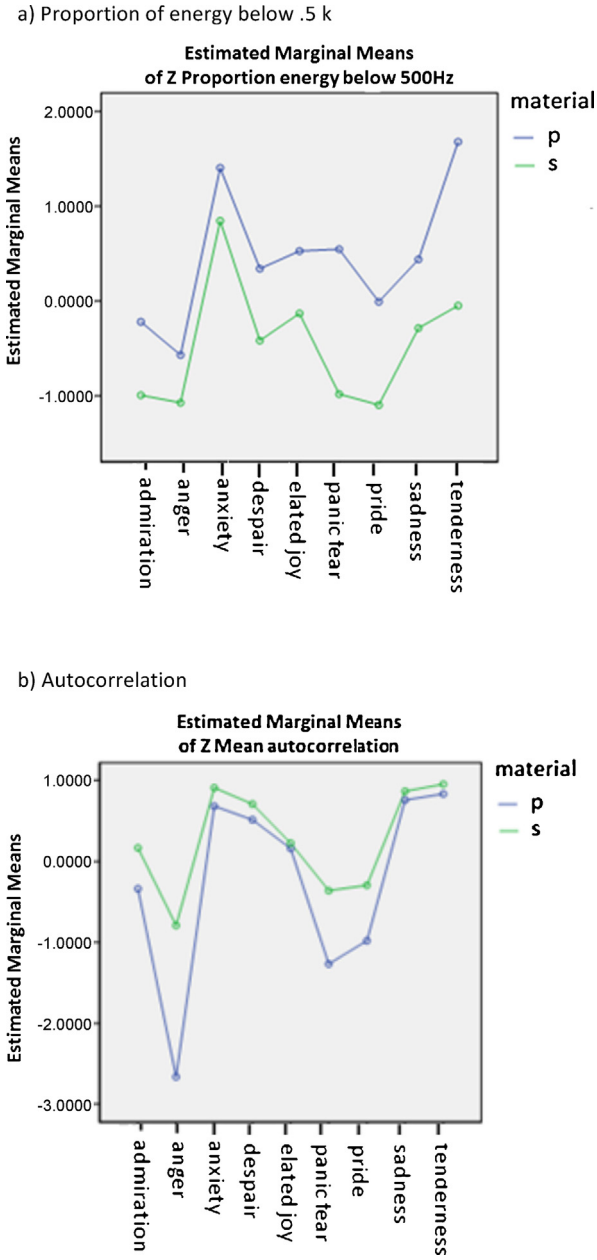

Fig. 1. Illustration of the ANOVA results – Emotion × Materials – for the *z*-scores of proportion of energy below .5 k and autocorrelation. (a) Proportion of energy below .5 k. (b) Autocorrelation

we have a very high degree of agreement between sung phrases and vocalises, suggesting that similar underlying mechanisms are at work.

Having accounted for the effects of the type of sung material, which is not the focus of the current paper, we will not explore these aspects further, especially as there is only one interaction effect. The ANOVA findings support the conclusion that similar underlying production mechanisms for differential emotion expression may be operative in both in singing vocalises on a single vowel sound (/a/) and singing phrases with meaningless phonemes, as suggested by the extremely high profile correlations reported above. In consequence, we further analyze both types of material together, increasing the statistical power of the following analyses.

Table 2
Profile correlations across emotion portrayals for the acoustic parameters of singers' productions for phrases and "vocalises" (a), actors' productions of sentences (b) and /aa/ vowels (c) in the GEMEP study (see Goudbeek and Scherer, 2010, for sentences; Patel et al., 2011, for /aa/s).

	a Singers' phrases/vocalises $N=9$	b Actors' sentences/singers' phrases $N=7$	c Actors' aa's/singers' vocalises $N=4$
Tempo	0.95	−0.40	
SIL	0.95	0.50	0.47
Prop. Energy < .5 k	0.80	0.61	
Prop. Energy < 1 k	0.87	0.74	
Hammarberg index	0.84	0.75	
Spectral flatness	0.90	0.46	
Alpha	0.85		0.63
$H1H2_{LTAS}$	0.85		0.89
Spectral centroid	0.84		
HNR	0.99	0.14	0.32
Autocorrelation	0.97	0.26	
Jitter	0.81	0.51	0.86
Shimmer	0.92	0.38	0.37

Note: The direction of the variables was adapted to correspond to the respective comparison sample (spectral flatness inverted in Goudbeek and Scherer, 2010; alpha inverted in Patel et al., 2011).

3.2. Parameter reduction

Given the strong effect sizes of the Emotion factor for many of the acoustic parameters shown in the ANOVAs described in the preceding section, it is of interest to compute post hoc comparisons to examine the exact patterns of the emotion differences on different acoustic parameters. However given the comparatively large parameter set, this would require a large number of comparisons with the associated danger of Type I errors. Therefore we decided to check for potential collinearity between parameters to determine whether parameter reduction is possible without losing too much information. Table 3 shows the correlation matrix for the acoustic parameters in this study. Inspection of the table reveals several clusters of strong correlations between subsets of parameters, suggesting a high degree of collinearity. In addition, the SIL parameter tends to correlate with almost all other parameters. This situation invites the use of an exploratory principal components analysis (PCA) to examine the degree of overlap between subsets of the parameters. To determine the factor structure, we chose the "elbow" criterion (the point at which the Eigenvalue curve flattens out) in a visual inspection of the scree plot (Fig. 2) and consequently extracted five factors and rotated them using the Varimax criterion. The rotated factor loadings, sorted for size, are shown in Table 4. The first factor consists of parameters that are linked to the energy distribution in the entire spectrum, specifically the spectral tilt or the balance between the upper and lower parts of the spectrum. In consequence, we labeled this factor Spectral balance. The second factor groups the parameters that are linked to the autocorrelation of the waveform, with low autocorrelation suggesting aperiodicity and noise (reflected in lower harmonic-to-noise ratio). The parameters jitter and shimmer, respectively measuring frequency and amplitude irregularity, also load on this factor which we therefore called Perturbation. Factor 3 is characterized by two variables that reflect the relative importance of the lower partials measuring the relative size of the lower harmonics. We labeled this factor Low Partial dominance. SIL also loads highly on this factor but has sizeable cross-loadings on other factors. The fourth factor has a single loading for spectral slope which is somewhat surprising as one might expect a relationship with spectral balance. However, as the correlation matrix in Table 3 shows there are only weak correlations with other parameters loading on the spectral balance factor (except a moderate relationship to spectral flatness). The fifth factor shows a single high loading, for Tempo.

Given this highly interpretable component structure and the high correlations shown between certain parameter subsets (see Table 3), we computed the following composite scores (means) for the parameters with dominant loadings on the first 3 components (the sign given in parentheses before some variable names indicate the direction in which it was integrated in the scale):

Table 3
Correlations between extracted acoustic parameters for the three singers.

	Tempo	SIL	Prop. Energy < .5 k	Prop. Energy < 1 k	Hammarberg index	Slope	Spectral flatness	Spectral centroid	Alpha	$H1H2_{LTAS}$	HNR	Autocorrelation	Jitter	Shimmer
Tempo	1.00	−0.58	0.39	0.61	0.50	−0.14	−0.48	−0.58	−0.64	0.47	0.78	0.68	−0.60	−0.72
SIL	−0.58	1.00	−0.75	−0.68	−0.59	0.32	0.66	0.73	0.68	−0.62	−0.73	−0.62	0.34	0.44
Prop. Energy < .5 k	0.39	−0.75	1.00	0.61	0.31	0.04	−0.22	−0.57	−0.61	0.79	0.53	0.34	−0.05	−0.28
Prop. Energy < 1 k	0.61	−0.68	0.61	1.00	0.78	−0.23	−0.65	−0.94	−0.97	0.51	0.73	0.65	−0.42	−0.51
Hammarberg index	0.50	−0.59	0.31	0.78	1.00	−0.39	−0.84	−0.85	−0.79	0.27	0.66	0.64	−0.49	−0.49
Slope	−0.14	0.32	0.04	−0.23	−0.39	1.00	0.62	0.32	0.23	0.09	−0.33	−0.39	0.30	0.22
Spectral flatness	−0.48	0.66	−0.22	−0.65	−0.84	0.62	1.00	0.76	0.66	−0.26	−0.73	−0.74	0.60	0.51
Spectral centroid	−0.58	0.73	−0.57	−0.94	−0.85	0.32	0.76	1.00	0.92	−0.46	−0.77	−0.76	0.49	0.52
Alpha	−0.64	0.68	−0.61	−0.97	−0.79	0.23	0.66	0.92	1.00	−0.53	−0.76	−0.64	0.45	0.55
$H1H2_{LTAS}$	0.47	−0.62	0.79	0.51	0.27	0.09	−0.26	−0.46	−0.53	1.00	0.62	0.42	−0.18	−0.48
HNR	0.78	−0.73	0.53	0.73	0.66	−0.33	−0.73	−0.77	−0.76	0.62	1.00	0.89	−0.75	−0.84
Autocorrelation	0.68	−0.62	0.34	0.65	0.64	−0.39	−0.74	−0.76	−0.64	0.42	0.89	1.00	−0.83	−0.76
Jitter	−0.60	0.34	−0.05	−0.42	−0.49	0.30	0.60	0.49	0.45	−0.18	−0.75	−0.83	1.00	0.73
Shimmer	−0.72	0.44	−0.28	−0.51	−0.49	0.22	0.51	0.52	0.55	−0.48	−0.84	−0.76	0.73	1.00

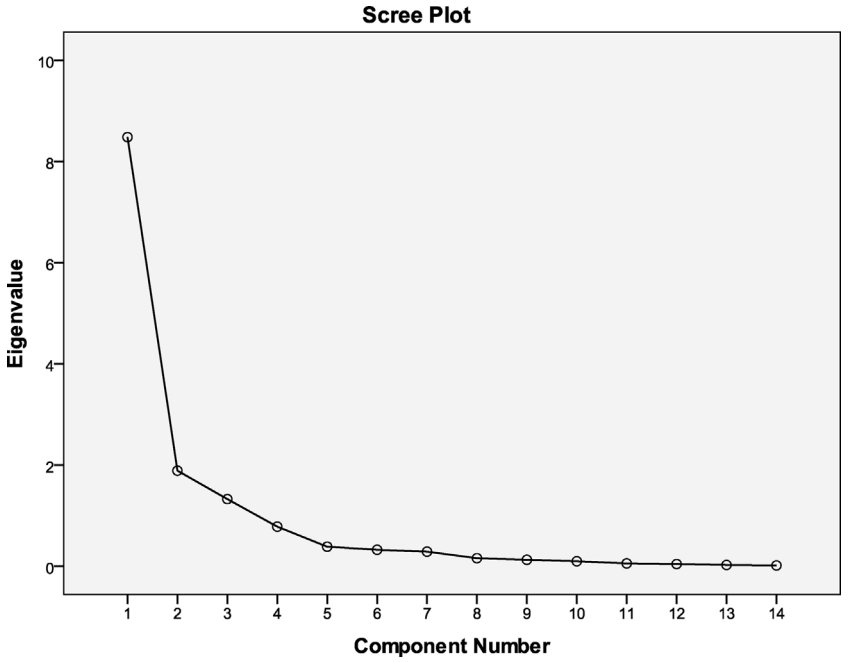

Fig. 2. Scree plot for the PCA of the extracted features.

Table 4
Varimax rotated factor loadings from a Principle Components Analysis of the extracted parameters.

	Component				
	1	2	3	4	5
Tempo	−0.31	−0.55	0.28	−0.05	**0.68**[a]
SIL	0.44	0.22	**−0.70**[a]	0.36	−0.10
Prop. Energy < .5 k	−0.31	0.01	**0.91**[a]	0.04	0.10
Prop. Energy < 1 k	**−0.85**[a]	−0.23	0.36	−0.05	0.21
Hammarberg index	**−0.85**[a]	−0.31	0.08	−0.25	0.01
Slope	0.15	0.15	0.07	**0.94**[a]	−0.04
Spectral flatness	0.62	0.46	−0.11	0.55	0.09
Spectral centroid	**0.84**[a]	0.32	−0.34	0.16	−0.06
Alpha	**0.83**[a]	0.26	−0.37	0.04	−0.25
$H1H2_{LTAS}$	−0.13	−0.27	**0.88**[a]	0.13	0.07
HNR	−0.42	**−0.72**[a]	0.44	−0.19	0.19
Autocorrelation	−0.41	**−0.80**[a]	0.24	−0.24	0.02
Jitter	0.24	**0.91**[a]	0.05	0.13	−0.03
Shimmer	0.20	**0.82**[a]	−0.23	0.05	−0.31

Note: The five factors explain 91.9% of the total variance.
[a] Loadings used to define factors are shown in bold.

- Spectral balance = (−) Hammarberg index, (−) Prop. Energy <1 k, Spectral centroid, Alpha. This composite represents the mean level difference between partials above and below 1 kHz; a high number reflects strong high partials (a relatively flat spectrum). Spectral balance is strongly influenced by vocal loudness (e.g., Sundberg and Nordenberg, 2006).
- Perturbation = Jitter, Shimmer, (−) HNR, (−) Autocorrelation. This composite combines different measures of aperiodicity, e.g., the dissimilarity between adjacent cycles (a high number represents a high degree of perturbation). Perturbation is a parameter that is regulated by the combination of glottal adjustment, subglottal pressure and sometimes also vocal tract constriction.

Table 5
Correlations between composite scales and single variables.

	SIL	Tempo	Residual perturbation	Residual low partial
Tempo	−0.45			
Residual_Perturbation	−0.04	−0.47		
Residual_Low_Partial_Dominance	−0.08	0.07	0.00	
Residual_Spectral_Balance	0.08	−0.25	0.47	−0.06

Note: SIL was partialed out of all three composite variables.

Table 6
ANOVA results for Emotion × Materials on *z*-scores of composite, residualized scores.

	Emotion		Phrases/vocalises		Interaction	
Residual low partial dominance	1.54	.26	41.12***	.54	0.78	.15
Residual spectral balance	2.41*	.36	2.73	.07	0.48	.10
Residual perturbation	4.86***	.53	6.26*	.15	1.16	.21

Note: SIL was partialed out of all three composite variables. Cell entries show *F* values and significance levels *<.05, **<.01, ***<.001, and effect sizes (eta squared).

- Low partial dominance = Prop. Energy <.5 k, $H1H2_{LTAS}$. This composite variable represents the mean level difference between the lowest and the higher spectrum partials; a high number reflects a steep negative slope, i.e. very strong low partials. Dominance of low partials is depending on glottal adduction, forceful adduction attenuating the lowest partials.

In addition to these composite scores, the following two variables were used separately – Tempo and SIL. We chose not to include the parameters spectral slope and spectral flatness in further analyses as they cross-load on all or some of three first factors and are thus difficult to interpret. While most of the above mentioned parameters belong to the frequency domain, Tempo is part of the temporal domain loading on a separate factor. This parameter is related to the motor system and the phonatory and articulatory planning.

SIL, measuring the energy or intensity of the voice, is a special case. Even though it is mainly controlled by subglottal pressure, it also is influenced by the frequency distance between the first formant the partial closest to it. It has a powerful influence on many of the acoustic parameters in the frequency domain, as demonstrated by the strong cross-loadings on the first and third factors in the rotated component matrix (Table 4). We chose to analyze it as a separate variable because of its pervasive effects on other acoustic parameters. The latter is shown by strong correlations of SIL with our composite variables. This suggests that SIL can be considered as a type of superfactor (Scherer and Fontaine, 2013), and thus we used regression analysis to partial SIL out of the three composite variables and to work with the resulting residual scores to examine the effects of the respective spectral measures independently of overall energy. The intercorrelations among the three residual composite scales (with SIL partialed out) and their correlations with the individual scales retained are shown in Table 5. The matrix shows that the variables in this set are quite independent.

The results of univariate ANOVAs (Emotion x Material) for the three residualized composite scales are shown in Table 6. They suggest that two of these parameter groups – spectral balance and perturbation – make a significant contribution, with reasonable effect sizes, to the differentiation of the emotions studied here that is independent of the overpowering effect of loudness (as measured by SIL). In contrast, low partial dominance does not contribute as much to emotion differentiation, it mainly distinguishes the production of the /a/ vocalises vs. the nonsense phrases. The latter effect is certainly due to the greater occurrence of low first formant values in the nonsense text as compare to the vowel /a/. Fig. 3 nicely illustrates this effect.

In order to evaluate the specific nature of the emotion-differentiating effects of our parameter set, we computed post hoc comparisons (using the Duncan criterion for homogeneous subgroups on the Emotion factor based on the ANOVAs shown in Tables 1 and 6) to determine the most appropriate differentiation of subsets of emotion for the respective composite variables and the three single parameters (SIL, tempo). The results of these post hoc comparisons are shown in Table 7.

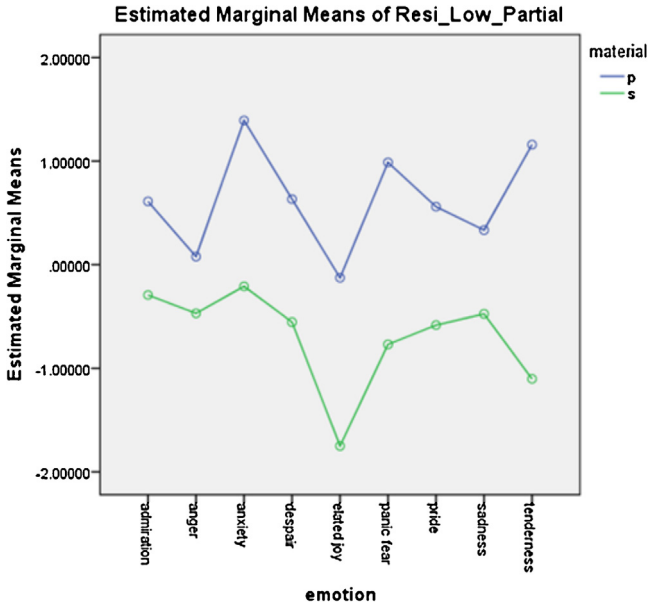

Fig. 3. Interaction Emotion × Material for the composite scale "low partials dominance".

As expected on the basis of the low effect size Low partials dominance does not provide clear separation of major emotion types. Spectral balance clearly separates anger (showing a rather flat, highly balanced spectrum) from sadness and fear on the other end of the continuum. This supports earlier findings in the literature that suggest that anger is generally characterized by strong energy in the higher partials (e.g., Banse and Scherer, 1996). Perturbation seems to separate the high arousal from the low arousal emotions, with a higher degree of perturbation and noise in the case of strong activation. This may be due to higher muscle tone and stronger subglottal pressure leading to greater variability

Table 7
Homogeneous groups of emotions according to post hoc comparisons of the ANOVA results (Duncan test) for vocalises and phrases combined reported in Table 1 (individual variables) and Table 6 (composite scales).

Acoustic parameter	Homogeneous subgroups arranged in the order of mean distances
Residual Low partials dominance	Anxiety >
	Admiration ≈ anger ≈ despair ≈ fear ≈ pride ≈ sadness ≈ tenderness >
	Joy
Residual Spectral balance	Anger >
	Admiration ≈ joy>
	Anxiety ≈ despair ≈ pride ≈ sadness ≈ tenderness >
	Fear
Residual Perturbation	Anger ≈ fear ≈ joy ≈ pride >
	Admiration >
	Anxiety ≈ despair ≈ tenderness >
	Sadness >
Tempo(duration)	Anger ≈ fear >
	Admiration ≈ pride >
	Joy >
	Despair >
	Anxiety ≈ sadness ≈ tenderness
SIL	Admiration ≈ anger ≈ fear ≈ pride >
	Despair >
	Anxiety ≈ joy ≈ sadness ≈ tenderness

Note: SIL was partialled out of all three composite variables. Within group emotions have been arranged in alphabetical order to avoid interpretation of nonsignificant order differences.

of phonation but also to the fact that the tempo of the sequence of phonatory and articulatory changes tends to be faster under high arousal. The latter hypothesis is clearly borne out by the post hoc comparison results, where Tempo clearly separates high and low arousal emotions. SIL separates the emotions basically into two groups, the difference being potentially linked to the habitual degree of intensity of the respective emotions. Anger, fear, and pride might tend to have high, anxiety, sadness, and tenderness relatively lower intensity in everyday life.

3.3. Profile similarity with speech results

In order to compare these findings with the results for spoken emotion expressions, we first tested the degree of agreement between the parameter profiles across the expressed emotions for the singers' *phrase* samples with the actors' vocal expression samples using nonsense *sentences* in the GEMEP study (both using the same nonsense syllable sequence; see Section 2 and Goudbeek and Scherer, 2010). The results (profile correlations for 11 acoustic parameters across seven emotions examined in both studies) are shown in Table 2, column b (raw *z*-scores for actors' phrases are shown in Table B in the Supplementary Material). The profile correlations of the singers' *vocalise* samples were compared to actors' productions of the *vowel /aa/* as analyzed in Patel et al. (2011) and are shown in column c of Table 2 (raw *z*-scores for actors' /aa/s are shown in Table C in the Supplementary Material). The results (profile correlations for 6 acoustic parameters across four emotions examined in both studies) show for most of the parameters medium (>.30) to strong (>.50) effect size (Cohen, 1992) correlations between the profiles. Two of the parameters linked to perturbation effects (especially the autocorrelation and the harmonics-to-noise ratio) do not attain this level which is probably due to the fact that while for the speech sentence stimuli the perturbation parameters (reflecting aperiodicity of the speech wave form) are less important than prosodic and vowel transition features, in the singers' productions these values are important because of the central role of vibrato in emotional expression. In contrast the same high correlation is found in the case of /aa/ sounds in speech and vocalise singing, possibly because actors also use aperiodicity as an expressive device in the absence of suprasegmental variation.

In consequence, except for differential use of perturbation features in sentence speech and phrase singing we find a high degree of similarity in the use voice production mechanisms for the expression of a large set of emotions for both speech and singing expressions. The areas of apparent divergence or independence, mostly linked to the importance of vibrato in singing, await further clarification and will be discussed in the conclusion.

In the interest of a more fine-grained analysis of the differences and similarities between emotion expression in speech and singing, we computed ANOVAs with post hoc comparisons (using Duncan test based homogeneous subgroups) for all those individual acoustic parameters and emotions that are shared between the singers' phrases and the spoken nonsense sentences (Goudbeek and Scherer, 2010) and the singers' vocalises and the actors' /aa/ bursts (Patel et al., 2011). The results are shown in Tables 8 and 9, respectively. Generally we find a rather high degree of similarity in both cases. The following comments are intended to comment on some of the more interesting or challenging differences between the speech and the singing samples.

In the case of the sentences/phrases we find that for Tempo anger is lower for the actors compared to the singers. This may be due to the distinction between violent rage, which is highly aroused with fast tempo, and threatening anger which can be expressed in a slow menacing but very powerful manner. Vocally, the latter requires the use of a variety of prosodic features such as emphasis of accents and rhythm changes. As the nature of the musical material allows less variation for the singers, they may have relied more strongly on the high arousal aspect of certain types of anger.

Another interesting difference is the differential use of loudness for elated joy in the case of the actor's spoken expression. This may be partly due to the instructions given to the actors and the design of the studies. In the case of spoken expressions actors had to enact both quiet pleasure and elated joy – most likely they used loudness to emphasize the difference. In contrast, the singers had only one joy category and it is likely that they portrayed more peaceful and quieter varieties of joy.

Finally, as in the earlier comparisons, one is struck by the rather large differences with respect to the perturbation variables showing much stronger Emotion effect sizes for the singers in the differentiation of emotion groups. There is clear evidence in our results that singers make much more use of vibrato, which is one type of perturbation, than speakers – possibly because of the restrictions impose by the musical score to be respected but possibly also because of the important role of vibrato in the history of singing and its use for expressive effects. While there is this strong difference in the level of perturbation in actors and singers, the respective patterns, in terms of which types of emotion

Table 8
Comparison of ANOVA results for phrases and homogeneous groups of emotions (according to post hoc comparisons, Duncan test) between singers' (this study) and actors' (Goudbeek and Scherer, 2010) portrayals of emotion.

Acoustic parameter	Phrases in the current study of emotion portrayals by professional opera singers (ANOVA results and order of homogeneous subgroups)	Phrases used for emotion portrayals by professional actors (see Goudbeek and Scherer, 2010) (ANOVA results and order of homogeneous subgroups)
Tempo	5.71**; .71 Anger ≈ fear > Anxiety ≈ despair ≈ joy ≈ pride > Sadness	2.38*; .19 Fear > Anxiety ≈ pride > Anger ≈ despair ≈ joy ≈ sadness
SIL	2.57; .52 Anger ≈ pride > Despair ≈ fear ≈ sadness > Anxiety ≈ joy	41.60***; .80 Anger ≈ fear ≈ joy > Despair ≈ pride > Anxiety > Sadness
Prop. Energy < .5 k	1.61 .41 Anxiety > Sadness ≈ despair ≈ fear ≈ joy ≈ pride > Anger	12.54***; .54 Anxiety ≈ sadness > Despair ≈ pride > Fear ≈ joy > Anger
Hammarberg index	1.59; .41 Anxiety > Fear ≈ despair ≈ joy ≈ pride ≈ sadness > Anger	13.57***; .56 Anxiety ≈ sadness > Despair ≈ fear ≈ pride > Joy > Anger
Spectral flatness	5.75**; .71 Sadness > Anxiety ≈ despair > Joy > Anger ≈ fear ≈ pride >	9.52**; .48 Sadness > Anxiety ≈ despair ≈ fear ≈ pride > Anger > Joy >
HNR	18.31***; .89 Anxiety ≈ sadness > Despair ≈ joy > Anger ≈ fear ≈ pride	3.58**; .25 Sadness > Anxiety> Anger ≈ despair ≈ fear ≈ joy ≈ pride
Autocorrelation	45.5***; .95 Anxiety ≈ despair ≈ joy ≈ sadness > Fear ≈ pride > Anger	3.83**; .27 Sadness > Anger ≈ anxiety ≈ despair ≈ fear ≈ joy ≈ pride
Jitter	6.72**; .74 Anger ≈ fear ≈ pride > Joy > Anxiety ≈ despair ≈ sadness	1.30; .11 No group
Shimmer	3.11*; .57 Anger ≈ fear ≈ joy ≈ pride > Anxiety ≈ despair ≈ sadness	3.33**; .24 Anger ≈ despair ≈ fear ≈ joy> Anxiety ≈ pride ≈ sadness

Note: Cell entries show *F* values and significance levels *<.05, **<.01, ***<.001, and effect sizes (eta squared).

show more or less perturbation, are rather similar – suggesting that the expression mechanism, and the signal value, are quite comparable.

The results of ANOVA post hoc comparison approach for singers' vocalises and actors' affect burst /aa/s (Patel et al., 2011) are shown in Table 9. Again, we find very similar patterns suggesting that similar expressive vocal devices are used in both cases. As to the exceptions, as in the case of phrases/sentences, the singers showed lower SIL for joy, confirming the suggestion that the absence of a contrast between differentially aroused versions of joy led them to generally express a quieter version of joy. In the case of $H1H2_{LTAS}$ we do not find a significant differentiation of subgroups in the case of the singers. However, the results show an ordering of the emotions similar to the actors with borderline significance.

Table 9
Comparison of ANOVA results for /aa/s and homogeneous groups of emotions (according to post hoc comparisons, Duncan test) between singer's vocalises (this study) and actors' (Patel et al., 2011) portrayals of emotion.

Acoustic parameter	Vocalises in the current study of emotion portrayals by professional opera singers (ANOVA results and order of homogeneous subgroups)	/aa/s used for emotion portrayals by professional actors (see Patel et al., 2011) (ANOVA results and order of homogeneous subgroups)
SIL	14.51***; .86 Anger ≈ fear > Sadness > Joy	141.81***; .92 Anger ≈ fear > Joy > Sadness
Alpha	5.73*; .71 Anger ≈ fear > Joy ≈ sadness	94.6***; .89 Anger ≈ fear ≈ joy > Sadness
$H1H2_{LTAS}$	2.95; .56 No groups	12.02***; .50 Sadness > Anger ≈ fear ≈ joy
HNR	33.86***; .94 Sadness > Joy > Anger ≈ fear	8.49***; .41 Fear ≈ joy ≈ sadness > Anger
Jitter	4.20; .64 Sadness > Anger ≈ fear ≈ joy	6.86**; .37 Sadness ≈ anger > Fear ≈ joy >
Shimmer	8.12*; .78 Sadness > Joy > Anger ≈ fear	23.77***; .66 Anger > Sadness > Fear ≈ joy

Note: Cell entries show *F* values and significance levels *<.05, **<.01, ***<.001, and effect sizes (eta squared).

4. Discussion and conclusions

The results of the acoustic analyses are very clear-cut: there are significant main effects of the enacted emotions on all of the parameters measured and for both types of sung material, reaching very respectable, in some cases extremely strong, effect sizes. This can be interpreted as signifying that the singers, despite the absence of emotionally meaningful lyrics, are able to modify their voice quality and the dynamic aspects of the musical rendering for expressive purposes. In addition, we found some significant main effects for the type of sung material. Thus, the presence of different vowels in the nonsense phrases seem to have affected the relative strength of the lower partials (which are, due to the formant structure), stronger in the case of the /aa/ vowel used for the vocalises. In addition, probably because of the presence of vowel–consonant transitions in the phrases, there was a larger amount of phonation irregularity for the phrases (see Fig. 1). Importantly, we found only one Emotion × Material interaction for autocorrelation (the phrase/vocalise differences being particularly prominent for anger and panic fear).

We had selected a number of acoustic parameters that are frequently employed in studies of vocal expression. As in earlier studies, we found several of these parameters highly intercorrelated because they are assessing similar dimensions in slightly different measurement operations. A principal components analyses revealed a factor structure that is quite similar to the ones found earlier (Goudbeek and Scherer, 2010; Patel et al., 2011). Because of this stability over different studies, speakers/singers, and materials, we decided to combine some of the parameters to composite scales – dominance of low partials, spectral balance, and perturbation (keeping loudness level and tempo as separate variables) to simplify the interpretation of the data. As these composite scales, particularly the first two, were strongly correlated with overall loudness level (SIL), we decided to partial out the latter superfactor and work with residual scores to obtain the specific effect of the respective dimension without the effect of loudness.

The ANOVAs showed significant emotion effects for the composite scales spectral balance and perturbation but not for low partials dominance. The effects for the individual parameters averaged for the latter composite scale had in fact been weaker in the individual analyses, resulting in their common denominator not allowing to strongly differentiate

the different emotions. Thus, despite the overall concordance of findings for both types of singing material, the vowel /aa/ may be more appropriate in future studies because fixed formant structure allows a higher degree of replicability, compared to varying formant frequencies in material with greater variety of vowels.

In addition to the overall ANOVAs we computed post hoc comparison to determine the specific effects of particular parameters on specific emotions. The results show a clear differentiation between some emotions at the extreme poles of the respective dimension with less clear-cut distinctions (with respect to differential levels of the parameter) for emotions situated in the mid range. We replicate the frequently reported finding that anger has relatively strong energy in the higher frequency range (higher spectral balance). Whereas low arousal emotions (like sadness) show little waveform irregularity, have a lower loudness level, and are sung with a slower tempo, perturbations, loudness, and tempo are increased for high arousal emotions like anger and fear.

While these results are interesting in their own right, especially as there is little in terms of comparable data in the literature, the main aim of this paper was to compare the findings with comparable data from actor-produced vocal emotion expressions in speech-like portrayals. In order to examine the degree of similarity between the two expression modalities, we ran profile correlations for the individual variables that had been measured in the respective studies over the shared set of emotions. In addition, we computed post hoc comparisons with homogeneous subgroups for the speech data in two different studies and compared the respective groups. Overall, we found a rather striking degree of agreement with somewhat of an exception for the perturbation dimension. It seems that singers make general more use of a specific type of phonatory irregularity in the form of vibrato (possibly because of the classic significance of this parameter in the domain of operatic singing). The difference is particularly pronounced in the case of anger. While singers use strong vibrato for high arousal emotions including anger, actors express anger generally with firm, regular phonation. Some actors tend to show perturbations in sadness (possibly tremor) whereas this is rarely the case for singers. One important factor involved in this dimension may be that singers, given the constraints of musical structure and composition, have fewer acoustic parameters to work with and thus privilege vibrato as a mechanism of choice.

Despite these very encouraging results, we need to mention a number of limitations in the current study. It would clearly be advantageous to have a larger number of singers representing a wider range of tessitura. However, based on pilot studies with singing students, we decided that the approach chosen could only be used with excellent singers benefitting from extensive experience on major opera stages. Clearly, this population is extremely difficult to contact and it is even more difficult to obtain their collaboration on what is an arduous and time consuming task. Therefore, we privileged expertise and experience over numbers for this initial study. There is no doubt that our results require future replication with a larger set of different voice types but numbers are always likely to be fairly small. Furthermore, from the standpoint of acoustic analysis, the duration of approx. 8 s for a voice sample is too short to expect a high degree of spectral stabilization, especially for the parameters based on the long-term spectrum. At the same time, it is difficult to see how this limitation can be overcome given the ecological constraints – emotions are short-lived episodes and strong expressions in voice and face often occur only at the apex of emotional arousal. Also, for reasons of comparison with the published results on vocal emotion expression in speech, a similar format for the enacted expressions needed to be adopted.

As a general conclusion we can retain that overall there are many similarities in vocal emotion expression across speech and singing, confirming theorizing that suggests a parallel evolution of speech and music from primitive affect bursts, sharing similar codes for the signaling of affect (Scherer, 1991, 2013b). It is to be hoped that future work will extend this type of inquiry to the physiological underpinnings of phonation and articulation in speech and singing in emotion expression, leading to a better understanding of the utility of different parameters by linking parameters directly to underlying production mechanisms. Another avenue for further development, especially when data for larger groups of singers and actors become available, is the use of machine learning to assess the degree of generalizability of the underlying acoustic profiles. A recent study by Weninger et al. (2013) shows that algorithms trained on emotional music are quite successful on emotional speech and vice versa, suggesting that this may be a very viable approach. Weniger et al. show that the effect generalizes, to some extent, even to environmental sounds. This suggests that the similarity of emotion coding in the audio domain may be quite robust and that, inconsequence, the results we obtained with recordings under studio conditions, would generalize, at least in large part, to real life recordings of many different kinds of music. Obviously, it is to be expected that noise and degraded acoustic conditions (low bandwidth channel, background noise, reverberation) as well as genre-specific vocal effects (e.g., belting), may affect the degree of generalizability.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.csl.2013.10.002.

References

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70 (3), 614–636.

Bänziger, T., Scherer, K.R., 2010. Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus. In: Scherer, K.R., Bänziger, T., Roesch, E.B. (Eds.), Blueprint for Affective Computing: A Sourcebook. Oxford University Press, Oxford, pp. 271–294.

Boersma, P., Weenink, D., 2010. Praat: doing phonetics by computer [Computer program], Version 5.1.43, August 2010 from http://www.praat.org/

Bryant, G.A., Barrett, H.C., 2008. Vocal emotion recognition across disparate cultures. J. Cogn. Cult. 8, 135–148.

Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159, http://dx.doi.org/10.1037/0033-2909.112.1.155.

Fonseca, N., 2011. Singing Voice Resynthesis Using Concatenative-based Techniques. University of Porto, Portugal (Unpublished doctoral dissertation).

Goto, M., Nakano, T., Kajita, S., Matsusaka, Y., Nakaoka, S.I., Yokoi, K., 2012. VocaListener and VocaWatcher: imitating a human singer by using signal processing. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, pp. 5393–5396.

Goudbeek, M., Scherer, K.R., 2010. Beyond arousal: valence and potency/control in the vocal expression of emotion. J. Acoust. Soc. Am. 128 (3), 1322–1336.

Howes, P., Callaghan, J., Davis, P., Kenny, D., Thorpe, W., 2004. The perception of vibrato in Western operatic singing: its relationship to measured vibrato onset, rate, and extent, listener preference, and emotional expression. J. Voice 18 (2), 216–230.

Jansens, S., Bloothooft, G., de Krom, G., 1997. Perception and acoustics of emotions in singing. In: Proceedings of the 5th Eurospeech, Rhodes, IV, pp. 2155–2158.

Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull. 129, 770–814.

Kenmochi, H., Ohshita, H., 2007. Vocaloid – commercial singing synthesizer based on sample concatenation. In: Proceedings of Interspeech 2007, Antwerp, pp. 4009–4010.

Kotlyar, G.M., Morozov, V.P., 1976. Acoustical correlates of the emotional content of vocalized speech. Soviet J. Phys. Acoust. 22, 208–211.

Laukka, P., Eerola, T., Thingujam, N.S., Yamasaki, T., Beller, G., 2013a. Universal and culture-specific factors in the recognition and performance of musical affect expressions. Emotion 13, 434–449.

Laukka, P., Elfenbein, H.A., Söder, N., Nordström, H., Althoff, J., Chui, W., Iraki, F.K., Rockstuhl, T., Thingujam, N.S., 2013b. Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. Frontiers Psychol. 4, 353.

Maynard Smith, J., Harper, D., 2003. Animal Signals. Oxford University Press, Oxford.

Mortillaro, M., Mehu, M., Scherer, K.R., 2013. The evolutionary origin of multimodal synchronisation and emotional expression. In: Altenmüller, E., Schmidt, S., Zimmermann, E. (Eds.), Evolution of emotional communication: from sounds in nonhuman mammals to speech and music in man. Oxford University Press, Oxford, pp. 3–25.

Patel, S., Scherer, K.R., Bjorkner, E., Sundberg, J., 2011. Mapping emotions into acoustic space: the role of voice production. Biol. Psychol. 87, 93–98.

Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.A., 2009. Factors in the recognition of vocally expressed emotions: a comparison of four languages. J. Phon. 37, 417–435.

Risset, J.C., 1991. Speech and music combined: an overview. In: Sundberg, J., Nord, L., Carlson, R. (Eds.), Music, Language, Speech, and Brain. Macmillan, London, pp. 368–379.

Sauter, D.A., Eisner, F., Ekman, P., Scott, S.K., 2010. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc. Natl. Acad. Sci. U.S.A. 107, 2408–2412.

Scherer, K.R., 1986. Vocal affect expression: a review and a model for future research. Psychol. Bull. 99, 143–165.

Scherer, K.R., Oshinsky, J.S., 1977. Cue utilization in emotion attribution from auditory stimuli. Motiv. Emot. 1, 331–346.

Scherer, K.R., 1991. Emotion expression in speech and music. In: Sundberg, J., Nord, L., Carlson, R. (Eds.), Music, Language, Speech, and Brain. Wenner-Gren Center International Symposium Series. Macmillan, London, pp. 146–156.

Scherer, K.R., 1995. Expression of emotion in voice and music. J. Voice 9 (3), 235–248.

Scherer, K.R., Banse, R., Wallbott, H.G., 2001. Emotion inferences from vocal expression correlate across languages and cultures. J. Cross-Cult. Psychol. 32 (1), 76–92.

Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Commun. 40, 227–256.

Scherer, K.R., Clark-Polner, E., Mortillaro, M., 2011. In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. Int. J. Psychol. 46 (6), 401–435.

Scherer, K.R., 2013a. The singer's paradox: on authenticity in emotional expression on the opera stage. In: Cochrane, T., Fantini, B., Scherer, K.R. (Eds.), The Emotional Power of Music. Oxford University Press, Oxford, pp. 55–73.

Scherer, K.R., Fontaine, J.R.J., 2013. Driving the emotion process: the appraisal component. In: Fontaine, J.R.J., Scherer, K.R., Soriano, C. (Eds.), Components of emotional meaning: a sourcebook. Oxford University Press, Oxford, pp. 186–209.

Scherer, K.R., 2013b. Emotion in action, interaction, music, and speech. In: Arbib, M.A. (Ed.), Language, music, and the brain: a mysterious relationship. MIT Press, Cambridge, MA, pp. 107–139.

Siegwart, H., Scherer, K.R., 1995. Acoustic concomitants of emotional expression in operatic singing: the case of Lucia in Ardi gli incensi. J. Voice 9 (3), 249–260.

Sundberg, J., 1978. Synthesis of singing. Swedish J. Musicol. 60, 107–112.

Sundberg, J., 1989. The Science of the Singing Voice. Northern Illinois University Press.

Sundberg, J., Iwarsson, J., Hagegard, H., 1995. A singer's expression of emotions in sung performance. In: Hirano, M., Fujimura, O. (Eds.), Proceedings of the Vocal Folds Physiology Conference 1994 (S. 217–232). Singular Publishing Group, San Diego, CA, pp. 217–229.

Sundberg, J., Nordenberg, M., 2006. Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. J. Acoust. Soc. Am. 120, 453–457.

Sundberg, J., Patel, S., Björkner, E., Scherer, K.R., 2011. Interdependencies among voice source parameters in emotional speech. IEEE Trans. Affect. Comput. 99, 2423–2426.

Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M., Scherer, K.R., 2013. On the acoustics of emotion in audio: what speech, music, and sound have in common. Frontiers Psychol – Emot. Sci. 4 292, 1–12.