Eyben *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:19
DOI 10.1186/s13636-015-0057-6

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Emotion in the singing voice—a deeper look at acoustic features in the light of automatic classification

Florian Eyben[1,3,8*], Gláucia L Salomão[2,5], Johan Sundberg[2,6], Klaus R Scherer[3] and Björn W Schuller[1,3,4,7]

Abstract

We investigate the automatic recognition of emotions in the singing voice and study the worth and role of a variety of relevant acoustic parameters. The data set contains phrases and vocalises sung by eight renowned professional opera singers in ten different emotions and a neutral state. The states are mapped to ternary arousal and valence labels. We propose a small set of relevant acoustic features basing on our previous findings on the same data and compare it with a large-scale state-of-the-art feature set for paralinguistics recognition, the baseline feature set of the Interspeech 2013 Computational Paralinguistics ChallengE (ComParE). A feature importance analysis with respect to classification accuracy and correlation of features with the targets is provided in the paper. Results show that the classification performance with both feature sets is similar for arousal, while the ComParE set is superior for valence. Intra singer feature ranking criteria further improve the classification accuracy in a leave-one-singer-out cross validation significantly.

Keywords: Emotion recognition; Singing voice; Acoustic features; Feature selection

1 Introduction

Automatic emotion recognition from speech has been a large research topic for over a decade. Early papers have covered psychological and theoretical aspects of emotion expression in speech (e. g., [1]) and presented early ideas for building systems to recognise emotions expressed in human speech (e. g., [2, 3]). In contrast, emotion recognition from the singing voice has largely been overlooked, although the expression of emotion in music and singing is a highly visible and important phenomenon [4]. In this paper, we apply methods from speaking voice emotion recognition to singing voice emotion recognition, evaluate classification performances for the first time, and take an in-depth look at important acoustic features.

The paper is structured as follows: the next section (2) gives an overview of related work and an in-depth introduction to the topic of vocal emotion recognition. The data-set of sung emotions is described in Section 3, followed by the description of the acoustic features in Section 4. A description of feature selection by correlation-based ranking and a discussion of the most highly ranked features is given in Section 5. The classification experiments and their results are discussed in Section 6 and the conclusions are drawn in Section 7.

2 Related work

The topic of speech emotion recognition has gained momentum in recent years. An early overview of basic methods is given in [5], while a basic comparison of performances on widely used speech emotion corpora in early studies is given in [6]. Following the world's first Emotion Recognition Challenge held at Interspeech 2009 [7], the methods have been extended and tranferred to many other, yet related, areas in follow-up challenges (e. g., sleepiness and alcohol intoxication [8] and conflict, emotion, and autism [9]).

The topic of recognition of emotions in the singing voice, on the other hand, has gained little attention (cf. [10–12]), although the fact that emotions are visible in acoustic properties of the voice has been frequently acknowledged [13, 14]. In particular, in music, emotions

*Correspondence: eyben@tum.de
[1]MISP Group, Technische Universität München, Theresienstr. 90, Munich, Germany
[3]Université De Genève, Geneva, Switzerland
Full list of author information is available at the end of the article

play a major role and singers must be able to easily express a wide range of emotions. There are a few existing studies that deal with enthusiasm in karaoke singing [15, 16], which is close to the emotional dimension of arousal, or target vocal tutoring systems [17]. Previous findings in [18] suggest that the expression of emotions in speaking and singing voice are related. Further, [12] concludes that similar methods and acoustic features can be used to automatically classify emotions in speech, polyphonic music, as well as emotions perceived by listeners in or associated by them with other, general sounds. This suggests that the methods for speech emotion recognition can be transferred to singing emotion recognition. Therefore, this paper investigates the performance of state-of-the-art speech emotion recognition methods on a data set of singing voice recordings and compares this to the performance of a newly designed acoustic feature set, which is based on findings in [18].

3 Singing voice database

A subset of the database of singing emotions was first introduced in [18]. Here, we use an extended set, where recordings from five additional professional opera singers have been added (eight in total) following the same protocol. In the full set—as used here—there are four male (two tenors, one countertenor, one barytone) and four female singers (two sopranos, two mezzos) in total. They were asked to portray the 11 emotion classes shown in Table 1 while singing the standard scale ascending and descending using vocalises (a) and a nonsense phrase ("ne kal ibam soud molen"). The sessions were recorded as a whole without pause, and manual segmentation was performed into the scale and phrase parts. In this way, a total of 300 instances was obtained. The distribution of instances across classes is nearly balanced (cf. Table 1).

Figure 1 shows plots of the pitch contours for one of the female singers singing the same scale ascending and descending in emotionally neutral, angry, sad, and proud styles. Clear differences among the emotions concerning the style and type of vibrato can be seen. For sadness, there is a large variation in the strength of vibrato and very little vibrato during the ascending scale, also the tempo is reduced. Most vibrato is found for anger, closely followed by pride, supporting the fact that this feature is likely an indication of arousal and enthusiasm [16, 19].

Table 1 Emotion classes and number of instances for each class, mappings to ternary arousal (0–2) and valence (− 0 +)

Emotion	Number of instances	Arousal	Valence
Neutral, no expression	24	1	0
Fear	30	2	−
Passionate love	24	1	+
Tense arousal	24	1	−
Animated joy	31	2	+
Triumphant pride	30	1	+
Anger	29	2	−
Sadness	30	0	−
Tenderness	30	0	+
Calm/serenity	24	0	0
Condescension	24	0	−

4 Acoustic features

We propose a feature set based on previous, careful analysis of acoustic parameters with respect to emotional expression in the singing and speaking voice as was presented in [18]. The parameters contained in the set (referred to as *EmoFt* henceforth) are listed in Table 2. The features are based on the principle of static analysis, i. e., a single feature vector is extracted for each analysis segment. In our case, a segment is a whole phrase or scale. Low-level descriptors (LLD) and their first-order delta (difference) coefficients are computed and statistical functionals are applied to the LLD contours in order to summarise the LLD over time, e. g., the LLD are aggregated over each segment into a number of summary statistics such as the mean value, the standard deviation, etc. This approach is adopted in the proposed feature set. Further, long-term average spectrum (LTAS)-based features used by [18] are added. Thereby, the aggregation is performed by computing the LTAS over a segment—as the arithmetic mean of the magnitude spectra of all frames (20 ms, see below) within the segment—and then computing a single vector of spectral properties from this LTAS (see Table 2 for details). These features are joint with the functionals of LLD by concatenation. Additionally, modulation spectra of F_0 and of the auditory loudness are computed and the locations of the maximum amplitude of each of these are added as two additional static features. The modulation spectrum is computed with a resolution of 0.25 Hz for a range from 0.25 to 32.0 Hz. Finally, the equivalent sound level (mean of frame-energy converted to dB) is added. In total, a 205-dimensional feature vector is obtained: 19 LLD (and 19 delta coefficients), each summarised by 5 functionals ($(19+19)\cdot 5 = 190$), 12 LTAS features, 2 modulation spectrum features, and the equivalent sound level, yield a total of 205 features.

For details on implementations of individual parameters, the reader is referred to the documentation of openSMILE and to [20]. The most important parameterization are given in the following: all spectral (including LTAS) and energy-related LLD are computed from 20-ms-long overlapping windows at a rate of 10 ms (50 % overlap). A Hamming window is applied prior to the FFT for these descriptors. F_0, jitter, and shimmer are computed from 60-ms-long overlapping windows at a rate of 10 ms. Before computing the FFT for F_0 computation, a

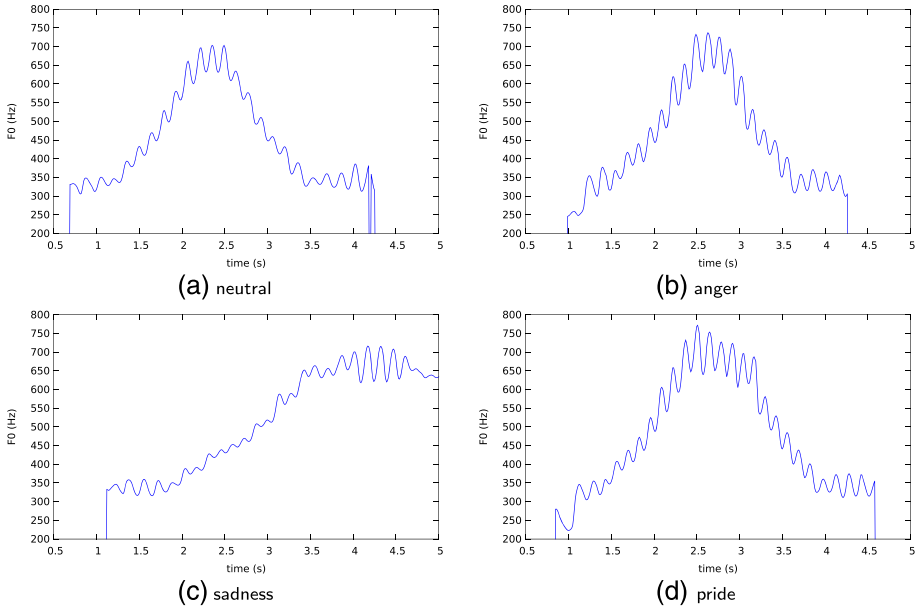

Fig. 1 Pitch (F_0) contour examples for the same scale sung by a female opera singer in different emotional styles: **a** neutral, **b** angry, **c** sad, and **d** proud; excerpts from 0.5 to 5 s are shown

Table 2 Two hundred five acoustic features in the proposed feature set (EmoFt): low-level descriptors (LLD) and functionals (brute-force combination) as well as features derived from the long-term average spectrum (LTAS) and three other features (see text)

19 LLD
Loudness, spectral flux and entropy
Energy in bands 0–0.5 and 0–1 kHz
Slope of log. power spectrum 0–1, 0–5, and 1–5 kHz
Alpha ratio (in dB), Hammarberg index (in dB)
MFCC 1–4, harmonics-to-noise ratio
F_0, prob. of voicing, jitter, and shimmer (local)
Five functionals
Arithmetic mean, standard deviation
Fifth and 95th percentile and range 5–95 %
Long-term average spectrum (LTAS), 27 bands
MFCC 1–4, spectral entropy
Energy in bands 0–0.5 and 0–1 kHz
Slope of log. band spectrum 0–1, 0–5, and 1–5 kHz
Alpha ratio (in dB), Hammarberg index (in dB)
Others
Equivalent sound level (in dB)
Frequency with maximum amplitude in modulation spectrum of F_0 and loudness

Gaussian window ($\sigma = 0.4$) is applied. F_0 is computed via sub-harmonic summation (SHS) followed by Viterbi smoothing. No windowing is performed for jitter and shimmer computation, which is performed in the time domain.

The equivalent sound level (LEq) is computed as the arithmetic mean (converted to decibels (dB) after averaging) of the frame-wise root mean-square (RMS) energy ($\mu_{\text{rms}E}$).

For the LTAS, the linear magnitude spectrum computed from the 20-ms frames is reduced to a linear 27-band power spectrum. The bands are 400 Hz wide (except for bands near the high and low borders of the frequency range from 0–5 kHz) with centers at multiples ($n = 0 \ldots 26$) of 187.5 Hz. The band spectra are averaged across all frames in a segment to obtain the LTAS. Harmonics-to-noise ratio (HNR) is computed via autocorrelation as the ratio of the first peak in the F_0 range to the peak at 0 delay in the autocorrelation function (ACF) (cf. [21]).

The alpha ratio is defined as the ratio of the energy below 1 kHz and between 1 and 5 kHz, the Hammarberg index is defined as the ratio of the highest energy peak in the 0–2 kHz region to that of the highest peak in the 2–5 kHz region [22]. Spectral slope measurements are conducted as described in [23] in Section 2.2.2 by least squares error fitting of a line to the given spectral power densities.

We compare the rather small EmoFt set to a state-of-the-art feature set used in the field of computational paralinguistics: the baseline feature set of the INTERSPEECH

2013 Computational Paralinguistics Challenge (ComParE) [9]. It was demonstrated in [12] that the features in this set provide robust, cross-domain assessment of emotion in speech, music, and acoustic events. For details on this set, we refer to [9] and [12]. In Tables 3 and 4, a detailed list of LLD and functionals contained in this set is provided. In total, the set contains 6373 features.

The motivation for this comparison of feature sets (in contrast to joining EmoFt and ComParE features to a single set), is twofold: first, the EmoFt set is based on prior work and experience of the authors, as well as psychological and acoustic studies regarding singing voice emotion (cf. [13, 14] for the spectral and prosodic parameters; and [12] for justification of lower order MFCCs)—it can be thus regarded as an "expert" designed feature set for the task of identifying emotions in the singing voice; the ComParE set is a brute-forced set, from another (yet closely related) domain (computational paralinguistics). Our goal is to compare both sets as they are, the "expert" set (EmoFt) vs. the "brute-force" set (ComParE). Second, due to this motivation of the sets, the two sets contain redundant descriptors, so simply merging is also sub-optimal.

All the acoustic features have been extracted with our openSMILE toolkit version 2.1 [24].

5 Feature selection

Feature selection is based on rankings of the features by the Pearson correlation coefficients (CC) of the features with the ternary arousal and valence labels. Three strategies for ranking-based feature selection are employed, namely ranking by absolute value of CC, by absolute value of CC after the features were normalised to 0 mean and variance 1 for every singer individually (SPKSTD-CC), and by a cross-domain correlation coefficient (Weninger-CDCC) introduced by Weninger et al. in [12].

The purpose of the CDCC measure is to weigh high correlation among single singers against correlation deviations across different singers. Thus, it ranks feature both by their correlation with the target and by the consistency of this correlation across different singers. Features which are not consistently highly correlated, and thus are not suitable for singer independent classification, are penalized. For S singers, it is defined for feature f as:

$$\mathrm{CDCC}_f^{(S)} = \frac{1}{(S-1)S}\left(\sum_{i=0}^{S}\sum_{j=i+1}^{S}\left|r_f^{(i)} + r_f^{(j)}\right|\right) - \frac{1}{(S-1)S}\left(\sum_{i=0}^{S}\sum_{j=i+1}^{S}\left|r_f^{(i)} - r_f^{(j)}\right|\right)$$

Table 3 Sixty-four low-level descriptors (LLD) of the ComParE feature set

Four energy-related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum (modulation loudness)
RMS energy, zero-crossing rate
Fifty-five spectral LLD
RASTA-style auditory spectrum, bands 1–26 (0–8 kHz)
MFCC 1–14
Spectral energy 250–650 Hz, 1 k–4 kHz
Spectral roll off points 0.25, 0.50, 0.75, 0.90
Spectral flux, centroid, entropy, slope
Variance, skewness, kurtosis
Psychoacoustic sharpness and harmonicity
Six voicing-related LLD
F_0 via sub harmonic summation (SHS) and Viterbi smoothing
Probability of voicing, logarithmic HNR by waveform matching
Jitter (local and delta), shimmer (local)

Table 4 Applied functionals in the ComParE feature set

Functionals applied to LLD / Δ LLD
Quartiles 1–3, three inter-quartile ranges
One percentile (≈ min), 99th percentile (≈ max)
Percentile range 1–99
Position of min / max, range (max–min)
Arithmetic mean[a], root quadratic mean
Contour centroid, flatness
Standard deviation, skewness, kurtosis
Rel. duration LLD is above 25 / 50 / 75 / 90% range
Rel. duration LLD is above 25 / 50 / 75 / 90 % range
Rel. duration LLD is rising
Rel. duration LLD has positive curvature
Gain of linear prediction (LP), LP coefficients 1–5
Mean, max, min, std. dev. of segment length[b]
Functionals applied to LLD only
Mean value of peaks
Mean value of peaks—arithmetic mean
Mean / std.dev. of inter peak distances
Amplitude mean of peaks, of minima
Amplitude range of peaks
Mean / std.dev. of rising / falling slopes
Linear regression slope, offset, quadratic error
Quadratic regression *a*, *b*, offset, quadratic error
Percentage of non-zero frames[c]

[a] Arithmetic mean of LLD/positive Δ LLD
[b] Not applied to voice related LLD except F_0
[c] Only applied to F_0

where $r_f^{(i)}$ is the correlation of feature f with the target (arousal, valence) for singer i. Feature reduction is performed by selecting the N features with highest rank.

In Table 5, the top three LLD (in combination with the functional the LLD was ranked highest with) obtained with each of the three strategies are shown for valence and for arousal.

In our results we find that delta (δ) coefficients of LLD are important when no singer normalisation is done, while only non delta LLD are among the top three with singer normalisation. Thus, the change in an LLD seems to be less affected by intra singer variability than the absolute value (e.g., by a speaker-dependent bias). When normalised, the LLD seems to be a better indicator of emotion than the deltas. Moreover, from EmoFt the pitch and voice quality (VQ) features dominate the top three, while from ComParE more spectral band and cepstral descriptors dominate (which are not contained in the EmoFt set)—yet still mixed with VQ and cepstral ones. This highlights the importance of the latter descriptors. The high ranking of spectral and cepstral descriptors can be attributed to their simple and robust extraction algorithms. Higher-level features like pitch and jitter are more affected by noise (even low levels of noise or non-proper voicing of sounds) and errors of the extraction algorithm (e. g., octave errors for F_0).

Table 5 Top three LLD with their highest ranked functional for CC-based ranking and CC-based ranking after singer normalisation of features (SPKSTD-CC) as well as CDCC-based ranking; Pearson correlation coefficients given in parentheses for each feature; EmoFt (top) and ComParE (bottom) feature sets

Arousal	Valence
SPKSTD-CC, EmoFt	
Jitter, fifth percentile (0.55)	F_0, 95th percentile (−0.21)
Shimmer, mean (0.52)	Jitter, range (−0.19)
Loudness, pos. mean (0.49)	Voice prob., mean (0.19)
Weninger-CDCC, EmoFt	
Jitter, fifth percentile (0.56)	Voice prob., fifth percentile (0.20)
Shimmer, mean (0.53)	F_0, fifth percentile (0.18)
F_0, range (0.49)	Loudness modulation, max freq. (0.16)
CC, EmoFt	
Jitter, fifth percentile (0.46)	F_0 δ, pos. mean (−0.18)
Shimmer, mean (0.44)	Jitter δ, 95th percentile (−0.17)
Loudness δ, pos. mean (0.43)	F_0, 95th percentile (−0.16)
SPKSTD-CC, ComParE	
Jitter, first quartile (0.60)	MFCC 5, third quartile (−0.322)
Log. HNR, pos. mean (−0.59)	log. HNR, skewness (−0.24)
Shimmer, first quartile (0.59)	RASTA f. band 10, LP gain (−0.24)
Weninger-CDCC, ComParE	
Jitter, first quartile (0.61)	MFCC 5, third quartile (0.26)
Shimmer, first quartile (0.59)	log. HNR, skewness (0.26)
Jitter DDP, second quartile (0.56)	MFCC 13, LPC2 (0.24)
CC, ComParE	
Loudness δ, pos. mean (0.51)	RASTA f. band 10, LPC 3 (−0.249)
Spec. centroid δ, IQR 2–3 (0.50)	Spec. variance, seg. len σ (0.20)
Sharpness δ, first quartile (−0.50)	log. HNR, qaud. reg. err. (−0.20)

Highly ranked for valence is the tenth RASTA-filtered auditory band, which is centered at 1287 Hz and has a bandwidth of 360 Hz (triangular filter on the Mel scale). The RASTA filter emphasises envelope modulations in the 4–8 Hz region. Therefore, it can be concluded that speech-range-modulated energy around 1.3 kHz is particularly relevant for the expression of valence. Other bands are also important, however not as single bands, but more the overall structure, as is underlined by the importance of spectral variance and HNR, both suggesting that the harmonic structure is important.

6 Experiments and results

We now describe the classfication experiments performed: classification of 11 emotion classes and three discrete levels of arousal and valence with both the full EmoFt feature set and the full ComParE feature set. Next, the results of the feature normalisation methods discussed above and the effects of ranking-based feature selection methods are explored. In all experiments, we apply support vector machines (SVMs) with linear kernel function as implemented in WEKA [25] with sequential minimal optimisation (SMO) as training algorithm, due to their good baseline performance in many related speech emotion recognition studies (cf., e. g., [9] and [8]). Other classifiers could have been also used and compared, however this study is a study on the relevance of acoustic parameters for classification in general. Thus, SVM are chosen as a possible classifier, of which we know it can handle the task sufficiently. It will be used first, to benchmark and compare the various feature selection and normalisation strategies. Model complexity constants C of 0.1 and 1.0 are used for these experiments. Next, to assess how much additional performance could be gained by classifier tuning, SVM model complexity and kernel functions are optimized systematically from a selected set.

In order to evaluate the cross-singer classification performance, we perform leave-one-singer-out (LOSO) cross validation in all experiments: all data from one singer constitutes the test set, the remaining seven singers constitute the training set. The experiment is then run eight times, using each singer once as the test set.

In order to scale all features to a common range and avoid numerical issues in linear SVM kernel evaluations, the feature vectors have to be normalised prior to SVM model training and evaluation. We investigate two

Table 6 Results (unweighted average recall (UAR)) with all features of proposed acoustic features set (EmoFt 200 features) and the INTERSPEECH 2013 ComParE feature set (6373 features) for ternary arousal and valence tasks; normalisation on training fold (mean/variance); leave-one-singer-out cross validation

[UAR %]	EmoFt		ComParE		Chance
SVM complexity *C*	0.1	1.0	0.1	1.0	-
Global (training fold) feature normalisation:					
Eleven emotion classes	22.4	24.4	28.0	28.0	9.1
Valence (three classes)	36.4	40.2	46.2	46.2	33.3
Arousal (three classes)	57.9	52.4	54.6	54.6	33.3
Per singer feature normalisation:					
Eleven emotion classes	23.6	28.1	38.2	38.2	9.1
Valence (three classes)	40.1	43.6	48.7	48.7	33.3
Arousal (three classes)	57.6	49.9	57.6	57.6	33.3

strategies for this to evaluate the influence of inter singer effects: baseline standardisation of each feature to mean 0 and variance 1 based on data from the training fold (STD) and per singer standardisation of each feature to mean 0 and variance 1 within the data of each singer (SPKSTD).

Table 6 shows the classification results obtained with the full EmoFt and ComParE sets with the two feature normalisation strategies. Results are reported in terms of unweighted average recall (UAR). UAR is computed as the unweighted average of the class-wise recall rates for N classes as follows:

$$\mathrm{UAR} = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{n_i} \tag{1}$$

where c_i is the number of correctly detected instances of class i and n_i the total number of instances of this class present in the evaluation partition. With a one-sided paired z-test, an upper bound for the significance of the results (with 300 instances) can be estimated: an absolute difference of 6.7 % is required for two results to be significantly different at $\alpha = 0.05$, and 9.4 % are required for significance at $\alpha = 0.01$. With this, all results are significantly above chance level. $C = 0.1$ is slightly better for arousal and EmoFt, while $C = 1.0$ seems to be better (for the harder) valence task. Notably, the two complexity settings show no difference on the ComParE set.

We further performed experiments with the same protocol as applied for results in Table 6 but with a reduced set of features by each of the three feature selection strategies (cf. Section 5). The number of retained features with high rank is varied from $N = 1$ to $N = 200$ in steps of 10. Results are plotted in Fig. 2 for ternary arousal classification and in Fig. 3 for ternary valence

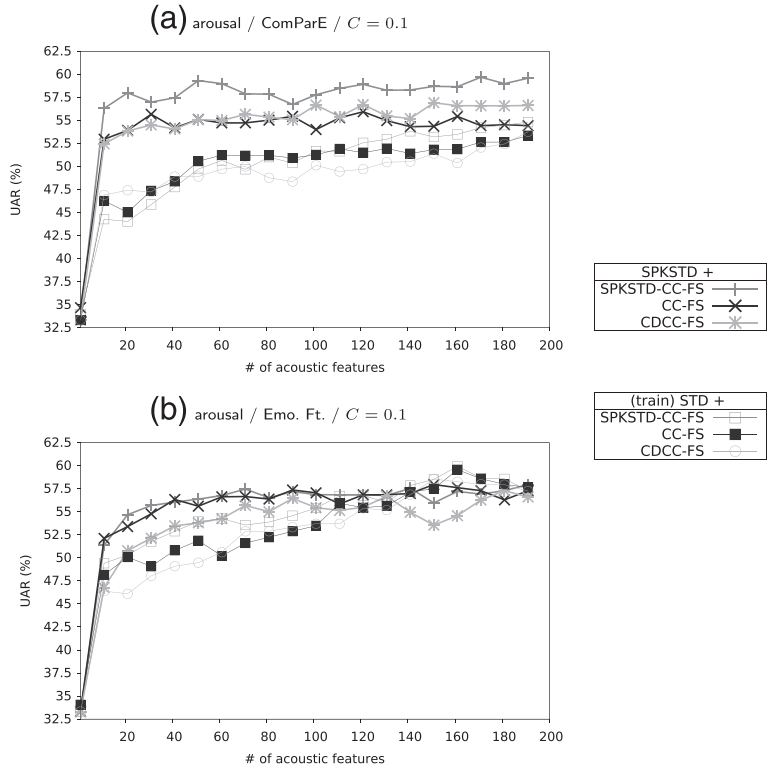

Fig. 2 Unweighted average recall (UAR) for ternary arousal; three feature selection (FS) methods (details in the text); two feature normalisation strategies for the classifier (per singer or for training set); LOSO; varying number of features (top 1–200 from ranking); linear kernel SVM; **a** ComParE feature set (*top*) and **b** EmoFt feature set (*bottom*)

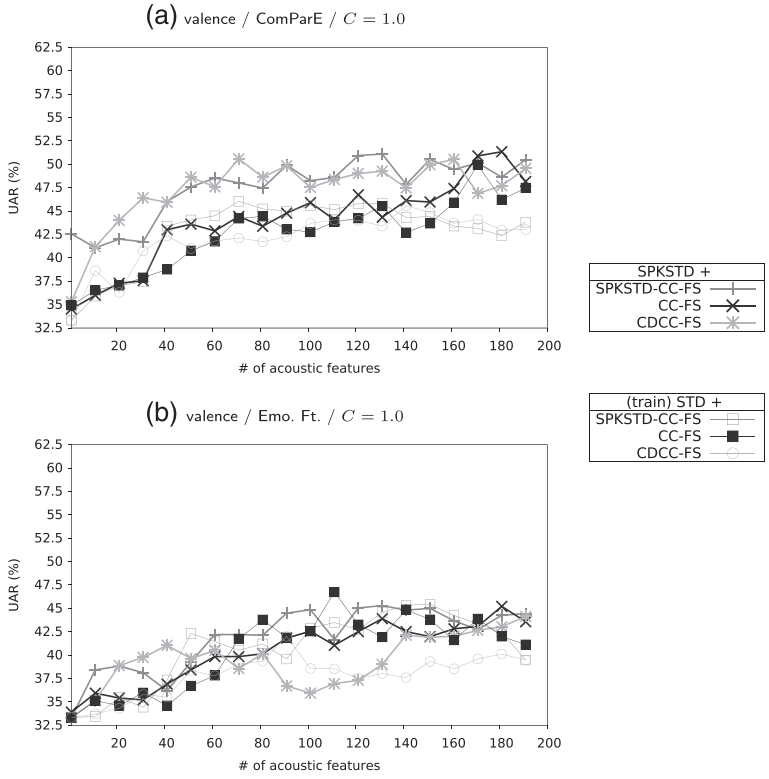

Fig. 3 Unweighted average recall (UAR) for ternary valence; three feature selection (FS) methods (details in the text); two feature normalisation strategies for the classifier (per singer or for training set); LOSO; varying number of features (top 1–200 from ranking); linear kernel SVM; **a** ComParE feature set (*top*) and **b** EmoFt feature set (*bottom*)

classification. It can be clearly seen that for arousal classification, very few features are required and the results converge very quickly, while for valence classification, adding more features generally improves the performance, especially with the larger ComParE set.

In terms of the best performing feature normalisation and feature selection strategies, no significant conclusion can be drawn but a clear tendency is visible in the plots, which is consistent with the results in Table 6: per singer normalisation is superior and both the Weninger-CDCC-based FS as well as the SPKSTD-CC-FS are superior to the simple CC-FS, which does not account for inter singer variability. Especially for the Weninger-CDCC-FS, a gain can be observed for small feature sub-sets (except for arousal and EmoFt). This gain vanishes, however, for a higher number of selected features. In the case of EmoFt, this is obvious, as all methods converge to the same set (all EmoFt features) at the end of the plot. Concluding, we

Table 7 Best results after parameter tuning for arousal and valence tasks with according settings (feature set, feature selection method, and percentage of features). Top, overall best settings; Mid, overall best with only the EmoFt set; Bottom, overall best with only the EmoFt set and linear kernel SVM

Dimension	UAR [%]	Setting
Overall best results		
Arousal	61.7	ComParE, CC-FS (50 %), linear, $C = 0.05$
Valence	52.9	ComParE, CC-FS (10 %), linear, $C = 1.0$
EmoFt set best results		
Arousal	60.1	EmoFt, CC-FS (50 %), RBF kernel, $C = 1.0$
Valence	48.0	EmoFt, CC-FS (50 %), RBF kernel, $C = 1.0$
EmoFt set, only linear kernel SVM, best results		
Arousal	58.0	ComParE, SPKSTD-CC (30 %), linear, $C = 0.05$
Valence	45.3	ComParE, all features, linear, $C = 1.0$

Table 8 Confusion matrix for the best arousal result, three levels of arousal (low, mid, high); top, classified as; left, ground truth emotion label

#	Low	Mid	High
Low	77	19	12
Mid	21	56	25
High	19	18	53

can say that there is no big difference between Weninger-CDCC-FS and SPKSTD-CC-FS, except for small feature sets, where it seems that the best ranked Weninger-CDCC features are superior to those ranked by other methods.

A deeper analysis of classifier parameters has been conducted in order to assess the potential of tuning parameters to the task. For feature normalisation, only the per singer standardisation was kept. In addition to a linear kernel SVM, a radial basis function (RBF) SVM was considered. The RBF gamma parameter was varied from 0.5 down to 0.00001 in steps of fifths and halves, i. e., 0.5, 0.1, 0.05, and 0.01. The SVM complexity parameter C was varied from 1.0 down to 0.00001 in steps of halves and fifths, i. e., 1.0, 0.5, 0.1, and 0.05. The three feature selection methods, CC, SPKSTD-CC, and Weninger-CDCC, as well as no feature selection were compared for all the above settings. For each feature selection method, a fixed number of selected features was varied over 10, 30, 50, and 70 %. In order to be able to systematically compare all feature selection methods at all classifier settings, the analysis was restricted to the ternary arousal and valence tasks. The best results for each dimension and the according settings are found in Table 7. It can be clearly seen that using a fraction of the ComParE feature set yields the best results, although not significantly better than the EmoFt set. For the (larger) ComParE set, the linear kernel SVMs are superior, while for the smaller EmoFt set, the RBF kernel appears to be the better choice, which is expected due to the initial small feature space dimensionality. For the overall best results, the according confusion matrices are shown in Table 8 (for arousal) and Table 9 (for valence). For arousal—as one would expect—confusions between low/mid and mid/high are slightly more frequent than between the extremes (low/high). In contrast, for valence, interestingly, confusions between the extremes (pos/neg) seem to be very frequent, while confusions with neutral seem to be rare. This is in line with findings that valence from acoustic parameters (for speech) is hard to identify, thus the high number of pos/neg confusions. For the singing voice, however, there seems to be a clearly distinct neutral valence style, though.

Table 9 Confusion matrix for the best valence result, three levels of valence (negative (neg.), neutral (neut.), and positive (pos.)); top, classified as; left, ground truth emotion label

#	Neg.	Neut.	Pos.
Low	79	5	53
Mid	10	26	12
High	50	11	54

7 Conclusions

We have successfully applied state-of-the-art speech emotion recognition methods to the problem of automatic recognition of emotions in the singing voice. Pitch- and jitter/shimmer-based features are found to be highly ranked in the proposed EmoFt feature set, while in the larger ComParE set, spectral band descriptors and MFCCs show an even higher correlation. Normalising features to zero mean and unit variance for each singer individually brings a consistent performance gain across all experiments, which is marginally significant.

In future work, we want to consider other feature ranking metrics, such as Bayes error and information gain and perform feature rankings on even larger and broader sets of acoustic features, but also using expert feedback on the implications of the most highly ranked features in order to design new descriptors which are better correlated with the problem and at the same time deepen our understanding of which features contribute to the expression of emotion in the singing voice.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The first author, FE, implemented the feature extraction, performed the experiments, and coordinated the writing of the text. The second author, GLS, helped to design the acoustic feature set, interpret the results, and proofread the manuscript. The third author, JS, provided details on the LTAS features used in [18], helped interpret the results, and was always available for fruitful discussions and provided valuable comments. The fourth author, KS, planned the experiments and data collection, collected and provided the database, and consulted the team on acoustic parameters. The fith author, BS, helped design the acoustic feature set and experiments and co-edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by an ERC Advanced Grant in the European Community's Seventh Framework Programme under grant agreement 230331-PROPEREMO (production and perception of emotion: an affective sciences approach) to Klaus Scherer and by the National Center of Competence in Research (NCCR) Affective Sciences financed by the Swiss National Science Foundation (51NF40-104897) and hosted by the University of Geneva as well as through an ERC Starting Grant No. 338164 (iHEARu) awarded to Björn Schuller.

Author details

[1]MISP Group, Technische Universität München, Theresienstr. 90, Munich, Germany. [2]Department of Speech Music Hearing, School of Computer Science and Communication, KTH (Royal Institute of Technology), Stockholm, Sweden. [3]Université De Genève, Geneva, Switzerland. [4]Department of Computing, Imperial College London, London, UK. [5]Department of Linguistics, Stockholm University, Stockholm, Sweden. [6]University College of Music Education, Stockholm, Sweden. [7]Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany. [8]audEERING UG (limited), Gilching, Germany.

Received: 20 February 2014 Accepted: 13 May 2015
Published online: 30 June 2015

References

1. R Cowie, in *Proc. of the ISCA Workshop on Speech and Emotion*, ISCA. Describing the emotional states expressed in speech (Newcastle, Northern Ireland, 2000), pp. 11–18
2. R Cowie, E Douglas-Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, J Taylor, Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)
3. J Liscombe, J Venditti, J Hirschberg, in *Proc. of Eurospeech*, ISCA. Classifying subject ratings of emotional speech using acoustic features (Geneva, Switzerland, 2003), pp. 725–728
4. KR Scherer, Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? J. New Music Res. **33**(3), 239–251 (2004)
5. D Ververidis, C Kotropoulos, Emotional speech recognition: resources, features, and methods. Speech Commun. **48**(9), 1162–1181 (2006)
6. B Schuller, B Vlasenko, F Eyben, G Rigoll, A Wendemuth, in *Proceedings of ASRU*. Acoustic emotion recognition: A benchmark comparison of performances (IEEE Merano, Italy, 2009). doi:10.1109/ASRU.2009.5372886
7. B Schuller, S Steidl, A Batliner, F Jurcicek, in *Proc. of INTERSPEECH*, ISCA. The Interspeech 2009 Emotion Challenge (Brighton, UK, 2009), pp. 312–315
8. B Schuller, A Batliner, S Steidl, F Schiel, J Krajewski, in *Proc. of INTERSPEECH*, ISCA. The INTERSPEECH 2011 speaker state challenge (Florence, Italy, 2011), pp. 3201–3204
9. B Schuller, S Steidl, A Batliner, A Vinciarelli, K Scherer, F Ringeval, M Chetouani, *et al*, in *Proc. of INTERSPEECH*. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism (ISCA Lyon, France, 2013), pp. 148–152
10. J Sundberg, J Iwarsson, H Hagegard, in *Vocal Fold Physiology: Voice Quality Control*, ed. by O Fujimura, M Hirano. A singer's expression of emotions in sung performance (Singular Press San Diego, CA, USA, 1995), pp. 217–229
11. J Sundberg, Emotive transforms. Phonetica. **57**, 95–112 (2000)
12. F Weninger, F Eyben, BW Schuller, M Mortillaro, KR Scherer, On the acoustics of emotion in audio: what speech, music and sound have in common. Front. Emotion Sci. **4**(Article ID 292), 1–12 (2013)
13. KR Scherer, Expression of emotion in voice and music. J. Voice. **9**(3), 235–248 (1995)
14. J Sundberg, *The Science of the Singing Voice*. (Northern Illinois University Press, DeKalb, 1989)
15. R Daido, S-J Hahm, M Ito, S Makino, A Ito, in *Proc. of ISMIR 2011, Miami, Florida*. A system for evaluating singing enthusiasm for karaoke (University of Miami, 2011)
16. R Daido, A Ito, M Ito, S Makino, Automatic evaluation of singing enthusiasm for karaoke. Compututer, Speech, and Language. Elsevier. **28**(2), 501–517 (March 2014). doi:10.1016/j.csl.2012.07.007
17. O Mayor, J Bonada, A Loscos, in *Proc. 121st AES Convention*. The singing tutor: expression categorization and segmentation of the singing voice (San Francisco, CA, 2006)
18. KR Scherer, J Sundberg, L Tamarit, GL Salomão, Comparing the acoustic expression of emotion in the speaking and the singing voice. Comput. Speech Lang. Elsevier. **29**(1), 218–235 (2013). doi:10.1016/j.csl.2013.10.002
19. T Nakano, M Goto, Y Hiraga, in *Proc. INTERSPEECH 2006*. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features (ISCA Pittsburgh, Pennsylvania, USA, 2006), pp. 1706–1709
20. F Eyben, Realtime speech and music classification by large audio feature space extraction. Dissertation, Technische Universität München, Germany, Springer Theses, Springer (2014)
21. B Schuller, *Intelligent Audio Analysis*. Signals and communication technology. (Springer, New York, 2013)
22. B Hammarberg, B Fritzell, J Gauffin, J Sundberg, L Wedin, Perceptual and acoustic correlates of abnormal voice qualities. Acta Otolaryngologica. **90**, 441–451 (1980)
23. L Tamarit, M Goudbeek, KR Scherer, in *Proc. of SPKD-2008*. Spectral slope measurements in emotionally expressive speech (ISCA, 2008). paper 007
24. F Eyben, F Weninger, F Gross, B Schuller, in *Proc. of ACM MM 2013, Barcelona, Spain*. Recent developments in openSMILE, the munich open-source multimedia feature extractor (ACM New York, NY, USA, 2013), pp. 835–838
25. M Hall, E Frank, G Holmes, B Pfahringer, P Reutemann, IH Witten, The WEKA data mining software: an update. ACM SIGKDD Explorations Newslett. **11**(1), 10–18 (2009)