IDENTIFYING EMOTIONS IN OPERA SINGING: IMPLICATIONS OF ADVERSE ACOUSTIC CONDITIONS

Emilia Parada-Cabaleiro[1] **Maximilian Schmitt**[1] **Anton Batliner**[1]
Simone Hantke[1,2] **Giovanni Costantini**[3] **Klaus Scherer**[4] **Björn W. Schuller**[1,5]
[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3] Department of Electronic Engineering, University of Rome Tor Vergata, Italy
[4] Department of Psychology, University of Geneva, Switzerland
[5] GLAM – Group on Language, Audio & Music, Imperial College London, UK
emilia.parada-cabaleiro@informatik.uni-augsburg.de

ABSTRACT

The expression of emotion is an inherent aspect in singing, especially in operatic voice. Yet, adverse acoustic conditions, as, e. g., a performance in open-air, or a noisy analog recording, may affect its perception. State-of-the art methods for emotional speech evaluation have been applied to operatic voice, such as perception experiments, acoustic analyses, and machine learning techniques. Still, the extent to which adverse acoustic conditions may impair listeners' and machines' identification of emotion in vocal cues has only been investigated in the realm of speech. For our study, 132 listeners evaluated 390 nonsense operatic sung instances of five basic emotions, affected by three noises (brown, pink, and white), each at four Signal-to-Noise Ratios (-1 dB, -0.5 dB, +1 dB, and +3 dB); the performance of state-of-the-art automatic recognition methods was evaluated as well. Our findings show that the three noises affect similarly female and male singers and that listeners' gender did not play a role. Human perception and automatic classification display similar confusion and recognition patterns: sadness is identified best, fear worst; low aroused emotions display higher confusion.

1. INTRODUCTION

Singing is a channel to communicate emotion that goes beyond culture or time, as shown by a variety of common musical representations across the world over centuries: as, e. g., lullabies [39] (typical expression of parental love) or spiritual chant [20] (typical expression of mystic feelings). In western music, the emotional expression in singing voice is inexorably linked to the *Italian Opera* which has had, from the XVIII century (through the development of the *belcanto* [38]) till the XIX century (with the advent of the *Melodramma Verdiano* [38]) a focus on the dramatic–emotional interpretation of the opera's characters [10].

The *Opera* was born in Italy at the beginning of the XVII century as an 'entertainment' [12]. Even though Opera is no longer the most common leisure activity, its cultural importance is still shown by thousands of 'opera performances' made every year—6,795 only in Germany for the 2015 / 2016 season [1] ; and by thousands of 'opera recordings' available in multi-media libraries—21,054 items only in the *Istituto Centrale per i Beni Sonori ed Audovisivi* (The National Italian Audiovisual Institute [2]). Yet, opera may face 'real-world' acoustic degradation, e. g., from open-air performances [3] or from analog recordings [22]. Indeed, improving the acoustics of an opera house is a central topic of sound engineering [2], as well as the application of digital signal processing solutions to the restoration of old recordings [11].

Even though emotion in opera singing has been studied from the perceptual [34], acoustic [32], and automatic recognition [7] point of view, it has not been evaluated so far up to which extent restricted acoustic quality affects the perception and classification of emotion in singing. In this regard, we present a perceptual study (based on a forced-choice categorical [6] and dimensional [28] test), performed by 132 Italian listeners, who evaluated 390 nonsense instances, sung by 6 professional opera singers (3 female), in 5 emotional states (hot anger, elated happiness, depressive sadness, panicked fear, and worried fear), subsequently masked by three noises (white, pink, and brown) at 4 signal-to-noise ratios (-1 dB, -0.5 dB, +1 dB, and +3 dB). The performance of state-of-the-art emotion recognition methods based on a Support Vector Machine classifier and `ComParE` features [36] is evaluated as well. In Section 2, related work is described; Sections 3 and 4 evaluate the database and the listening test; Section 5 discusses the results for the machine learning approach; finally, Section 6 outlines conclusions and future work.

[1] `http://operabase.com/top.cgi?lang`
[2] `http://opac2.icbsa.it/vufind/`

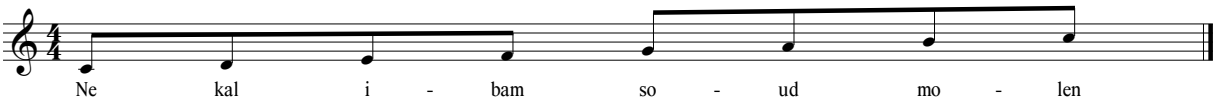

Figure 1: The nonsense utterance 'Ne kal ibam soud molen' sung in an ascending scale for each emotional state.

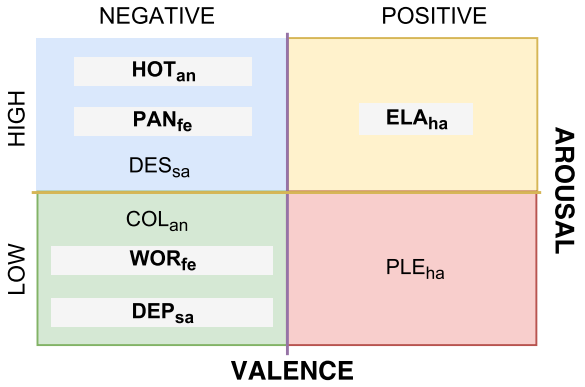

Figure 2: Correspondence between emotion categories and the bi-dimensional model of the five 'real' labels, i. e., hot anger (HOT$_{an}$), elated happiness (ELA$_{ha}$), depressive sadness (DEP$_{sa}$), panicked fear (PAN$_{fe}$), and worried fear (WOR$_{fe}$), in bold; and the three dimensional 'distractors', i. e., cold anger (COL$_{an}$), pleasured happiness (PLE$_{ha}$), and desperate sadness (DES$_{sa}$), considered in the perception test.

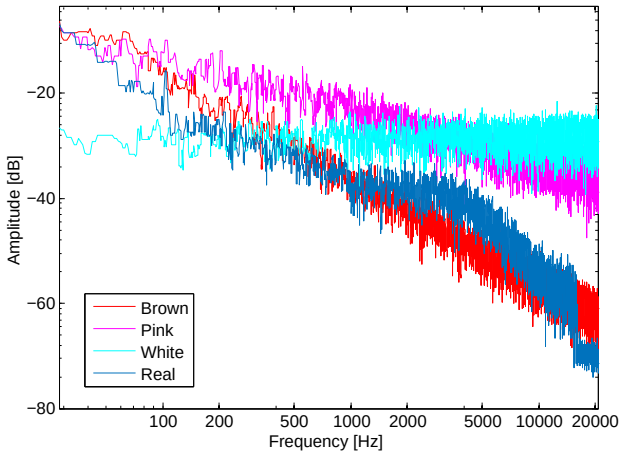

Figure 3: A comparison of the spectral distribution between 0–2 kHz and -80–0 dB, for the brown, pink, white, and real noise.

2. RELATED WORK

Even though emotions are typically expressed through the voice, emotional singing has received little attention compared to emotional speech [29]. Yet, the similarity between both channels (i. e., speech and singing [19, 33]) has recently encouraged researchers to analyse the expression and perception of sung emotional content [4]. Methods typically used in emotional speech research [1] have also been applied to singing—with special attention to the operatic voice—such as acoustic evaluation [23, 27, 32, 37] or perception assessment [15, 16, 34]. Furthermore, in the realm of affective computing, state-of-the-art machine learning techniques, typically used in audio signal processing for speech emotion recognition, have also been applied to the study of the *a cappella* singing voice [7, 40].

In the assessment of emotional speech, it has been shown that listeners' perception, acoustic feature analysis, and machine learning techniques, are affected by noisy backgrounds [25, 35], which are typical of 'real-world' environments and recordings. Yet, although singing mostly takes place in adverse acoustic conditions, the extent to which these may impair a listener's ability to perceive its inherent emotion, and how the robustness of automatic systems for emotion recognition in singing might be impaired, has not been, to the best of our knowledge, assessed so far.

3. METHODOLOGY

3.1 An Emotional Corpus of *a Cappella* Opera Singing

We took into account a selection of sentences from a dataset of the emotional singing voice [7, 33] in which professional opera singers performed a variety of sentences in different emotional states correlated to several levels of arousal (intensity) and valence (hedonistic value). Since linguistic meaning may influence listener perception of the emotional content, in order to avoid such a bias [31], the nonsense sentence *ne kal ibam soud molen!* has been considered. For a gender-balanced distribution of voice types, six singers have been selected: three females (two sopranos and one mezzosoprano) and three male (two tenors and one countertenor), who produced five times the nonsense sentence with an ascending scale (cf. Figure 1), each time expressing a different emotional state.

Following previous research on the perception of emotion in operatic voice [16], four basic emotions have been considered: anger, with high arousal (intensity), i. e., hot anger; happiness, high aroused, i. e., elated happiness; sadness, low aroused, i. e., depressive sadness; and fear, with both high arousal, i. e., panicked fear, and low arousal, i. e., worried fear (cf. Figure 2). Thus, considering one nonsense sentence, expressed in five emotional states by six singers, 30 'clean' stimuli in total have been employed.

3.2 Manipulation Techniques

The perception of emotion in speech is especially compromised by pink, and to a lesser extent by white and brown noise [25]. In Figure 3, the spectrum of a 'real' background noise, digitised from a 'no-musical fragment' of an LP recording [3], is compared with brown, pink, and white noise. The 'real' noise displays higher energy in the lowest frequencies, presenting a negative slope of approximately 6 dB per octave up to 1 kHz, a constant area from 1 kHz to 3 kHz, and a fall of energy of approximately 10 dB per octave above 3 kHz. Its acoustic characteristics makes it most similar to brown noise, which presents a negative

[3] Recording of the aria *Vissi d'amore* (Puccini's *Tosca*), interpreted by Giannina Arangi and produced in 1932 by *Columbia records*.

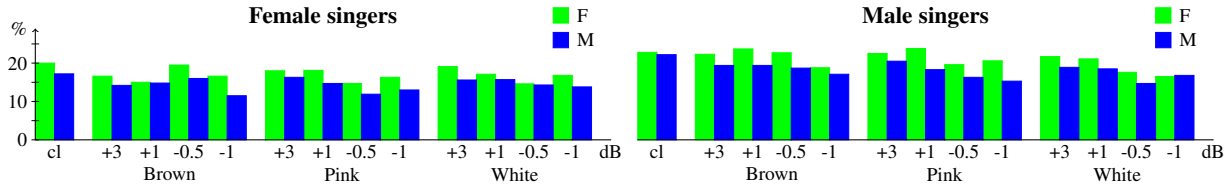

Figure 4: Mean accuracy in % of the 'real' emotions (cf. caption Figure 2) perceived by female (F) and male (M) listeners; in clean (cl) conditions, background noises (brown, pink, and white), and 4 SNR (-1 dB, -0.5 dB, +1 dB, and +3 dB); sung by females and males.

slope of around 6 dB per octave ($1/f^2$ noise); slightly similar to pink, with a negative slope of approximately 3 dB per octave (1/f noise); and dissimilar to white, whose flat spectrum presents all the frequencies at the same level. Note that the given comparison aims at exemplifying potential similarities between 'real' and 'artificially generated' noise; noise from different recordings may display higher similarity with pink, white, or other noise types.

We evaluated listeners' perception of emotion in adverse acoustic conditions by applying four Signal-to-Noise Ratio (SNR) levels (-1 dB, -0.5 dB, +1 dB, +3 dB) and three noises (brown, pink, white) to the 'clean' samples. The noises, normalized to -1 dB, have been artificially generated and mixed (at the specified SNR value) in *Matlab R2014a* [21]. Given 6 singers, 1 sentence sung in 5 emotional states, 3 noises, and 4 applied SNR levels yields 6 x 1 x 5 x 3 x 4 = 360 'noisy' samples plus 30 'clean' samples = 390 stimuli in total.

4. PERCEPTION STUDY

4.1 Emotion Measurement

The two prominent models considered to evaluate listeners' perception of emotional speech, i. e., the categorical [6], which identifies each emotional state with a specific category, and the dimensional [28], which identifies each emotional state within a continuous hyper-space characterised by dimensions—commonly arousal (from low to high) and valence (from negative to positive)—have already been applied to the perceptual evaluation of emotion in singing [16,26]. Yet, which of them would be more suitable to evaluate listeners' perception of emotion, is still an open question in both the musical domain [5] and speech research [18]. Both models have been taken into account for the perception test, i. e., each of the 4 considered basic emotions—anger, happiness, sadness, and fear (cf. Section 3.1)—has been defined in the bi-dimensional space, by having a level of arousal and valence (cf. Figure 2).

Five of these eight emotional categories (hot anger, elated happiness, depressive sadness, panicked fear, and worried fear), are 'real' emotions effectively expressed by the singers in the dataset. The other three (cold anger, pleasured happiness, and desperate sadness), so-called 'distractor labels' [24]—emotion categories not displayed in the evaluated data, have the purpose to 'distract' the listeners by minimising the chances of performing 'discrimination' rather than 'recognition' [30]. Furthermore, disgust and surprise (the remaining two basic emotions—in addition to anger, fear, sadness, and happiness—amongst those known as 'big six' [6]), have also been considered as 'distractors', without indicating a specific dimensional level; we thus present a balanced set of perceptual choices: five 'real' emotions and five 'distractors'.

4.2 Listening test setup

In total, 132 Italian listeners (55 f, 77 m, mean age 20.7 years, standard deviation 2.5 years) took part in the perception study. The participants were all students of the engineering faculty of the 'Tor Vergata' university (Rome) and received credits for their participation. To avoid fatigue, the 390 stimulus were similarly distributed into four sessions, each designed to last not longer than 30 minutes. Out of the 132 listeners, 101 had no musical instruction, 27 were self-taught in piano or guitar, 4 had studied in the conservatory—piano (2), flute, and accordion. Their musical interest was mostly in pop (65 listener), rock (45 listeners), and hip-hop (22 listeners); other genres as, e. g., Italian music, heavy-metal, or classic were underrepresented (less than 10 listeners). Since none of them had studied singing or demonstrated interest in opera, we consider them as a unique group of non-experts.

The test was designed as a forced-choice task; the ten emotion categories were presented and the participants could choose one out of them after listening to each stimulus (an initial training was provided). The test was hosted on a browser based interface (accessible from any computer) provided through the gamified crowd-sourcing platform *iHEARu-PLAY* [14]. To ensure a consistent listening environment, the participants were instructed to use earphones. Although the listeners had the possibility of listening to each stimulus indefinitely, they were encouraged to answer spontaneously to the randomized samples.

4.3 Results and discussion

Emotions were identified best in clean conditions; female listeners were slightly more accurate than male; emotions in male voices were somewhat better identified than in female voices (cf. Figure 4). Listeners' and singers' gender-related differences turned out to be not significant. In the former case, the biggest distance, i. e., female and male listeners evaluating male voices in pink noise at -1 SNR (21.6% vs 17.6%), corresponds to a p value in Pearson Chi-square of $= .47$ (way above the conventional threshold for significance of $p < .05$). In the latter case, the biggest distance, i. e., female and male voices perceived by male listeners in clean conditions (17.2 % vs 22.2 %) did not yield a significant difference either ($p = .37$). Thus, the further evaluations will not consider gender.

%	Real emotions					Distractor labels					#
	HOT$_{an}$	*ELA*$_{ha}$	*DEP*$_{sa}$	*PAN*$_{fe}$	*WOR*$_{fe}$	*COL*$_{an}$	*PLE*$_{ha}$	*DES*$_{sa}$	*DIS*	*SUR*	
HOT$_{an}$	26.6	14.7	05.5	03.2	05.2	26.4	9.2	02.8	01.7	04.5	6
ELA$_{ha}$	13.1	18.4	07.6	04.4	05.9	19.7	18.3	04.1	03.3	05.2	6
DEP$_{sa}$	01.0	03.6	42.0	03.6	05.7	07.0	12.3	21.4	01.2	02.2	6
PAN$_{fe}$	09.8	06.1	22.8	06.8	06.6	19.7	12.0	09.6	02.3	04.3	6
WOR$_{fe}$	12.6	08.8	14.4	08.2	08.5	21.2	14.2	05.1	02.2	04.9	6
total	63.1	51.7	92.3	26.1	31.9	93.9	66.4	43.0	10.7	21.1	30

Table 1: Listeners' perception (in %) of the clean instances (#), considering 'real' emotions and 'distractors' (cf. Figure 2, disgust—DIS, and surprise—SUR). Each row gives the 'reference', darker cells indicate higher % ; listeners' and singers' gender is not considered.

%	*HOT*$_{an}$	*ELA*$_{ha}$	*DEP*$_{sa}$	*PAN*$_{fe}$	*WOR*$_{fe}$	mean
cl	26.6	18.4	42.0	06.8	08.5	20.5
br	13.7	10.9	42.9	04.8	06.5	15.8
pi	12.9	09.5	45.1	05.2	11.5	17.0
wh	10.0	09.1	49.7	04.8	08.5	16.4

Table 2: Perception accuracy (in %) of HOT$_{an}$, ELA$_{ha}$, DEP$_{sa}$, PAN$_{fe}$, and WOR$_{fe}$ (cf. Figure 2), in clean (cl) and noisy background: brown (br), pink (pi), white (wh) at -1 dB SNR. Mean accuracy is given; each row gives results for 30 instances.

%	*HOT*$_{an}$	*ELA*$_{ha}$	*DEP*$_{sa}$	*PAN*$_{fe}$	*WOR*$_{fe}$	*COL*$_{an}$
cl	63.1	51.7	92.3	26.1	31.9	93.9
br	36.3	41.4	112.9	26.3	32.4	109.1
pi	44.1	37.9	138.1	26.7	27.7	107.2
wh	36.3	34.8	125.1	25.6	28.8	117.7

Table 3: Sum of columns (in %) 'perceived as' for the 'real' emotions: HOT$_{an}$, ELA$_{ha}$, DEP$_{sa}$, PAN$_{fe}$, WOR$_{fe}$; the 'distractor' COL$_{an}$ (cf. Figure 2), in clean (cl) and -1 dB SNR background: brown (br), pink (pi), white (wh); each row encodes 30 instances.

The results for clean conditions show that the emotional state most accurately perceived is DEP$_{sa}$ (42.0%), followed by HOT$_{an}$ (26.6%), and ELA$_{ha}$ (18.4%); worse recognised were WOR$_{fe}$ (08.5%) and PAN$_{fe}$ (06.8%). HOT$_{an}$ was mainly confused with COL$_{an}$, ELA$_{ha}$ with PLE$_{ha}$, and DES$_{sa}$ with DEP$_{sa}$ (cf. Table 1), suggesting that listeners discriminate better between two different emotions than between two arousal levels of the same emotion. The 'distractors' DIS and SUR have been rarely chosen (less than 5.5 %). Confusion between different emotions within the same arousal level took mostly place between HOT$_{an}$ vs ELA$_{ha}$ (high arousal) and WOR$_{fe}$ vs COL$_{an}$ (low arousal); this can be explained by the acoustic similarities between them. In Figure 5, the *Chroma* [4] representation of emotional singing performed by a female singer (soprano) displays that HOT$_{an}$ and DEP$_{sa}$ are expressed differently. HOT$_{an}$, as shown in acted speech [13], is expressed through articulated prosody, acoustically characterised by a strong decay in amplitude and lower slope declinations, which is displayed by a richer spectrum on partials with less differences between the energy of low and high frequencies. DEP$_{sa}$, on the contrary, is expressed through sustained amplitude for each note, which concentrates more energy in F0 and less in higher harmonics. ELA$_{ha}$ presents a spectrum and articulation at mid point between the previous ones.

As expected (apart from a rare exceptions in the perception of female voices at -0.5 dB SNR in brown noise), listeners' accuracy decreases with the increment of noise (cf. Figure 4), i. e., higher SNR (-1 dB and -0.5 dB) yielded lower accuracy. By evaluating the perception of emotion in clean and -1 dB SNR conditions, (cf. Table 2), HOT$_{an}$ and ELA$_{ha}$ were affected most by noise, DEP$_{sa}$ less, WOR$_{fe}$ and PAN$_{fe}$ were perceived similarly to clean background. The three noises affected perception in a similar way: brown slightly more (15.8 % mean accuracy), pink less (17.0 % mean accuracy). Yet, the higher level of accuracy in pink and white noises is due to an improvement—caused by an increment in the confusion towards low aroused emotions—in the accuracy of DEP$_{sa}$, rather than to a lower detriment in the overall accuracy. This phenomenon relates to an acoustic 'flattening' by the noise of the characteristics typical of each emotion, causing perception as sustained, with lower energy, and attenuated articulation, i. e., similarly to low aroused emotions. Indeed, the *chromogram* for HOT$_{an}$, ELA$_{ha}$, and DEP$_{sa}$, masked by pink noise at -1 dB (cf. Figure 6), displays comparable acoustic representation for the three emotions.

To evaluate such phenomena, for each confusion matrix—obtained by the perception in clean and -1 dB SNR conditions—the sum of the columns has been computed, by that counting for each emotion all the responses 'identified as' (cf. 'total' in Table 1). Confirming previous findings [25], the confusion in background noise mostly increases for the low aroused emotions DEP$_{sa}$ and COL$_{an}$, and decreases for the high aroused HOT$_{an}$ and ELA$_{ha}$ (cf. Table 3). No meaningful differences are displayed for the other emotions across conditions.

5. AUTOMATIC RECOGNITION

5.1 Methods

We employed state-of-the-art methods for emotion recognition of vocal cues by applying a Support Vector Machine (SVM) classifier with linear kernel, from the open–source toolkit LIBLINEAR [9], and the `ComParE` 2013 challenge features set [36], extracted with OPENSMILE [8]. Since our goal is to evaluate how background noises may affect the classification performance in general, only state-of-the-art methods for automatic recognition of emotion in the operatic voice [7] have been taken into account.

For speaker independence, we split the 390 instances into three sets (A, B, and C), considering for each 130 in-

[4] Chroma features have been extracted by OPENSMILE [8].

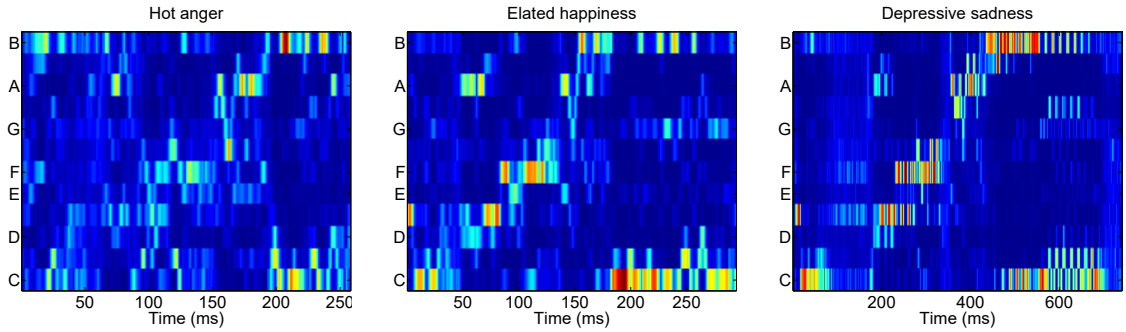

Figure 5: Chroma representation of the instances expressing: Hot_{an}, ELA_{ha}, and DEP_{sa} (from left to right); sung by one of the soprano. The y axis gives the C natural scale; the x axis the time in miliseconds. Dark blue indicates the lower level of energy, red the higher.

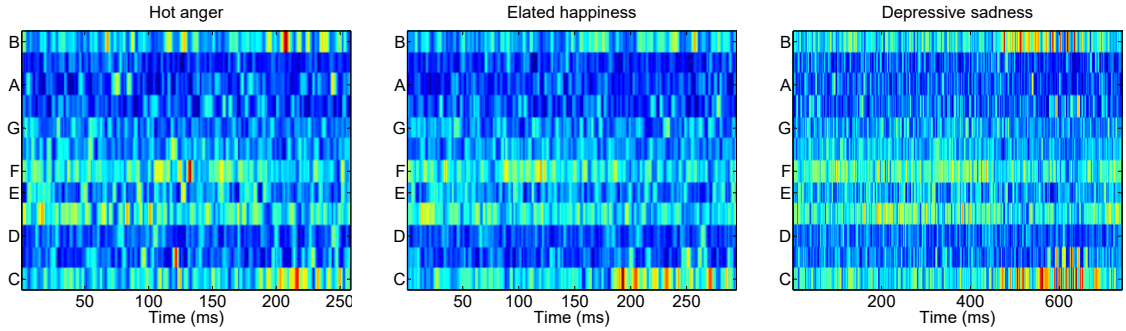

Figure 6: Chroma representation of the instances given in Figure 5 masked by pink noise at -1 dB SNR.

stances sung by two different singers (one female and one male), and performing the experiments in two phases—development and test. For the development phase we considered one set as training (e. g., A), and another as test (e. g., B); 30 levels of complexity (from 2^{30} to 2^{0}) have been tested to optimise the SVM performance. In the test phase, we merged the sets A and B for training and considered the set C as test; the complexity which achieved best results in the development phase was taken into account as optimisation parameter for the SVM. This procedure was carried out with the six possible permutations between the three sets, and the results were averaged.

We performed binary classification on five classes, i. e., each class was recognised against the other four. In the training phase, the minority class was upsampled to match the sample size of the remaining classes together; for each noise, all the SNR were considered together. We employed the whole `ComParE` 2013 features set [36], encompassing 6374 acoustic features in total: 64 low-level descriptors—LLD, and several functionals [7], in four subsets: mel-frequency cepstral coefficients—mfcc (1,400 features), spectrum (4,300), prosody (183), and voice quality (390).

5.2 Results and discussion

The classification of five classes (cf. Table 4) mirrors the perception findings (cf. Table 2) for all the feature sets: DEP being classified best, HOT and ELA in between, and PAN and WOR worse. The mfcc sub-set performs best, showing the highest Unweighed Average Recall (UAR), i. e., the mean average of the recall per class over the six permutations. In order to visualise these results, in Figure 7, a 2-dimensional *Non-Metrical Multi-Dimensional Scaling* (NMDS, [17]) solution is given. It shows a non-metrical visual representation of the optimal distances between the evaluated categories. DEP, since best recognised—thus classified as different—is more distant to the other classes in all the emotional constellations.

%	HOT	ELA	DEP	PAN	WOR	UAR
`ComParE`	26.3	28.2	77.6	07.0	18.6	31.5
mfcc	34.6	30.1	71.8	07.7	26.3	34.1
spec	26.9	25.0	82.0	03.8	12.2	30.0
prosody	17.3	22.4	48.1	08.3	21.1	23.5
vq	34.6	27.6	53.2	11.5	14.7	28.3

Table 4: Test classification accuracy and Unweighed Average Recall (UAR) in % for the 'real' emotions (HOT, ELA, DEP, PAN, WOR, cf. Figure 2), considering the four conditions—clean and the three noises—together, for each feature set: `ComParE`, mfcc, spectrum (spec), prosody, and voice quality (vq).

%	`ComParE`	mfcc	spec	prosody	vq	mean
cl	35.0	40.0	36.6	23.3	25.0	32.0
br	31.6	35.4	30.8	23.7	28.7	30.0
pi	32.0	36.2	28.3	22.9	25.4	29.0
wh	30.0	29.1	29.1	23.7	31.6	28.7

Table 5: UAR for test in % for each feature set (cf. caption of Table 4), in each condition: clean (cl), brown (br), pink (pi), white (wh). In noisy background the 4 SNR are considered together.

The feature set with the best performance (mfcc, cf. Table 4) displays an arousal related pattern, the high aroused emotions (HOT, ELA, PAN), clustered together, the low aroused (WOR, DEP) more distant. This may relate to the level of energy: higher in the former, lower in the latter (cf. Figure 5). The decline in UAR goes together with the condensation of the emotions in the 2-dim space, as

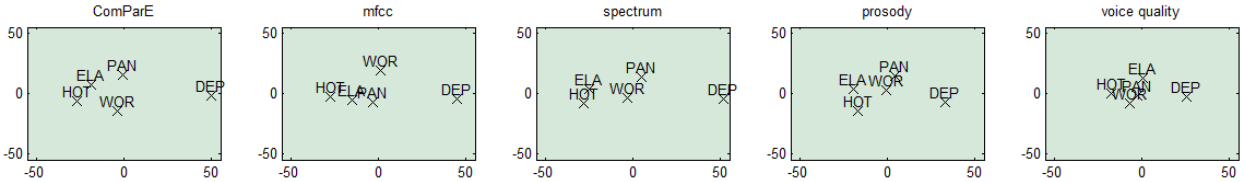

Figure 7: 2-dim NMDS solution for the five feature sets in the classification of the five 'real' emotions, considering the 390 instances (cf. caption of Table 4). Kruskal's stress is given as a measure of fit for 2-dim and 1-dim solution respectively: `ComParE` (6.3e-07, 4.0e-05); mfcc (3.4e-07, 0.1); spectrum (1.3e-17, 7.5e-16), prosody (6.9e-07, 9.2e-07); voice quality (1.3e-16, 4.8e-05).

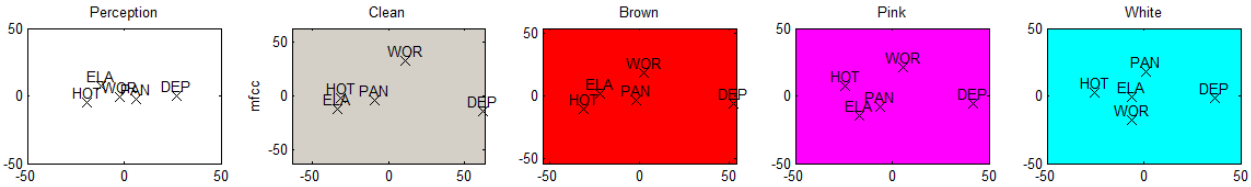

Figure 8: 2-dim NMDS solution for listeners' perception (in clean condition) and classification (in clean and noisy conditions) with mfcc features, of the five 'real' emotions. Kruskal's stress is given for 2-dim and 1-dim solution respectively: Perception (4.3e-17, 6.6e-07); Clean (1.2e-16, 7.3e-07); Brown (1.3e-16, 0.02), Pink (3.9e-07, 1.3e-05); White (6.9e-07, 3.0e-04).

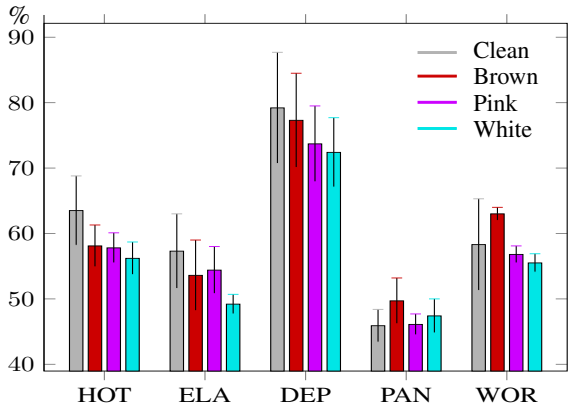

Figure 9: UAR and std in % for binary classification, i. e., each emotion against the other four, in the four conditions (cf. caption of Figure 8), for the mfcc sub-set.

prominently shown for the voice quality sub-set (cf. Figure 7). As for listeners' perception (cf. Table 2), for mfcc, `ComParE`, and to some extent for spectral features, highest accuracy was achieved in clean condition, medium in pink and brown, lower in white (cf. Table 5). Prosodic and voice quality features performed worst, which relates to both the considered musical instances and to the operatic technique. On the one hand, the sung melodic contour is the same for all utterances and emotions; thus, there are no degrees of freedom for pitch leftover for the marking of emotions. On the other hand, opera singing is characterised by the 'projection' of the voice—a high control of articulation (and by that, mfcc configurations), and a weak use of different voice qualities when expressing emotions, in contrast, for instance, to modern actors or pop singers.

In Figure 8, an NMDS visualisation for perception (in clean condition) and mfcc classification (in the different backgrounds) is given. The confusion in the perceptual constellation relates mainly to the low accuracies achieved in the listening test, which is given mostly by the use of 'distractors'. As shown in Table 5, classification in clean background yields the highest UAR, which is visually mirrored by the arousal-related pattern previously described, i. e., high aroused emotions clustered together, low aroused distant (DEP more, WOR less); this is more or less preserved for brown and pink noise but not for white noise with lowest UAR, cf. Table 5.

The binary classification (cf. Figure 9) confirms again the perceptual findings (cf. Table 2): DEP best recognised, HOT and ELA at a medium level, PAN worse. WOR is better classified than perceived, which relates to the spread of the listeners' responses motivated by the 'distractor' COL_{an}. Indeed, WOR—having the same arousal—was mainly misclassified by the listeners as COL_{an}, thus decreasing the perception accuracy of the former. White noise seems to affect binary classification more which might suggest that higher frequencies (more masked in white noise) could be more relevant for the identification of emotion in singing; lower frequencies (more masked in pink and brown noises), since related to pitch—thus to the melodic contour, which is the same for all the samples—might be less relevant for the emotional understanding in this specific study but not in general.

6. CONCLUSIONS

The present study shows that brown, pink, and white noises affect similarly the perception of emotion in operatic singing: the lower the SNR, the lower the perception. Gender seems not be an influential factor, neither for singers nor for listeners. In general, perception and classification shows analogous emotional constellations regardless the background, sadness being identified best, fear worst. The use of 'distractors' influences listeners' perception, affecting even more the accuracy of fear, an emotion which seems not to have a typical expression in singing; thus it is worse identified and easily confused. Voice quality features perform worst, mfcc best. In the former, this relates to the voice 'projection' inherent to opera (which minimise the differences between emotions), in the latter, to the relevance of energy per band to discriminate between sung emotions. Listeners' low accuracy suggests that identifying emotion in opera singing may be challenging for non trained subjects; thus, musically trained listeners will be considered in future investigations.

7. ACKNOWLEDGEMENT


This work was supported by the European Unions's Seventh Framework and Horizon 2020 Program under grant agreement No. 338164 (ERC StG iHEARu).

8. REFERENCES

[1] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, p. 614, 1996.

[2] L. Beranek, *Concert halls and opera houses: Music and acoustics*. New York, NY: Springer, 2012.

[3] M. Cognini, A. Farina, and R. Pompoli, "L'acustica dell'anfiteatro romano Arena di Verona," in *Proc. of Acoustics and Recovery of Spaces for Music*. Ferrara, Italy: ACM, 1993, pp. 105–118.

[4] E. Coutinho, K. R. Scherer, and N. Dibben, "Singing and emotion," in *The Oxford handbook of singing*, G. Welch, D. M. Howard, and J. Nix, Eds. Oxford, UK: OUP, 2014.

[5] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[6] P. Ekman, "Expression and the nature of emotion," *Approaches to Emotion*, vol. 3, pp. 19–344, 1984.

[7] F. Eyben, G. L. Salomão, J. Sundberg, K. R. Scherer, and B. W. Schuller, "Emotion in the singing voice – a deeper look at acoustic features in the light of automatic classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 1, pp. 1–9, 2015.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM MM*, 2010, pp. 1459–1462.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *JMLR*, vol. 9, pp. 1871–1874, 2008.

[10] M. García and D. V. Paschke, *A complete treatise on the art of singing*. New York, NY: Da Capo, 1975.

[11] S. Godsill, P. Rayner, and O. Cappé, "Digital audio restoration," in *Applications of digital signal processing to audio and acoustics*, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Springer, 2002, pp. 133–194.

[12] D. J. Grout and C. V. Palisca, *A history of western music*. New York, NY: Norton, 2001.

[13] M. Guzman, S. Correa, D. Muñoz, and R. Mayerhoff, "Influence on spectral energy distribution of emotional expression," *Journal of Voice*, vol. 27, pp. 129–139, 2013.

[14] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. of WASA*. Xi'an, China: IEEE, 2015, pp. 891–897.

[15] P. Howes, J. Callaghan, P. Davis, D. Kenny, and W. Thorpe, "The relationship between measured vibrato characteristics and perception in western operatic singing," *Journal of Voice*, vol. 18, pp. 216–230, 2004.

[16] S. Jansens, G. Bloothooft, and G. de Krom, "Perception and acoustics of emotions in singing," in *Proc. of Eurospeech*. Rhodes, Greece: ISCA, 1997, pp. 2155–2158.

[17] J. Kruskal and M. Wish, *Multidimensional Scaling*. London, U.K.: Sage University, 1978.

[18] P. Laukka, "Vocal expression of emotion: discrete-emotions and dimensional accounts," Ph.D. dissertation, Acta Universitatis Upsaliensis, 2004.

[19] S. R. Livingstone, K. Peck, and F. A. Russo, "Acoustic differences in the speaking and singing voice," in *Proc. of Meetings on Acoustics*, Montreal, QC, 2013, pp. 1–5.

[20] I. W. Mabbett, "Buddhism and music," *Asian Music*, vol. 25, no. 1/2, pp. 9–28, 1993.

[21] I. Mathworks, "Matlab: R2014a," *Natick*, 2014.

[22] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. of ISMIR*, Curitiba, Brazil, 2013, pp. 83–88.

[23] V. P. Morozov, "Emotional expressiveness of the singing voice: The role of macrostructural and microstructural modifications of spectra," *Logopedics Phoniatrics Vocology*, vol. 21, pp. 49–58, 1996.

[24] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Comm.*, vol. 16, pp. 369–390, 1995.

[25] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, S. Hantke, and B. Schuller, "The Perception of Emotions in Noisified Non-Sense Speech," in *Proc. of Interspeech*. Stockholm, Sweden: ISCA, 2017, pp. 3246–3250.

[26] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, S. Hantke, and B. W. Schuller, in *Proc. of DLFM*. ACM, 2017, pp. 29–36.

[27] E. Rapoport, "Emotional expression code in opera and lied singing," *JNMR*, vol. 25, pp. 109–149, 1996.

[28] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[29] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Comm.*, vol. 40, pp. 227–256, 2003.

[30] ——, "Vocal communication of emotion: A review of research paradigms," *Speech Comm.*, vol. 40, pp. 227–256, 2003.

[31] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology*, vol. 32, pp. 76–92, 2001.

[32] K. R. Scherer, J. Sundberg, B. Fantini, S. Trznadel, and F. Eyben, "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," *JASA*, vol. 142, pp. 1805–1815, 2017.

[33] K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salomão, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Computer Speech & Language*, vol. 29, pp. 218–235, 2015.

[34] K. R. Scherer, S. Trznadel, B. Fantini, and J. Sundberg, "Recognizing emotions in the singing voice," *Psychomusicology: Music, Mind & Brain*, vol. 27, pp. 244–255, 2017.

[35] B. Schuller, D. Arsić, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Proc. of Speech Prosody*. Dresden, Germany: ISCA, 2006, pp. 276–289.

[36] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of Interspeech*. Lyon, France: ISCA, 2013, pp. 148–152.

[37] H. Siegwart and K. R. Scherer, "Acoustic concomitants of emotional expression in operatic singing: The case of Lucia in Ardi gli incensi," *Journal of Voice*, vol. 9, pp. 249–260, 1995.

[38] J. Stark, *Bel canto: A history of vocal pedagogy*. London, U.K.: University of Toronto Press, 2003.

[39] S. Trehub, A. Unyk, and L. Trainor, "Adults identify infant-directed music across cultures," *Infant Behavior and Development*, vol. 16, pp. 193–211, 1993.

[40] B. Zhang, G. Essl, and E. M. Provost, "Recognizing emotion from singing and speaking using shared models," in *Proc. of ACII*. IEEE, 2015, pp. 139–145.