The Perception of Emotion in the Singing Voice

The Understanding of Music Mood for Music Organisation

Emilia Parada-Cabaleiro
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Chair of Embedded Intelligence for Health Care
and Wellbeing, Augsburg University, Germany
emilia.paradacabaleiro@informatik.uni-augsburg.de
Emilia.ParadaCabaleiro@uni-passau.de

Alice Baird
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Chair of Embedded Intelligence for Health Care
and Wellbeing, Augsburg University, Germany
Alice.Baird@uni-passau.de

Anton Batliner
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Chair of Embedded Intelligence for Health Care
and Wellbeing, Augsburg University, Germany
Anton.Batliner@uni-passau.de

Nicholas Cummins
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
Chair of Embedded Intelligence for Health Care
and Wellbeing, Augsburg University, Germany
nicholas.cummins@ieee.org

Simone Hantke
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
MISP Group, MKK,
Technische Universität München, Germany
Chair of Embedded Intelligence for Health Care
and Wellbeing, Augsburg University, Germany
Simone.Hantke@uni-passau.de

Björn W. Schuller
GLAM – Group on Language, Audio & Music,
Imperial College London, UK
Chair of Embedded Intelligence for Health Care
and Wellbeing, Augsburg University, Germany
Chair of Complex & Intelligent Systems,
Universität Passau, Germany
schuller@ieee.org

ABSTRACT

With the increased usage of internet based services and the mass of digital content now available online, the organisation of such content has become a major topic of interest both commercially and within academic research. The addition of emotional understanding for the content is a relevant parameter not only for music classification within digital libraries but also for improving users experiences, via services including automated music recommendation. Despite the singing voice being well–known for the natural communication of emotion, it is still unclear which specific musical characteristics of this signal are involved such affective expressions. The presented study investigates which musical parameters of singing relate to the emotional content, by evaluating the perception of emotion in electronically manipulated *a cappella* audio samples. A group of 24 individuals participated in a perception test evaluating the emotional dimensions of arousal and valence of 104 sung instances. Key results presented indicate that the rhythmic-melodic contour is potentially related to the perception of arousal whereas musical syntax and tempo can alter the perception of valence.

CCS CONCEPTS

•**Applied computing** → *Digital signal processing; Digital libraries and archives;*

KEYWORDS

Music mood, Perception of emotion, A cappella singing, Digital signal processing.

ACM Reference format:
Emilia Parada-Cabaleiro, Alice Baird, Anton Batliner, Nicholas Cummins, Simone Hantke, and Björn W. Schuller. 2017. The Perception of Emotion in the Singing Voice . In *Proceedings of 4th International Workshop on Digital Libraries for Musicology, Shanghai, China, October 28, 2017 (DLFM'17),* 8 pages.
DOI: 10.1145/3144749.3144756

1 INTRODUCTION

The unprecedented growth in the popularity and availability of smart devices makes intelligent solutions for efficient information retrieval more important than ever before. Music is an everyday accompaniment for user experiences, and its power to evoke emotional states is well established [25]. For this reason Music Information Retrieval (MIR) systems and Music Digital Libraries (MDL) have often focused on emotional content [20, 45].

Further, music is used not only by individuals or small user groups enjoying private and intimate moments, but also in public spaces to create a desired atmospheres. The purpose of such public sonic environments can often be used to influence behaviour in locations, such as restaurants or shopping centres, for example, inducing consumerism [4, 38]. Considering these factors, there is an increasing interest in the understanding of musical mood labels [18] and music emotion recognition [44, 45].

Music and speech are communication channels well known for their ability to express and transmit emotions [25, 39]. The emotional power of both has been the object of theorisation since the time of the ancient Greek and Roman philosophers – most notably, Pythagoras and Cicero [5, 6]. The singing voice, as a combination of both music and speech channels, is therefore an intensely emotive communication medium. Indeed, across all cultures, the singing voice is a natural and spontaneous way for humans to express a variety of feelings. Some of these, amongst others, are parental love (lullaby), social identity (work songs), and mystic connections (religious chant).

In most cases, the singing voice is the focal element of folk music (traditional and urban [3]), and vocal music is by far the most popular music genre, being streamed on services including *Spotify*, *YouTube*, and *SoundCloud*. Indeed, the top 200 streamed songs on Spotify currently contain singing voices[1]. Furthermore, the psychological impact of singing voices has also been shown in several areas, such as therapy [50] or pedagogy [55] among others.

Human perception has also been used extensively to evaluate and identify which acoustic features relate to emotion [22, 47, 53]. Specifically, in order to evaluate exactly which features relate to the vocal signal (and not to the instrumental accompaniment), several studies have examined the acoustic characteristics of emotionally labelled *a cappella* singing samples [42, 47]. Furthermore, the acoustic characteristics of music have also been evaluated in a range of automated music emotion classification systems [44, 45].

Nevertheless, it is still unclear which *musical features* of the singing voice have a capacity for evoking emotion. In perceptual studies of emotion in speech, electronic techniques have been applied in order to mask specific acoustic parameters of the vocal signal [23, 40]. However, to the best of our knowledge, such techniques have not been employed in the evaluation of emotions in the singing voice. Further, the musical genre most commonly considered in singing voice emotional studies is opera [42, 47], while other genres more popular in MDL, such as folk, rock, or popular music, have been overlooked.

With this in mind, the present work applies several electronic techniques—*random-splicing*, *reversing*, and *global tempo manipulation*—to a cappella singing voice samples from the genre of Italian folk, in order to mask the specific musical parameters of *rhythmic-melodic contour*, *musical syntax*, and *tempo*. Our goal is to evaluate the extent to which these musical parameters influence the perception of emotion. Improving the understanding of the emotions evoked in the singing voice can contribute to the development of MIR systems based on affective-based metadata.

The rest of the manuscript is structured as follows: Section 2 evaluates the related work; Section 3 presents the perception study; Section 4 describes the dataset; Section 5 discusses the results, and Section 6 offers the conclusion.

2 RELATED WORK

There is a diverse and increasing range of studies that evaluate the understanding of music mood for MIR applications [20, 28]. These studies, in line with others from music psychology [13, 25, 59], focus on the exploration of the perception of emotion in music, evaluating the extent to which factors such as a listener's cultural background [20, 27], lyrics [48], or musical genre, and instrumentation [28] influence the perception of music in terms of mood.

The perception of emotion in speech [31, 39] has also been studied extensively. Evidence of acoustic [29, 34, 42] and physiological [49] connections between emotional expression in singing and speech have also been shown. A variety of electronic masking techniques, such as random-splicing or low-pass filtering, have been performed in perceptual studies with the aim of evaluating which specific acoustic parameters are involved in emotional speech [23, 40]. However, only bandpass-filtering has been applied to manipulate the a cappella singing voice for the same purpose [30].

Opera is by far the most popular music style for acoustic and perceptual studies focusing on the emotional content of a cappella singing voices. The acoustic characteristics of an operatic voice and its relation to the perception of emotion have been evaluated in [22, 47, 53]. The operatic voice has also been considered in the automatic recognition of emotion [15]. However, the perception of emotion in other singing styles has rarely been investigated. One such example is [11], which evaluated the enthusiasm of amateur singers in karaoke.

3 PERCEPTION OF EMOTION IN MUSIC AND SPEECH

There is evidence of a deep bond between emotional expressions in music and speech [7, 24, 58]. Studies which previously have evaluated listener perception of emotion in both (speech and music) mainly refer to the categorical [14] and dimensional [37] models of emotions. Despite their popularity, both of these models present several issues.

The main problem of the categorical model, with emotions as discrete classes, is that the given categories do not always have a one-to-one correspondence with specific musical and vocal expressions. In the perception of emotions in music, some categories like happiness are identified with high agreement whereas others like jealousy are not. This suggests that certain discrete emotional categories may not be commonly induced via music [25]. In the perception of emotional speech, everyday emotional experiences are usually not clearly linked to concrete sensations but are the result of the simultaneous manifestation of different feelings [9].

The dimensional models that place emotions in a continuous hyperplane demarcated by different 'dimensions' try to resolve this difficulty. The most commonly used is the bi-dimensional model [37], based on the dimensions of valence (related to hedonistic value) and arousal (related to intensity). Nevertheless, several studies that evaluate emotional speech and the perception of emotions in music consider this model insufficient to describe all the possible emotional states [12, 25].

However, given the intrinsic characteristic of music to continuously change over the time, time-continuous dimensional annotations of emotions are considered better suited for the task of musical evaluation [25]. Tools such as FeelTrace, developed for real-time simultaneous perception of arousal and valence, have been used to evaluate both music and speech [8]. Despite this, the simultaneous evaluation of two dimensions has been considered by some authors as too challenging [36], reducing consequently the accuracy of listener responses. The real-time evaluation of both dimensions independently, for instance by GTrace[2] [8], has been shown to achieve good levels of agreement for the perception of emotion in both music [43] and speech [36]. Continuous annotation requires samples of longer duration [10, 36, 43], which can impede the evaluation of specific musical parameters like musical syntax, that depend on short 'musical sentences' whose length is usually around 10 sec.

In addition, mood models specially designed to evaluate the perception of emotion, according to preferences in contemporary music listening environments, have been developed for MIR applications [19] and shown to be more suitable than the categorical and dimensional models of emotion. Nevertheless such models are not sufficient for mirroring the real preferences and requirements

[1] https://spotifycharts.com/regional (last updated 29/6/2017).

[2] https://sites.google.com/site/roddycowie/work-resources

of users when evaluating cross-culturual perception [20]. Considering all of this in the present study, we employ a bi-dimensional independent evaluation of arousal and valence, using a rating-scale with 10 distinct levels.

3.1 Perception Test

We conducted a bi-dimensional test in order to evaluate the perception of valence and arousal in relation to the specific musical parameters of rhythmic-melodic contour, musical syntax, and tempo. The dimensions have been tested independently from each other, over two different rounds. Both tasks were divided into 13 sections, with responses saved at the end of each sub-task in order to provide the opportunity for a break. The sessions lasted around 1 hour; they were performed in an acoustically treated recording studio using a browser-based interface provided through the gamified crowdsourcing platform *iHEARu-PLAY* [17].

We created identical conditions and equipment set-up for all of the participants. Listeners had the possibility to repeat each sung chunk (presented in a randomised order) and the associated annotation indefinitely. However, in order to promote similar conditions during the task, the listeners were encouraged to answer following their first impression. The test was performed by 24 listeners (age between 18 and 30 years, standard deviation 3.5 years). The participants were students or employee volunteers from the University of Passau (Germany) from different nationalities: 13 German, 4 Indian, 2 British, 2 Tunisian, 1 Iranian, 1 Russian, and 1 Spanish.

Previous research has shown that lyrics can influence the perception of emotion in music, especially when there is a contradiction with the valence dimension of the message given by lyrics and music [26]. Considering that our methodological approach refers to the dimensional model of emotions, in order to avoid a linguistic information bias, people with Italian language proficiency above basic conversation, i. e., the ability to communicate through everyday simple expressions, were excluded from the test. Finally, in order to avoid influencing their response, the listeners were not informed as to the purpose of the research.

Additionally, the listeners completed a questionnaire in regards to their musical knowledge, interest, and level of understanding from the linguistic message of each sample heard during the test. Only one of the listeners had vocal training, and their musical skills were varied from no musical education, to instrument self-training, or computer music knowledge. The musical interest of the listeners have been shown to vary a lot as well, including electronic, pop, latin, indi, rock, hip hop, classical, or Indian classical music, among others. None of the listeners declared that they could understand the linguistic content of the presented chunks during the test.

Given the heterogeneity of the responses relating to musical knowledge and interest, and considering that this sparse information did not show a correlation to patterns of cultural background, gender, or age, we decided to not take this information into account in our analysis.

4 ELECTRONICALLY MASKED DATASET OF A CAPPELLA SINGING VOICES

In the presented study we utilised a dataset of 104 sung chunks (52 for each gender), sung by six singers (three males and three females), with similar length (11.8 sec average, standard deviation 2.9 sec). From these, 26 sung chunks are presented in their original version, i. e., without artificial manipulation, and the remaining 78 have been electronically processed following three manipulation techniques: 26 random-spliced, 26 reversed, and 26 global tempo manipulated. The total duration of the dataset is 19 min and 43 sec and is freely available upon request for research purposes.

Table 1: Each singer is identified with a different ID (1-6) and their gender: female (F) and male (M). The number of original samples performed by each singer, their musical tempo, and the total duration is also given.

Singer ID	# Samples	Tempo	Duration (min)
$1-F$	2	*Allegro*	0.22
$2-F$	7	*Adagio*	1.16
$3-F$	4	*Adagio*	0.57
$4-M$	4	*Adagio*	0.37
$5-M$	5	*Allegro*	1.06
$6-M$	4	*Adagio*	0.48

4.1 Italian Folk Singing Voice Corpus

We considered six Italian folk songs; both music and lyrics were originally composed, in the musical style of *Canzone Romana*. This traditional Italian folk music genre, predominately written in tonal harmony, presents a regular musical structure, and is characterised by the utilisation of lyrics written in the *Roman* Italian dialect. Developed as a popular form of spontaneous communication, *Canzone Romana* is a music style that people find easy and comfortable to sing, even without specialist vocal training.

Each of the song has been performed by a different singer, three males and three females. From the six songs, a total of 26 sung chunks (13 for each gender) have been chosen. A description of the original data considering singer, number of sung chunks per song, tempo (the speed of a musical composition), and duration is given in Table 1. The samples used were generously donated by the composer and have no copyright restrictions. All involved in this creative exchange have given their full consent for the free use of these samples for non-commercial purpose[3].

4.2 Evaluated Musical Parameters

We evaluated three musical parameters: (i) rhythmic-melodic contour, (ii) musical syntax, and (iii) tempo. These have been chosen as they have previously been shown to be relevant for the perception of emotional content [25]. In addition, we also consider these parameters as particularly suitable to be disrupted through artificial manipulation, while keeping the naturalness of the voice intact, a property that is crucial when evaluating the perception of a cappella samples. We consider that the naturalness of the voice would indeed be affected if considering the artificial manipulation of other musical features more related to the vocal signal, such as e. g., articulation or vibrato.

The *rhythmic-melodic contour* is a pattern characterised by the changes in the pitch (ups and downs) and by the relative length

[3] The corpus is freely available upon request.

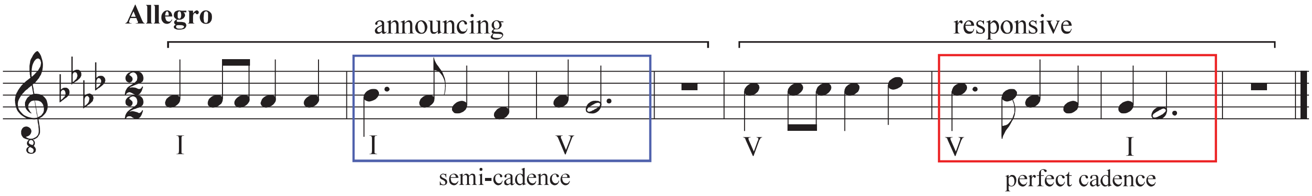

Figure 1: A transcription of a song chunk sung by singer number 5. This excerpt is an example of the use of announcing and responsive phrases typical in the *Canzone Romana* genre. The rhythmic-melodic contour of the announcing phrase is harmonically suspended, as shown by the semi-cadence I-V (highlighted in blue), which creates a sense of expectation in the listener. The responsive phrase, on the other hand, is harmonically conclusive, as shown by the perfect cadence V-I (highlighted in red), which gives the listener the acoustic sensation that the excerpt is finished. The rhythmic-melodic contour of the responsive phrase is a progression of the announcing phrase with minimal variations.

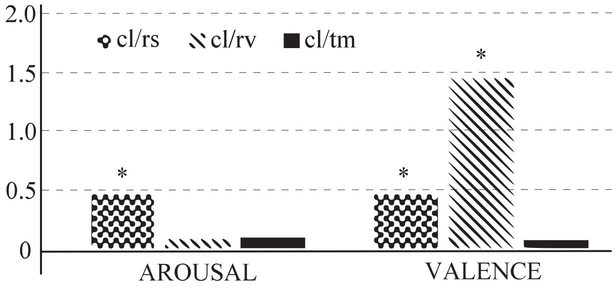

Figure 2: Mean differences between clean (cl) and manipulated signals: random-splicing (rs), reversing (rv), and global tempo manipulation (tm). Results are given for the perception of arousal and valence considering the whole dataset (104 samples encoded in each bar); starred results indicate $p < .05$ in Tukey's post hoc test.

of the notes in time. *Musical syntax* is the internal musical structure; this parameter is related to the harmonic discourse (given by the changes of tonal functions over the time-line) as well as to the musical sentences (made up of two phrases: announcing and responsive). *Tempo* is the speed of a musical composition. Two different tempos have been considered: fast tempo, known as *Allegro* (songs performed by singers 1 and 5), and slow tempo, known as *Adagio* (songs performed by singers 2, 3, 4, and 6).

A typical example of the considered repertoire presents a rhythmic-melodic contour characterised by basic and flowing rhythms and melodies. This is shown by the use of simple syncopations created by dotted quarter-notes and eight-notes, as well as by the dominance in repeated notes and step-wise motions—meaning that the melodic contour is performed by repeating a note or in conjunction with, so each note is followed by the note immediately upper or lower (cf. Figure 1). The musical syntax is articulated in sentences of two symmetric phrases (announcing and responsive), being usually harmonically suspended and conclusive respectively, and determined by a very simple and static harmonic rhythm.

In Figure 1 a sung chunk is symbolically represented and annotated to visually illustrate the evaluated musical parameters. The tempo considered in the presented example is Allegro, which generally means 120 BPM, i. e., 120 quarter-notes per minute.

4.3 Manipulation Techniques

The three signal processing techniques of random-splicing, reversing, and global tempo manipulation have been chosen to mask the musical parameters (rhythmic-melodic contour, musical syntax, and tempo). The first two techniques (random-splicing and reversing) have been extensively used in emotional speech perceptual studies [41]. The third (global tempo manipulation) has been chosen for its specific musical application, considering previous research that evaluates how the overall tempo affects the perception of expressiveness [35].

Random-splicing (*rs*) is performed through segmentation and random re-reordering; it should disrupt the rhythmic-melodic contour. *Reversing* (*rv*) can mask the original musical syntax by altering the natural flow of the musical sentence structure (announcing followed by responsive) as well as the original harmonic discourse. *Global tempo manipulation* (*tm*), performed through a variation of speed, can invert the original tempo. It should be noted that each manipulation technique has been specially chosen to modify the natural expression of each musical parameter, whilst keeping intact its essential elements. This means that the elements typical of each musical parameter are still perceptible in the manipulated samples, yet displayed in an unnatural way according the musical context.

We used the commonly available DAW Logic Pro[4]. Random-splicing was performed on each sample by manually segmenting into chunks of 0.5 sec and random re-ordering. As well as this, each sample was reversed, a process which physically reverses the original samples. Finally, global tempo manipulation was applied making those samples previously classified as Adagio 50% faster, and those classified as Allegro 25% slower. These values were chosen as they produced minimal artefacts on the samples and also were not an obvious alteration from the original sample.

5 RESULTS AND DISCUSSION

The mean values of clean (*cl*) vs manipulated results are given in Figure 2; values below the conventional threshold of $p < .05$

[4] Apple Inc., *Logic 9.1.8*, 2012.

Table 2: Mean differences between clean (cl) vs random-splicing (rs) for arousal and valence for both male and female voices. Values in bold indicate $p < .05$ in Tukey's post hoc test.

cl/*rs*	**arousal**	**valence**
female voices	0.14	**0.74**
male voices	**0.80**	0.22

Table 3: Mean differences between clean (cl) and random-splicing (rs), reversing (rv), and global tempo manipulation (tm) signals for arousal. Values in bold indicate $p < .05$ in Tukey's post hoc test.

arousal	ID1	ID2	ID3	ID4	ID5	ID6
cl/*rs*	0.19	0.23	0.16	**1.03**	**0.63**	**0.79**
cl/*rv*	0.30	0.11	0.11	0.37	0.14	**0.87**
cl/*tm*	0.06	0.06	0.16	0.30	0.46	0.12

obtained by Tukey's post hoc tests from an ANOVA are marked by a star[5]. As can be seen, the rhythmic-melodic contour (disrupted by *rs*) is related to the identification of both dimensions (arousal and valence). The effect sizes *d* for this result (*cl* vs *rs*) shows a small effect (d=0.25 for arousal, and d=0.23 for valence). Musical syntax (disrupted by *rv*) displays a stronger tendency in the identification of valence giving d=0.73 for *cl* vs *rv*. Interestingly, tempo does not seem to play a role.

As for the singer's gender, our test reveals the rhythmic-melodic contour to potentially be an important parameter (cf. Table 2), however, with different effects in each gender. For male voices, the rhythmic-melodic contour appears to be related to the identification of arousal (d=0.47), whereas for female voices, it appears related to the recognition of valence (d=0.32). We speculate that these differing effects are linked to the properties of the spectra typical of each gender. In female voices, the long-term average spectrum is usually in a higher position, the third formant being especially prominent [30]. This provides female voices the brightness necessary to push the voice over other acoustic layers, e. g., competing orchestral instruments during an opera. This brightness is present even when no other acoustic layers are considered, allowing the listeners to, better, perceive the arousal level of female voices, even in the *rs* manipulated samples.

Rhythmic-melodic contour and the perception of valence in female voices could potentially be linked to the notion that in Western cultures, women are allowed to be more expressive than men, for both positive and negative emotions [16]. With this in mind, we could consider the female samples in our corpus as being more characteristic in terms of valence than those sung by males. Therefore, *rs* manipulation, which disrupts the rhythmic-melodic contour and alters natural expressiveness, potentially has a greater effect on the female samples.

Table 4: Mean differences between clean (cl) and random-splicing (rs), reversing (rv), and global tempo manipulation (tm) signals for valence. Values in bold indicate $p < .05$ in Tukey's post hoc test.

valence	ID1	ID2	ID3	ID4	ID5	ID6
cl/*rs*	**1.02**	0.39	**1.16**	0.67	0.22	0.32
cl/*rv*	**2.41**	0.35	**1.51**	**2.30**	**1.78**	**1.58**
cl/*tm*	**1.40**	0.34	**0.87**	0.37	**0.85**	0.42

Table 5: Mean differences between clean (cl) vs random-splicing (rs) for arousal (male voices: m) and valence (female voices: f), recognised by German and non German listeners. Values in bold indicate $p < .05$ in Tukey's post hoc test.

cl/*rs*	**valence (f)**			**arousal (m)**		
	ID1	ID2	ID3	ID4	ID5	ID6
German	0.94	0.50	0.57	**0.62**	0.29	0.21
Non German	**1.05**	0.57	**1.48**	**1.33**	**0.83**	**0.91**

The differences of the means for arousal and valence, *cl* vs manipulated, for each singer, are presented in Table 3 and Table 4. The effect sizes *d* for these results vary from 0.41 (ID5-arousal for *cl*/*rs*) to 1.48 (ID1-valence for *cl*/*rv*). Given the specificity of the considered classes, these results display a more stable tendency compared to those previously described. Results indicate that musical syntax (disrupted by *rv*) is related to the identification of valence in both genders (cf. Table 4). This is because the implicit harmony contained in the melodic line has an essential value in the positive/negative musical meaning [25], and this is potentially lost when applying *rv*, which disrupts the natural harmonic sequence.

Although tempo can be associated with the expression of emotions in music [25], when considering the whole dataset, our analysis shows that tempo is not related to the perception of emotions (cf. Figure 2). However, this musical parameter has shown to be relevant in the identification of valence for Allegro samples (ID1 and ID5) (cf. Table 4), as indicated by the pronounced mean difference. This outcome supports studies showing that music performed in fast tempo is perceived as more positive, whereas a slow tempo is not related to valence [21].

5.1 Cross-cultural Evaluation

Given the small size of our listener group, especially the non-German sub-groups, we consider the cultural-dependent interpretation of the obtained results as an indicator for future work. The analysis presented in this section should therefore be considered as speculative and subject to further experimentation, with a larger cross-cultural sample size.

First, we considered two groups of listeners: German (G: 13 participants) and non German (¬G: 11 participants). Contrasting these two groups, our analysis displays that the perception of both emotional dimensions by ¬G is more affected by all the manipulation techniques than for G listeners (cf. Table 5 and Table 6). The effect sizes *d* for these results vary from 0.33 (ID4-arousal for *cl*/*rs* identified by G listeners) to 1.75 (ID1-valence for *cl*/*rv* identified by ¬G listeners).

In Table 5, the mean differences for *cl* vs *rs* are given for the evaluation of arousal in male voices and valence in female voices.

[5] In this paper, we report basics for traditional Null Hypothesis Testing (NHT) but refrain from a full-fledged NHT analysis, due to its inherent problems [56]; instead, we employ effect sizes [52].

Table 6: Mean differences between clean (cl) vs reversing (rv) and clean (cl) vs global tempo manipulation (tm) for valence, recognised by German and non German listeners. Values in bold indicate p $<$ .05 in Tukey's post hoc test.

cl/rv	ID1	ID2	ID3	ID4	ID5	ID6
German	**1.56**	0.23	**1.08**	**1.59**	**1.17**	**0.76**
Non German	**2.65**	0.55	**1.74**	**2.58**	**1.94**	**2.00**
cl/tm	ID1	ID2	ID3	ID4	ID5	ID6
German	**1.23**	0.05	0.77	0.35	0.58	0.93
Non German	**1.47**	0.23	0.92	0.38	**1.00**	0.15

Table 7: Mean differences between clean (cl) vs random-splicing (rs) for arousal (male voices: m) and valence (female voices: f), as recognised by Indian (IN), Tunesian and Iranian (TI), and British, Spanish, Russian (EU— European) listeners. Values in bold indicate p $<$ .05 in Tukey's post hoc test.

cl/rs	valence (f)			arousal (m)		
	ID1	ID2	ID3	ID4	ID5	ID6
IN	**1.18**	0.62	0.34	**1.96**	**1.78**	**1.04**
TI	1.23	0.10	**3.05**	**3.04**	1.39	**1.86**
EU	**2.37**	1.98	1.11	0.57	0.36	1.20

Table 8: Mean differences between clean (cl) vs reversing (rv) for valence, recognised by IN, TI, and EU (cf. Table 7) listeners. Values in bold indicate p $<$ .05 in Tukey's post hoc test.

cl/rv	ID1	ID2	ID3	ID4	ID5	ID6	mean
IN	**3.06**	0.50	0.80	**3.60**	**3.04**	**2.67**	2.27
TI	**2.56**	0.24	**2.23**	**2.58**	0.74	**1.75**	1.68
EU	**2.00**	0.92	**1.51**	0.75	1.07	**2.23**	1.41

Table 9: Mean differences between clean (cl) vs global tempo manipulation (tm) for valence, recognized by IN, TI, and EU (cf. Table 7) listeners. Values in bold indicate p $<$ .05 in Tukey's post hoc test.

cl/tm	ID1	ID2	ID3	ID4	ID5	ID6	mean
IN	**1.28**	0.35	**2.32**	0.01	**1.84**	0.68	1.06
TI	**2.08**	0.40	0.39	**1.28**	0.79	0.60	0.92
EU	**2.82**	0.32	0.60	0.87	1.03	0.01	0.94

The perception of G listeners is clearly less affected in both cases. The perception of valence in samples manipulated by *rv* is affected for both groups (cf. Table 6). Again, this influence is more prominent for ¬G but yields a marked mean difference in the recognition of both samples by the G listeners as well. Considering samples in Allegro tempo, the perception of valence is also affected by *tm* (cf. Table 6), having a predominant effect over the ¬G listeners in comparison to the G group.

It has previously been shown that a listener's musical-cultural background can influence their perception of emotion in music [51]. Indeed, even though music information research is often characterised as being western-centric [46], the influence of a listener's cultural-musical background in their perception of emotion in music is a relevant topic [1, 20, 27]. Therefore, in order to evaluate if there is a relationship between culture and listeners perception of musical parameters in relation to emotional content, we analysed the perception of ¬G listeners considering the musical convention typical of their cultural background.

Furthermore, psychoanalytical models for music-therapy [2] support the notion that even in today's global society, music genres typical of a region play a key role in forming the *cultural musical-identity* of persons born and raised in that region. From this perspective, we can speculate that a listener's musical-identity is still connected to their native culture influencing how they perceive and respond to music, even for listeners exposed to and immersed in western-dominated musical environments.

We split the ¬G listeners into three sub-groups: Indian (IN), Tunisian and Iranian (TI), and British, Spanish, and Russian (EU, i. e., European). Even though German listeners belong to the EU sub-group, given the unbalanced number of listeners obtained by the combination of German with British, Spanish and Russian, combining them would have caused further imbalance. Therefore, in our evaluation, when comparing to the ¬G group, we considered G listeners as an independent and culturally characterised group, thus leaving it aside when evaluating the sub-group cross-cultural analysis, i. e., the comparison between IN, TI, and EU sub-groups.

The three subgroups of listeners can be characterised, from an ethno-musicological point of view, by three different musical traditions. The IN group relates to the Indian musical tradition, whereas the TI group relates to the *maqam* phenomenon, i. e., an improvisatory process typical of countries from North Africa, the Near East, and Central Asia (known as *dastgah* in Iran and as *maqam* in the Tunesia [54]). Although in melodic therms, these two musical traditions are linked to microtonal scales, aspects such as the improvisatory models typical of each tradition make them unique [32]. Finally, the EU group is the only one to relate to the Western musical tradition. This group, even considering in-group differences, such as Russian musical identity [33], is generally dominated by tonal scales, melodic system which clearly differ from that used by IN and TI groups.

Our results show that arousal perception of male voices is influenced by *rs* for TI and especially for IN, but not for EU listeners (cf. Table 7). Rhythmic-melodic contour is also related to valence in female voices but to a lesser extent than for arousal. The cultural sub-groups show similar results, finding difficulties in the perception of only one singer (cf. Table 7). The perception of valence is affected when musical syntax is disrupted for all the sub-groups in the following order: IN (more affected), TI (mildly), and EU (less), as shown by the mean values (cf. Table 8). This tendency does not show up for the mean values of tempo: manipulation of Allegro samples influences the perception of valence for all the sub-groups (cf. Table 9) roughly to the same extent.

In summary, our results indicate that, in general, arousal perception in a capella singing does not depend on musical parameters, with the exception of the rhythmic-melodic contour in male voices. The perception of valence, on the other hand, is related to all the musical parameters, evaluated; especially in the IN and TI subgroups.

Musical syntax has been shown to be an essential parameter in the identification of valence, most prominently in the IN and TI subgroups and markedly in the EU subgroup.

As none of the listeners had previous knowledge of the Italian language, we speculate that these differences relate to differing cultural musical backgrounds. The G and EU listeners potentially found it easier to evaluate the affective content of the Italian folk music, even in the manipulated conditions. This could be due, in part, to these cultures sharing a similar musical background based on diatonic musical scales; the IN and TI listeners on the other hand have a musical background based on quarter-tone scales. The presented findings support previously presented outcomes in the evaluation of cross-cultural perception of emotion in music [20, 27], which also indicate that a listener's cultural background influences their perception of emotion in music.

5.2 Implications of the presented work for MDL and MIR research and applications

As culture influences the perception of mood in music [19], it has been suggested that there is a need for a redefinition of existing musical mood models [20]. For example, the five cluster model proposed for previous Music Information Retrieval Evaluation eXchange (MIREX) Audio Mood Classification (AMC) tasks [18, 19], is no longer considered suitable for cross-cultural evaluations [20]. The results presented in Section 5.1 indicate that specific musical parameters, including rhythmic-melodic contour, musical syntax, and tempo, also play a role in the perception of emotion in music. Therefore we speculate that the identification of these parameters will contribute to the development of musical features suitable for Music Information Retrieval (MIR) and Musical Digital Libraries (MDL) systems which use mood as a metadata label.

Moreover, given the acoustic similarities between speech and singing [58], it is not surprising that state-of-the-art methodologies common in speech-based emotion recognition have been successfully exploited to automatically classify emotions in a cappella singing [15]. However, the role of musical features has yet to be considered in such paradigms. Given the results presented in Section 5.1 suggest that musical parameters also have a role in cross-cultural musical mood perception of the singing voice, we speculate the inclusion of musical features will aid the development of accurate and robust automated music mood recognition systems [44, 45, 57]. This could have an important impacts on the advancement of MIR and MDL systems specifically optimised for dealing with differing musical traditions.

6 CONCLUSIONS

In this contribution, we evaluated the extent to which the specific musical parameters of rhythmic-melodic contour, musical syntax and tempo affect perception of emotion in singing. Such an understanding is an essential step in the development of MIR systems that relate to areas of perception and cognition. The presented outcomes are key steps towards developing affective MDL. Such developments will consider not only acoustic and linguistic features, but also the role of musical parameters in the development of robust and user-centred systems for music mood classification.

The presented results indicate that the rhythmic-melodic contour is linked more to arousal perception of male voices and less so to valence in females voices. The recognition of valence was shown to be specially related to musical syntax and Allegro tempo (fast speed). Although these aspects generalise across all the listeners, the Europeans listeners (Germans, British, Spanish, and Russians), displayed greater similarities in the perception of arousal and valence when comparing between the clean and manipulated samples.

In future work, we will further evaluate the relationship between singer's gender and the perception of arousal and valence. We will also develop further perception tests considering larger groups of listeners, coming from a variety of musical traditions, for a deeper understanding of the influence aspects such as musical familiarity and novelty, or cultural background have, on the perception of emotion in the singing voice.

Since specific emotional models have been developed in the context of MIR and MDL for classification and annotation tasks, we will utilise the initial results presented to perform further perception test, taking a methodological approach that is currently more influential in MIR, as e. g., MIREX 5-cluster model [19]. This and future findings will contribute deeper insights into emotional perception of the singing voice with the aim of developing advanced MIR systems designed around emotional content.

7 ACKNOWLEDGEMENT

This work was supported by the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA). We would also like to thank all involved in the creation of the folk Italian dataset used in the presented research, as well as all the participants for their time and effort.

REFERENCES

[1] L-L. Balkwill, W. F. Thompson, and R. Matsunaga. 2004. Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. *Japanese Psychological Research* 46, 4 (2004), 337–349.

[2] R. O. Benenzon. 1991. *Teoría de la musicoterapia: Aportes al conocimiento del contexto no-verbal.* Editorial Mandala, Madrid, Spain.

[3] P. V. Bohlman. 1988. *The study of folk music in the modern world.* Indiana University Press, Bloomington, IN, USA.

[4] D. Buxton. 1983. Rock music, the star-system and the rise of consumerism. *Telos* 1983, 57 (1983), 93–106.

[5] M. T. Cicero, J. M. May, and J. Wisse. 2001. *Cicero: On the ideal orator (De Oratore).* Oxford University Press, Oxford, UK.

[6] G. Comotti. 1991. *Music in Greek and Roman culture.* Johns Hopkins University Press, Baltimore, MD, USA.

[7] E. Coutinho, J. Deng, and B. W. Schuller. 2014. Transfer learning emotion manifestation across music and speech. In *International Joint Conference on Neural Networks (IJCNN).* IEEE, Beijing, P. R. China, 3592–3598.

[8] R. Cowie, C. Cox, J-C. Martin, A. Batliner, D. Heylen, and K. Karpouzis. 2011. Issues in data labelling. In *Emotion-oriented systems: The humaine handbook*, P. Petta, C. Pelachaud, and R. Cowie (Eds.). Springer, Berlin, Germany, 213–241.

[9] R. Cowie, E. Douglas-Cowie, and C. Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks* 18 (2005), 371–388.

[10] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 2000. FEELTRACE: An instrument for recording perceived emotion in real time. In *Proceedings of the Tutorial and Research Workshop (ITRW) on Speech and Emotion.* ISCA, Newcastle, UK, 19–24.

[11] R. Daido, S. Hahm, M. Ito, S. Makino, and A. Ito. 2011. A System for evaluating singing enthusiasm for karaoke. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Miami, FL, USA, 31–36.

[12] L. Devillers, L. Vidrascu, and L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18 (2005), 407–422.

[13] H. Egermann, N. Fernando, L. Chuen, and S. McAdams. 2014. Music induces universal emotion-related psychophysiological responses: comparing Canadian listeners to Congolese Pygmies. *Frontiers in psychology* 5 (2014), 1–9.

[14] P. Ekman. 1984. Expression and the nature of emotion. *Approaches to Emotion* 3 (1984), 19–344.

[15] F. Eyben, G. L. Salomão, J. Sundberg, K. R. Scherer, and B. W. Schuller. 2015. 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* 1 (2015), 1–9.

[16] A. H. Fischer. 1993. Sex differences in emotionality: Fact or stereotype? *Feminism & Psychology* 3 (1993), 303–318.

[17] S. Hantke, F. Eyben, T. Appel, and B. W. Schuller. 2015. iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing. In *Proceedings of the 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA) held in conjunction with the 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Xi'an, P. R. China, 891–897.

[18] X. Hu and J. S. Downie. 2007. Exploring mood metadata: Relationships with genre, artist and Usage metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Vienna, Austria, 67–72.

[19] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. 2008. The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Philadelphia, PA, USA, 462–467.

[20] X. Hu and J. H. Lee. 2016. Towards global music digital libraries: A cross-cultural comparison on the mood of Chinese music. *Journal of Documentation* 72, 5 (2016), 858–877.

[21] G. Ilie and W. F. Thompson. 2006. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception: An Interdisciplinary Journal* 23 (2006), 319–330.

[22] S. Jansens, G. Bloothooft, and G. de Krom. 1997. Perception and acoustics of emotions in singing. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*. ISCA, Rhodes, Greece, 2155–2158.

[23] T. Johnstone and K. R. Scherer. 2000. Vocal communication of emotion. In *Handbook of emotion*, M. Lewis and J. M. Haviland-Jones (Eds.). Vol. 2. Guilford, New York, NY, USA, 220–235.

[24] P. N. Juslin and P. Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* 129 (2003), 770.

[25] P. N. Juslin and P. Laukka. 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research* 33 (2004), 217–238.

[26] M. Kazuma. 2009. The influence of the meaning of lyrics on the expressed emotion of music valence. *Systematic Musicology* (2009), 53–58.

[27] K. Kosta, Y. Song, G. Fazekas, and M. B. Sandler. 2013. A Study of Cultural Dependence of Perceived Mood in Greek Music.. In *Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Curitiba, PR, Brazil, 317–322.

[28] J. H. Lee, T. Hill, and L. Work. 2012. What does music mood mean for real users?. In *Proceedings of the 2012 iConference*. ACM, Toronto, ON, Canada, 112–119.

[29] S. R. Livingstone, K. Peck, and F. A. Russo. 2013. Acoustic differences in the speaking and singing voice. *Proceedings of Meetings on Acoustics* 19, 1 (2013), 035080.

[30] V. P. Morozov. 1996. Emotional expressiveness of the Singing Voice: The role of macrostructural and microstructural modifications of spectra. *Logopedics Phoniatrics Vocology* 21 (1996), 49–58.

[31] I. R. Murray and J. L. Arnott. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America* 93 (1993), 1097–1108.

[32] B. Nettl. 1974. Thoughts on improvisation: A comparative approach. *The Musical Quarterly* 60, 1 (1974), 1–19.

[33] A. G. Piotrowska. 2016. The place of Russian music on the multicultural map of Europe. *Muzikologija* 21 (2016), 109–122.

[34] E. Rapoport. 1996. Emotional expression code in opera and lied singing. *Journal of New Music Research* 25 (1996), 109–149.

[35] B. H. Repp. 1995. Quantitative effects of global tempo on expressive timing in music performance: Some perceptual evidence. *Music Perception: An Interdisciplinary Journal* 13 (1995), 39–57.

[36] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, Shanghai, P. R. China, 1–8.

[37] J. A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39 (1980), 1161–1178.

[38] K. Ryu and S. S. Jang. 2007. The effect of environmental perceptions on behavioral intentions through emotions: The case of upscale restaurants. *Journal of Hospitality & Tourism Research* 31, 1 (2007), 56–72.

[39] K. R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40 (2003), 227–256.

[40] K. R. Scherer, S. Feldstein, R. N. Bond, and R. Rosenthal. 1985. Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research* 14 (1985), 409–425.

[41] K. R. Scherer, D. R. Ladd, and K. E. Silverman. 1984. Vocal cues to speaker affect: Testing two models. *Journal of Language and Social Psychology* 5 (1984), 1346–1356.

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

[43] E. Schubert. 1999. Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology* 51 (1999), 154–165.

[44] B. W. Schuller, J. Dorfner, and G. Rigoll. 2010. Determination of nonprototypical valence and arousal in popular music: features and performances. *EURASIP Journal on Audio, Speech, and Music Processing* 1 (2010), 1–19.

[45] B. W. Schuller, F. Weninger, and J. Dorfner. 2011. Multi-Modal Non-Prototypical Music Mood Analysis in Continuous Space: Reliability and Performances.. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Miami, FL, USA, 759–764.

[46] X. Serra. 2011. A multicultural approach in music information research. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Miami, FL, USA, 151–156.

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

[48] A. Singhi and D. G. Brown. 2014. On Cultural, Textual and Experiential Aspects of Music Mood.. In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Taipei, Taiwan, 3–8.

[49] H. Spencer. 2015. The origin and function of music. In *The routledge reader on the sociology of music*, J. Shepherd and K. Devine (Eds.). Routledge, London, UK, 210–238.

[50] J. C. Stemple and E. R. Hapner. 2014. *Voice therapy: Clinical case studies*. Plural Publishing, San Diego, CA, USA.

[51] C. J. Stevens. 2012. Music perception and cognition: A review of recent cross-cultural research. *Topics in Cognitive Science* 4, 4 (2012), 653–667.

[52] G. M. Sullivan and R. Feinn. 2012. Using effect size – or why the p value is not enough. *Journal of Graduate Medical Education* 4 (2012), 279–282.

[53] J. Sundberg, J. Iwarsson, and H. Hagegård. 1995. A singer's expression of emotions in sung performance. In *Vocal fold physiology: Voice quality control*, O. Fujimura (Ed.). Singular Pub. Group, San Diego, CA, USA, 217–229.

[54] H. H. Touma. 1971. The maqam phenomenon: An improvisation technique in the music of the Middle East. *Ethnomusicology* 15, 1 (1971), 38–48.

[55] C. Ware. 1998. *Basics of vocal pedagogy: The foundations and process of singing*. McGraw-Hill, Boston, MA, USA.

[56] R. L. Wasserstein and N. A. Lazar. 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* 70 (2016), 129–133.

[57] F. Weninger, F. Eyben, and B. W. Schuller. 2014. On-line continuous-time music mood regression with deep recurrent neural networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Florence, Italy, 5412–5416.

[58] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer. 2013. On the acoustics of emotion in audio: What speech, music and sound have in common. *Frontiers in Psychology, section Emotion Science, Special Issue on Expression of emotion in music and vocal communication* 4 (2013), 1–12.

[59] M. Zentner, D. Grandjean, and K. R. Scherer. 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion* 8 (2008), 494–521.