Audio Features Affected by Music Expressiveness

Article · May 2016

3 authors:

Alberto Introini
University of Milan
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Giorgio Presti
University of Milan
6 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Giuseppe Boccignone
University of Milan
64 PUBLICATIONS **743** CITATIONS

SEE PROFILE

Audio Features Affected by Music Expressiveness

Experimental Setup and Preliminary Results on Tuba Players

Alberto Introini
Dipartimento di Informatica
Università degli Studi di Milano
39, Via Comelico, Milano, Italy

Giorgio Presti
Dipartimento di Informatica
Università degli Studi di Milano
39, Via Comelico, Milano, Italy
giorgio.presti@unimi.it

Giuseppe Boccignone
PHuSe Lab
Dipartimento di Informatica
Università degli Studi di Milano
39, Via Comelico, Milano, Italy

ABSTRACT

Within a Music Information Retrieval perspective, the goal of the study presented here is to investigate the impact on sound features of the musician's *affective intention*, namely when trying to intentionally convey emotional contents via expressiveness. A preliminary experiment has been performed involving 10 tuba players. The recordings have been analysed by extracting a variety of features, which have been subsequently evaluated by combining both classic and machine learning statistical techniques. Results are reported and discussed.

Keywords

Music Information Retrieval; Performance Analysis; Sentiment Analysis; Music and Emotions; Tuba

1. INTRODUCTION

The sound and music computing literature has investigated several approaches relevant for *Music Information Retrieval* (MIR) tasks, and a prominent one is the identification of emotions expressed by music. Yet, when such perspective is embraced a deceptively simple question should be answered: which facets of music (such as: harmony, rhythm and timbre) and which combinations of these features can be effectively exploited for emotion identification?

In order to make a principled step in such direction, here we present a preliminary study on how audio features are likely to be modulated while musicians try to intentionally convey emotional contents through expressiveness.

Reviews made by [6] and [19] revealed that there is not a dominant feature among the others able to distinguish different emotions, instead a large set of features is necessary to get some result. As to expressivenes, [9] found a set of relevant *score-independent* features (such as: *Roughness*, *Attack time*, *Peak level* and *Notes per second*). In [3] a great variety of features is used to achieve emotion classification and are also related to which emotion they outline,

e.g. *Dynamics*, *Timbre* and *Articulation*. Research in [18] was aimed at finding which aspects of the signal other than *brightness* and *attack time* were relevant in timbre-evoked emotions, and determined the relevance of *odd-even ratio*, a feature strictly related to the harmonic properties of the signal. Eventually, [16] suggest the choice of *tempo*, *dynamics*, and *articulation* as key feature-classes for studying music expressiveness, while demonstrating that musicians play in a different way whether they are trying to *express* an emotion rather than when *feeling* an emotion.

In the study reported here, as a novel contribution, we set up an experiment leading the musician to mostly exploit on *emotional contagion* and *auditory sensations* rather than *musical expectancy*, *episodic memories* or others subjective phenomena. Then, we have exploited a mix of statistical analysis (*ANalysis Of VAriance*, ANOVA) and machine learning techniques [1], such as dimensionality reduction (*Principal Component Analysis*, PCA) and automatic classification (*Support Vector Machines*, SVM), to analyse the emotional content conveyed by the audio features.

2. BACKGROUND AND RATIONALES

The music ability to elicit emotions intrigues researchers since a long time [2]. Juslin and Västfjäll reviewed several works related to this topic, and, most important, identified six mechanisms through which music can evoke emotions [5]:

1. Brain stem reflexes: Emotion is induced by *sound* itself because one or more acoustical characteristics of the music are taken by the brain stem to signal a potentially important event; reflexes account for the impact of *auditory sensations*

2. Evaluative conditioning: Listened music is repeatedly paired with other positive or negative stimuli.

3. Emotional contagion: The listener "perceives" the emotional expression of the music, and then "mimics" this expression internally, which leads to an induction of the same emotion by means of either peripheral feedback from muscles, or a more direct activation of the relevant emotional representations in the brain.

4. Visual imagery: The listener conjures up visual images while listening to the music; experienced emotions are the result of a close interaction between the music and the images.

5. Episodic memory: The music evokes a memory of a particular event in the listener's life.

6. Music expectancies: A specific feature violates, delays, or confirms the listener's expectations about the continua-

tion of the music.

It is clear that mechanisms 2, 4, 5 and 6 are strictly subjective, while 1 and 3 rely on some degree of objective (or at least shared) aspects.In particular, emotional contagion is likely to be deeply rooted in simulation and motor-based mechanisms of emotion and thus relying on a shared manifold between subjects [17]. Under this rationale, 1 and 3 have been considered the mechanisms to be principally elicited in setting up the experiment.

3. EXPERIMENTAL SETUP

Ten tuba players have been asked to play 7 times a music fragment composed *ad hoc*, each time trying to express a different emotion.

We choose the six *basic emotions* namely, *anger, disgust, fear, happiness, sadness* and *surprise* [4], in spite of the ones cited in [14], since the mechanisms we are investigating are closer to low-level emotions rather than the high-level ones involved in *music expectancies* [5]. Nevertheless, since not all basic emotions are usually embedded in music performances, we expect some discordance among the performers (e.g. when asking for *disgust*).

The players came from different geographical areas and cultural and musical backgrounds: a tuba teacher, a graduated professional, a graduating student, a semi-pro, 3 young mid-level students and 3 amateurs. Each performer had sufficient skills to perform the melodic fragment: being able to express emotions regardless of the technical aspects was a basic condition in order to add expressiveness to the performance.

Materials: Music Fragment and Instrument. Materials choice aimed at maximizing the influence of shared mechanisms (*brain stem reflexes* and *emotional contagion*) over acoustic outcomes with respect to other mechanisms (cfr. Section 2).

A music fragment (shown in Figure 1) has been composed *ad hoc*. It presents no hints about known tunes, this should avoid emotions evoked by *episodic memories* and likely *Evaluative conditionings*. Moreover, it sounds more like a *succession of pitches* rather than a *melody*, almost resembling a serial sequence: this should help avoiding any *music expectancy*. Just a small variance in note durations has been introduced to let some room for the performer to add expressiveness without stimulating much rhythmic expectancies. Note that the score has been written with neither key signatures nor time indications, dynamics and agogics. Performers are used to read scores containing tempo, phrasing and articulation signs; a score like this should guarantee more degrees of freedom to the interpreters.

Since for this experiment an instrument capable of great expression is needed, a wind instrument seems the best choice. Brasses, in particular, are very close to the human voice from both a technical and a timbral point of view. We chose the tuba since its sound is generated by the musician and then amplified, tuned and filtered by the instrument: this allows the artist to take control over many aspects of the timbre, enabling great expressiveness. Of course, many other brasses share the same advantages but our knowledge of this instrument allowed to conduct a sound analysis taking into account its potential and capabilities.

Data acquisition Once the performer was comfortable with the setup, 7 performances have been recorded: 6 while the musician tried to express basic emotions and a *neutral*

Figure 1: The *ad hoc* composed music fragment

one, which served as a baseline reference.

To prevent systematic bias in collecting data due to a common sequence pattern, the 6 basic emotions have been considered in random order. For each emotion the performer has been given given some time to enter the right mood, and then the performance was recorded and listened back until the musician was satisfied of the outcome.

The neutral recording was the last task, made after a small debriefing phase and with a metronome to a fixed tempo of 85 BPM: the attention given to the metronome and to a rigorous execution was meant to leave no space for emotions in the performance, thus flattening the expressiveness.

All recordings have been made in the same environment (a small concert hall), with the same equipment and with the same settings. The aim was limiting the influence of ambience on musicians while keeping the level differences consistent across the performances.

All signals have been recorded as mono PCM 44.100kHz 24bit files, the hardware and software tools used are: an *AKG C 419* Clip Microphone for wind instruments; an *M-AUDIO MobilePre USB* sound card; a laptop PC running *Cockos Reaper v4.77 x64* digital audio workstation. To prevent biases in the computation of some features, each recorded track has been manually processed in order to set leading and trailing silence to 0.5 seconds. Since the microphone was attached to the instrument, almost no reverberation was captured by the recordings (e.g. the ratio of *dry* vs. *wet* signal is very high).

4. DATA ANALYSIS AND RESULTS

Feature extraction has been performed within the Matlab environment, using both toolboxes (such as the *MIRToolbox* [8] and the *Timbre Toolbox* [10]) and custom algorithms, such as [12]. Table 1 shows the complete feature set together with the results obtained. Below we briefly discuss briefly how each feature maps over the tuba performance.

- Beats Per Minute (BPM) corresponds to the speed of the execution.

- Root Mean Square (RMS) is a measure of the signal energy; here it reflects how *loud* the instrument has been played.

- Low Energy (LOW) is the percentage of signal with a level below the average [15]. This is a good measure of the amount of amplitude modulation and *legatos*.

- Attack leap (ATK) is a measure of the dynamic range of the attack phase of the sound [8]. This is influenced by transients, *staccatos* and abrupt dynamic changes, in opposition to *legatos* and soft attacks.

- Harmonic energy (HAE), Noise Energy (NOE), Noisiness (NSN) and **Harmonic Spectral Deviation (HRD)** measure the periodic vs. non-periodic component of the sound [10, 7]; in our context is intended to measure the balance between pitched notes and buzz.

- Brightness (EBF) by and large measures the amount of

high-frequencies in the signal. Many algorithms have been proposed to measure EBF, for our task the one described in [12] and implemented in [13] gives the best results. In tuba performance many aspects of sound production may influence EBF.

- Tristimulus (T1-3), a concept borrowed from colour perception studies, is triple which in MIR is used for measuring the contribution of the fundamental, the second and third harmonic, and all the remaining components on the overall sound [11]. In our context, T1 and T3 can be thought as how many harmonics are excited.

- Inharmonicity (INH) and **Roughness (ROH)** are respectively a measure of the partials that fall outside the harmonic series and a measure of the intrinsic dissonance of the sound [10]. They both measure harshness of sound and may be good indicators of *sforzato*, *singing through the instrument* or other peculiar techniques.

- Odd-Even Ratio (OER) is the ratio between the energy of odd and even harmonics. Usually it depends only on the instrument, but it may also reflect the intention of the performer to obtain a *softer* or a *bitter* sound.

Folllowing [10], time-varying features are collapsed into scalar values through their medians (M) and interquartile ranges (IQR). Since a certain amount of subjectivity and cultural differences are present also in the *contagion* mechanism, and since the technical skills of the performers were not as uniform as expected, all features were normalized by performer. This allowed to evaluate the relation among different emotions rather than the artist's skills or instrument manufacture. The only feature that was considered by its non-normalized value[1] is BPM (denoted $\mathbf{BPM_{nn}}$ in Table 1), since it showed some predictive accuracy also in the non-normalized version.

ANOVA was performed on each feature to test statistical significance when distinguishing between emotions. All features can be considered statistically significant ($p < 0.05$), except for $\mathrm{ATK_{IQR}}$ and $\mathrm{HRD_{IQR}}$, which were thus discarded from further analyses.

We also analysed the feature space to test when and where features means were significantly different among the emotions; in other words we investigated which classes were distinguished by each feature. The outcomes are graphically shown in Table 1, each dot meaning that the feature can help distinguishing the corresponding pair of classes. The table shows that each pair of emotions is distinguishable by at least one feature.

A redundancy analysis over the remaining 24 features was then performed via PCA [1], showing that 4 principal components (PCs) have an eigenvalue greater than one, explaining 82.1% of the variance. To explain more than 90%, 7 PCs are necessary. An ANOVA test run over all PCs revealed that $\mathrm{PC_{1-3}}$ are useful to differentiate basic emotions, but none of the remaining are statistically significant. Moreover, always according to ANOVA, no PC can help distinguishing *disgust* vs. *surprise*. Feature coefficients related to $\mathrm{PC_{1-3}}$ are shown in Table 1.

To assess the goodness of ANOVA and PCA results in terms of predictive accuracy, we automatically classified recordings through an SVM [1], then we considered the *F*-Scores obtained from the confusion matrices generated according to

a *leave-one-out* strategy [1] over the 70 available recordings.

The test considered: the whole feature set (called 24F in Table 2), sets of relevant PCs (respectively 7PC, 4PC, and 3PC) and different subsets of features.

Clearly, according to ANOVA, we expect to see lower *F*-scores for emotions with less features binding them. As to PCA, a failure in isolating the right components will result in better *F*-Scores when using feature subset rather than PCs sets.

SVM behaved as expected for what concerns the 24F set, presenting smaller *F*-Scores for *fear*, *disgust* and *surprise*, while holding a greater accuracy in other classes, thus confirming ANOVA results. SVM scores seem to be higher with reduced feature sets, suggesting the need of dimensionality reduction. Unfortunately PCA seems to fail to isolate the right components: 7F, 4F, and 3F present better results than 7PC, 4PC, and 3PC.

5. CONCLUSIONS

We outlined an experiment to investigate the effect of expressiveness on audio features, where a set of musicians was asked to play a music fragment composed *ad hoc* while trying to express different emotions. Recorded data have been processed by extracting 26 features subsequently analysed via ANOVA, PCA and an SVM classifier. Overall preliminary results look promising.

The ANOVA test discarded 2 features and gave a broad view of which features can distinguish each pair of emotions. PCA on the one hand revealed the presence of redundancy but, on the other hand for what concerns our goals, failed in isolating the correct descriptors, probably due to the small size of our dataset (70 recordings and 24 features). Different dimensionality reduction techniques will be explored in future works.

SVM classification confirmed the reliability of ANOVA results, in particular the confusion between some classes of emotions, which needs to be addressed to understand whether the cause of the cluttering is a matter of chosen features or a matter of eliciting mechanisms.

6. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[2] M. Budd et al. *Music and the emotions: The philosophical theories*. Routledge, 2002.

[3] T. Eerola, O. Lartillot, and P. Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *ISMIR*, pages 621–626, 2009.

[4] P. Ekman and K. Scherer. Expression and the nature of emotion. *Approaches to emotion*, 3:19–344, 1984.

[5] P. N. Juslin and D. Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(05):559–575, 2008.

[6] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[7] J. Krimphoff et al. Characterization of the timbre of complex sounds. 2. Acoustic analysis and psychophysical quantification. *J. de Physique*, 4:625–628, 1994.

---

[1] Actually, the value of $\mathrm{BMP_{nn}}$ is normalized among all recordings – instead of the per-musician normalization used elsewhere – to match other features' range

Table 1: 26 chosen features, with emotion pairs and PCA coefficients.
A: Anger, D: Disgust, F: Fear, H: Happiness, S: Sadness, U: Surprise, N: Neutral.

	AS	AN	SU	AH	UN	AF	HN	HS	FN	SD	SN	DF	HF	AD	AU	HU	DN	UF	HD	DU	SF	#	PC1	PC2	PC3
BPM	•		•		•					•							•	•	•	•		8	0,17	0,03	**0,41**
BPM-nn			•	•	•			•		•								•				6	0,17	0,06	**0,41**
RMS	•	•	•	•		•		•	•					•	•							9	0,25	-0,21	-0,07
LOW	•	•	•	•	•		•									•					•	8	0,23	0,16	-0,01
ATK-M	•	•	•		•				•													5	0,18	-0,01	**0,30**
ATK-IQR																						0	·	·	
EBF-M	•	•	•	•		•		•	•		•			•	•			•				11	0,24	-0,22	-0,05
EBF-IQR	•	•	•	•	•		•						•									7	0,24	0,07	-0,04
T1-M	•	•	•	•		•		•	•			•										8	-0,23	0,26	0,06
T1-IQR	•					•	•				•	•					•					6	-0,02	**0,35**	-0,13
T3-M	•	•	•	•		•		•	•			•		•								9	0,26	-0,16	-0,16
T3-IQR	•	•		•		•			•								•		•			7	0,18	-0,10	-0,32
HAE-M		•		•	•		•			•	•		•									7	-0,23	-0,22	0,12
HAE-IQR		•		•	•											•						4	-0,20	-0,17	**0,26**
NOE-M		•		•	•		•			•	•		•			•						8	-0,22	-0,22	0,18
NOE-IQR				•																		1	-0,16	-0,11	**0,35**
NSN-M	•	•	•		•									•						•		6	0,24	0,17	0,19
NSN-IQR	•	•	•		•	•	•	•		•				•								9	0,25	0,08	0,20
HRD-M		•	•		•			•			•	•		•			•					8	-0,21	-0,26	0,09
HRD-IQR																						0	·	·	
INH-M	•	•	•		•		•															5	0,23	0,18	0,13
INH-IQR	•	•	•					•	•													5	0,22	0,04	**0,22**
ROH-M	•				•				•		•				•							5	0,11	-0,32	-0,09
ROH-IQR	•	•		•		•			•		•			•	•							8	0,22	-0,22	-0,15
OER-M	•		•			•		•	•		•	•					•	•				9	-0,18	**0,30**	-0,07
OER-IQR	•					•	•	•	•								•	•				9	-0,05	**0,38**	-0,08
PC1	•	•	•	•	•	•	•	•		•				•			•	•			•	13			
PC2	•			•	•			•			•	•	•		•							8			
PC3			•				•	•					•				•		•			6			
	20	18	16	15	14	13	11	11	10	9	8	7	7	6	6	6	5	5	4	2	2	#			

Table 2: *F*-Scores of SVM classification
SVM trained with different sets of features and PCs.

	24F	7F	7PC	4F	4PC	3F	3PC
Ang	78%	**95%**	58%	72%	69%	62%	69%
Hap	60%	82%	57%	63%	73%	71%	64%
Sad	67%	76%	42%	80%	63%	62%	55%
Dis	32%	56%	35%	40%	0%	27%	14%
Sur	21%	63%	35%	50%	27%	71%	27%
Fea	32%	47%	33%	50%	21%	0%	0%
Neu	82%	82%	67%	83%	77%	69%	71%

7F: BPM, LOW, RMS, ROH_M, ROH_{IQR} HAE_M OER_M.
4F: BPM, EBF_{IQR}, EBF_M, HRD_M; **3F:** NSN_{IQR}, NOE_M, $T3_M$

[8] O. Lartillot, P. Toiviainen, and T. Eerola. A Matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*, pages 261–268. Springer, 2008.

[9] L. Mion and G. D. Poli. Score-independent audio features for description of music expression. *IEEE Trans. Audio, Speech, and Language Processing*, 16(2):458–466, 2008.

[10] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams. The timbre toolbox: Extracting audio descriptors from musical signals. *J. Acoustical Society of America*, 130(5):2902–2916, 2011.

[11] H. F. Pollard and E. V. Jansson. A tristimulus method for the specification of musical timbre. *Acta Acustica united with Acustica*, 51(3):162–171, 1982.

[12] G. Presti and D. Mauro. Continuous Brightness Estimation (CoBE): Implementation and its possible applications. In *10th Proc. Int. Symp. Comp. Music Modeling and Retrieval (CMMR)*, pages 967–974, 2013.

[13] G. Presti, D. Mauro, and G. Haus. TRAP: TRAnsient Presence detection exploiting continuous brightness estimation (CoBE). In *Proceedings of the 12th Sound and Music Computing Conference (SMC 2015), Maynooth, Ireland*, pages 379–385, 2015.

[14] E. Schubert. Update of the hevner adjective checklist. *Perceptual and motor skills*, 96(3c):1117–1122, 2003.

[15] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech and*

Audio Processing,, 10(5):293–302, 2002.

[16] A. G. Van Zijl, P. Toiviainen, O. Lartillot, and G. Luck. The sound of emotion: The effect of performers' experienced emotions on auditory performance characteristics. *Music Perception: An Interdisciplinary Journal*, 32(1):33–50, 2014.

[17] J. Vitale, M.-A. Williams, B. Johnston, and G. Boccignone. Affective facial expression processing via simulation: A probabilistic model. *Biologically Inspired Cognitive Architectures Journal*, 10:30–41, 2014.

[18] B. Wu, A. Horner, and C. Lee. Musical timbre and emotion: The identification of salient timbral features in sustained musical instrument tones equalized in attack time and spectral centroid. In *Proc. 40th Int. Comp. Music Conf.(ICMC)*, pages 928–934, 2014.

[19] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Trans. Intelligent Systems and Technology (TIST)*, 3(3):40, 2012.