OBJECTIVE CHARACTERIZATION OF AUDIO SIGNAL QUALITY: APPLICATIONS TO MUSIC COLLECTION DESCRIPTION

Dominique Fourer, Geoffroy Peeters

OBJECTIVE CHARACTERIZATION OF AUDIO SIGNAL QUALITY: APPLICATIONS TO MUSIC COLLECTION DESCRIPTION

Dominique Fourer *Geoffroy Peeters*

UMR STMS (IRCAM - CNRS - UPMC)

dominique@fourer.fr, geoffroy.peeters@ircam.fr

In this paper, we propose a set of audio features to describe the quality of an audio signal. Audio quality is here considered as being modified by the chain of processes/effects applied to the individual instrument tracks to obtain the final mix of a musical piece. Thus, the quality also depends on the mastering processes applied to the final mix or the signal degradation caused by MP3 compression. To evaluate our proposal, we created a large set of artificial mixes and also used real-world studio mixes. Using unsupervised and supervised classification methods, we show that our proposed audio features can detect the processing chain. Since this processing chain applied in professional studio has evolved over the years, we use our audio features to directly predict the decade during which a music track was recorded.

Index Terms— audio quality, music information retrieval, audio reverse-engineering, database indexing, music remixing.

1. INTRODUCTION

Audio signal quality can be related to subjective and objective audio signal attributes resulting from a sophisticated digital signal processing chain. Despite, a consistent definition of audio quality has not yet been offered, researchers agree to say that it depends on a combination of transformations applied to audio signal since studio recording (or pure synthesis) to the resulting final mix obtained after mastering [1]. Knowing the audio quality of a music track is full of interest for applications such as music streaming or playlist generation since it allows to decide which sound file (when several occurrences of the same music track exist in a database) has the better quality or should be discarded.

Among the first works related to the objective description of the audio quality, the standard ISO/IEC 15938-4 (MPEG-7 Audio) [2] proposes a set of informative features to describe the audio content and the signal quality. More recently, audio quality has re-gained interest. In 2011, [3, 1] propose to use a set of audio quality features to estimate the decade during which a musical piece was recorded and help the navigation in large music collection. In 2015, [4] performs a set of perceptual experiments in which users judge the audio quality through listening tests. This leads to an audio quality lexicon. [5, 6, 7] propose a set of audio quality features to predict the results of the perceptual experiments using a machine learning approach.

In this paper, we extend this approach, *i.e.* we propose an objective description of the audio quality. We aim at describing the audio signal content related to the mixing process and the signal quality. Hence, this approach is directly related to the audio signal reverse engineering problem [8, 9] which finds applications in music description, audio branding [10], automatic playlist generation or automatic music mixing [11].

The paper is organized as follows. In Section 2, we address problem of the objective description of the audio quality by first describing the set of considered alteration effects. We then present the set of proposed audio signal quality features. In Section 3, we apply this framework to automatically predict the audio signal alterations and to automatically predict the music decade. We finally discuss the results and present future works in Section 4.

2. TOWARDS OBJECTIVE AUDIO QUALITY ASSESSMENT

A music audio signal is the result of the mixing of a set of effects and transformations applied on separated tracks (elementary signals) in order to obtain an artistic mixture [11]. These transformations, which are subjectively applied in studio by sound engineers, depend on the targeted music medium and are often difficult to reverse. Because of this, audio mixing reverse engineering recently gained interest. It was addressed for example in [8, 9]. Furthermore, after studio mixing, audio signals can also be degraded by signal transformations resulting from users manipulation (*e.g.* remixing, resampling, lossy compression, etc.). This results in a loss of quality, which can be characterized for example, by an addition of noise or a reduction of the content frequency bandwidth.

Hence, the audio quality characterization problem addressed here consists in either obtaining cues about the signal mixing process or (ideal case) recovering the exact signal properties related to the transformations which have been applied to the signal.

This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement n^o 688122.

2.1. Considered audio signal alteration effects

In the present work we only consider a restricted set of signal alterations. Those however cover a wide range of commonly used audio transformations, as often addressed in the music processing literature [12, 11, 8, 13, 1, 14].

2.1.1. Dynamic range control

Dynamic range control is a non-linear effect, which modifies the overall loudness perception of the audio signal. This can be done by amplifying the volume of quiet sounds and reducing the volume of loud sounds. The result is a transformation of the dynamic range of the input signal. A compressor or an expander, are often parameterized by a detection threshold L and a gain ratio R. Delay parameters τ_v^{att}, τ_v^{rel} and τ_g^{att}, τ_g^{rel} can also be used to obtain smoothed detection and gain functions. Furthermore, different compressor parameters can also be separately applied on arbitrary chosen signal frequency bands [12]. In our experiments, we apply a compressor-expander (compander) on a Linear Instantaneous (LI) mixture using SoX with parameters depending on the signal profile, as detailed in Table 1.

Table 1. SoX profiles used for dynamic range control.

Profile name	SoX parameters
speech	compand 0.02,0.20 5:-60,-40,-10 -5 -90 0.1
streaming	compand 0.3,1 6:-70,-60,-20 -5 -90
speech/music	compand 0.1,0.3 -60,-60,-30,-15,-20,-12,-4,-8,-2,-7 -2
music 1	compand 0.3,1 -90,-90,-70,-70,-60,-20,0,0 -10 0 0.2
music 2	compand 0.3,1 6:-70,-60,-20 -5 -90 0.2

2.1.2. Spatialization

In our experiment we only consider two-channels mixtures. We denote by $s_1[n]$ and $s_2[n]$ the left and right channels discrete-time signals and by $S_1[n,m]$ and $S_2[n,m]$ their discrete Short-Time Fourier Transforms (STFT). The considered stereophonic transformations are described as follows.

- **Mono** effect consists in duplicating both channels: $s_2[n] = s_1[n]$.
- **Amplitude panning** aims at simulating the direction of arrival of a mixture by changing its amplitude on each channel. For a given azimuth $\theta \in [-\frac{\pi}{2}, +\frac{\pi}{2}]$ and a monophonic input signal $x[n]$, the left and right channels are given by

$$\begin{pmatrix} s_1[n] \\ s_2[n] \end{pmatrix} = \begin{pmatrix} \sin\left(\frac{\theta}{2} + \frac{\pi}{4}\right) \\ \cos\left(\frac{\theta}{2} + \frac{\pi}{4}\right) \end{pmatrix} x[n]. \quad (1)$$

  If $\theta = 0$ then $s_1[n] = s_2[n]$, if $\theta = -\frac{\pi}{2}$ then $s_1[n] = x[n]$, $s_2[n] = 0$, if $\theta = \frac{\pi}{2}$ then $s_1[n] = 0$, $s_2[n] = x[n]$.
- **Phase panning** changes the delay between both channels. It can be implemented by transforming the STFT phase. If we denote by $\Delta\phi$, the phase lag parameter, this effect is obtained by $S_2[n,m] = S_1[n,m]e^{j\Delta\phi}$. $s_2[n]$ is then obtained by inversion of the STFT.
- **Head Related Transfer Function (HRTF)** filtering aims at simulating the binaural perception of a source signal which arrives from a given direction. This effect is simply obtained by convolving the source signal $x[n]$ by the left and right impulse responses corresponding to the given azimuth θ. Our experiments use the CIPIC HRTF database [15].

2.1.3. Lossy audio compression

We simulate this alteration by encoding the original audio mixture in the MP3 audio file format [16] (which is the most popular format for audio storage, transfer and playback). For this, we used the LAME encoder with four different quality profiles corresponding to the following bitrates: 16 kbs, 64 kbs, 128 kbs and 320 kbs.

2.1.4. Content alteration

We also consider two content alteration effects:

- **Resampling** consists in changing the number of samples. Down-sampling reduces it while up-sampling increases it. This results in an increase or a decrease of the original signal frequency bandwidth. It is often related to a loss of signal quality (in particular for down-sampling). Our original sampling rate is 44.1 kHz. We consider four profiles in our experiments: 8 kHz (down), 16 kHz (down), 32 kHz (down), 96 kHz (up).
- **Noise addition** is simply achieved by merging (addition) the original signal with a white Gaussian noise signal. We used five different Signal-to-Noise Ratio (SNR) values: -15 dB, -5 dB, 10 dB, 20 dB and 45 dB.

2.2. Audio quality features

For the purpose of describing the audio quality, we collect previously proposed audio quality features from -[1] (average spectrum), -[2] (monophony detector, cross-channel correlation, relative delay, balance, DC-offset, frequency bandwidth, background noise-level), -[3] (dynamic histogram, cochleagram difference, spectral stereo phase spread) and -[17] (spectral entropy). For reason of restricted length of the paper, we refer the readers to the respective publications for a detailed description. The entire set of features used in this study is summarized in Table 2. Features corresponding to a time series (DH, AS, SE and BW) are statistically summarized by their mean, median, Inter-Quartile Range (IQR), standard deviation, skewness, kurtosis, minimum, maximum, entropy and slope over time. This leads to 10 scalars for each frame-based feature. For DH and AS, we also compute the centroid and the position of the maximum. The CD feature is represented by a matrix $\mathcal{D}$ of size $M \times N$ (M denotes frequency bands and N time-frames). It is summarized by 5 scalars expressed as $\text{CD}_1 = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}|\mathcal{D}_{m,n}|$, $\text{CD}_2 = \sigma\left(\frac{1}{N}\sum_{n=0}^{N-1}|\mathcal{D}_{m,n}|\right)$, $\text{CD}_3 = \frac{1}{MN}\sum_{n=0}^{N-1}\left|\sum_{m=0}^{M-1}\mathcal{D}_{m,n}\right|$

SoX - Sound eXchange: `http://sox.sourceforge.net/`

LAME encoder: `http://lame.sourceforge.net/`

and $\mathrm{CD}_4 = \sigma\left(\frac{1}{M}\sum_{m=0}^{M-1}\mathcal{D}_{m,n}\right)$, where $\sigma(x)$ denotes the standard deviation of the time series x.

Each sound file is represented by a vector of features, which is used as the input of a classification method.

Table 2. List of proposed audio signal quality features.

Feature name	Label	Designation	#
Dynamic histogram [3]	DH	mixture dynamic range	12
Average spectrum[1]	AS		12
Cochleagram difference [3]	CD	stereo quality	5
Spectral Stereo Phase Spread [3]	SSPS		1
Monophony detector[2]	isMono		1
Cross-channel correlation[2]	CCCor		1
Relative delay[2]	RDelay		1
Balance [2]	Bal		1
DC-offset [2]	DCOff	signal content	1
Root Mean Squared amplitude	aRMS		1
Spectral Entropy [17]	SE		10
Frequency bandwidth[2]	BW		10
Background noise level [2]	BNL		1
Total number of features			57

3. NUMERICAL EXPERIMENTS

In order to validate our proposed audio quality features, we use the following two experimental scenarios.

3.1. Scenario 1: prediction of audio signal alterations

In this scenario, we used the Medley dataset [18] which provides 122 music pieces available in two forms: 1) studio mix (which is a stereo high-quality artistic mixture with effects applied to the various instrument tracks), 2) the set of separate multi-track instruments which allows to build a flat monophonic mix (named LI mix). On those, we apply a set of 27 different signal alteration effects. Those are detailed in Section 2.1 and summarized in Table 3. It should be noted that several sets of parameters can be used for a given effect to obtain several instances (*e.g.* dynamic range compression uses 5 different profiles).

Table 3. List of considered simulated audio alteration.

Effect name (# of classes)	Profiles	#
Dynamic range control (7)	no compression (LI mix)	1
	reference studio mix	1
	dynamic range compression (SoX)	5
Spatialization (5)	reference studio mix	1
	mono	1
	amplitude panning	4
	phase panning	4
	HRTF	4
Lossy compression (5)	original WAV file	1
	MP3 compression (LAME encoder)	4
Content alteration (10)	resampling	5
	addition of a white Gaussian noise	5

The goal of our experiment is to automatically recognize the alteration effects applied to the audio, using our audio features (*cf.* Table 2). We consider this as a set of classification tasks with 7 classes for the dynamic range control, 5 for spatialization, 5 for lossy compression and 10 for content alteration. We try to solve these tasks using both supervised and unsupervised classification.

3.1.1. Supervised classification

We tested the following supervised classification algorithms: K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA)[19] and multiclass (one-against-all) Radial Basis Function (RBF) kernel Support Vector Machine (SVM) [20]. We performed a 3-fold cross-validation, with randomly partitioned equal sized folds. For the KNN method, we used $K = 9$ (which was empirically found to provide the best results) and a city-block distance (also named Manhattan) which provided better results than an Euclidean one.

The classification results are indicated in Table 4 in terms of class-recall and global accuracy. The corresponding confusion matrices are indicated in Tables 5 (a)-(d).

Table 4. Supervised classification results for each task in terms of recall and of accuracy.

Method	Dynamic range control class name							**Accuracy**
	no co.	stud.	spee.	stream.	spe./mus.	mus.1	mus.2	
KNN	0.36	0.80	0.23	0.08	0.26	0.44	0.06	0.32
LDA	0.72	0.98	**0.65**	**0.48**	**0.89**	**0.96**	**0.27**	**0.71**
SVM	**0.90**	**0.99**	0.48	0.37	0.23	0.95	0.09	0.57

Method	Spatialization class name					**Accuracy**
	stud. mix	mono	amp. pan.	phs. pan.	HRTF	
KNN	0.31	0.34	0.90	0.85	0.98	0.83
LDA	0.94	**1**	0.97	0.57	**1**	0.86
SVM	**0.96**	0.89	**1**	**0.97**	0.99	**0.98**

Method	Lossy compression class name					**Accuracy**
	orig. wav	mp3 320kbs	mp3 128kbs	mp3 64kbs	mp3 16kbs	
KNN	0.34	0.20	0.20	0.99	**1**	0.55
LDA	0.73	**0.80**	**0.85**	**1**	**1**	**0.88**
SVM	**0.75**	0.59	0.43	**1**	0.99	0.75

Method	Content alteration class name										**Acc.**
	8kHz	16kHz	32kHz	44kHz	96kHz	-15dB	-5dB	10dB	20dB	45 dB	
KNN	0.83	0.72	0.51	0.25	0.32	**1**	**1**	0.90	0.61	0.24	0.64
LDA	0.87	**0.89**	**0.81**	0.55	**0.68**	**1**	**1**	**0.98**	**0.94**	**0.77**	**0.85**
SVM	**0.90**	0.80	0.70	**0.57**	0.65	0.99	**1**	0.89	0.66	0.46	0.76

The table shows that SVM and LDA outperform the KNN method in all cases. Best results are obtained to predict spatialization classes (98%), then lossy compression (88%), then content alteration (85%) and finally dynamic range (71%). It should be noted that discriminating the dynamic range classes is a much harder problem but interestingly, our method achieves (but with difficulties according to Table. 5 (a)) to discriminate the "streaming" and the "music 2" profiles which use almost identical SoX parameters.

3.1.2. Unsupervised classification

For the unsupervised classification, we use a simple k-means algorithm [21] using again the city-block distance. The number K of clusters is fitted to the number of classes to discriminate (*cf.* Table 3). For the unsupervised case, we used an automatic feature selection algorithm in order to detect the most relevant features to be used for each classification task. For this, we used the Inertia Ratio Maximization using Feature Space Projection (IRMFSP) [22] algorithm. Thus, the most informative features are indicated in Table 7.

In Table 6 we indicate the results of the unsupervised classification in terms of cluster purity [23] (a cluster containing a single class has a purity of 1). Using less than 7 signal features, we obtain purities above 0.6 for all classification tasks.

Table 5. Confusion matrices obtained using the best classification method for each prediction task.

(a) LDA method applied to dynamic range control effect prediction.

ref. class	Classified as						
	no comp.	stud. mix	spee.	stream.	spe./mus.	mus.1	mus.2
no comp.	**88**				4	30	
stud. mix		**119**	1		1	1	
spee.			**79**	14	8		21
stream.			9	**59**	1		53
spe./mus.			4		**109**		
mus.1					5	**117**	
mus.2			40	**49**			33

(b) LDA method applied to dynamic range control effect prediction.

ref. class	Classified as									
	8kHz	16kHz	32kHz	44kHz	96kHz	-15dB	-5dB	10dB	20dB	45 dB
8kHz	**106**	13								3
16kHz	4	**119**		6	1	1				3
32kHz		2	**99**	5	14					2
44kHz	1	6	7	**67**	19					22
96kHz		2	15	19	**83**					3
-15dB						**122**				
-5dB							**122**			
10dB								**119**	3	
20dB								2	**115**	5
45dB		3		20					5	**94**

(c) LDA method applied to lossy compression detect.

ref. class	Classified as				
	wav	320kbs	128kbs	64kbs	16kbs
orig. wav	**89**	25	8		
mp3 320kbs	18	**97**	7		
mp3 128kbs	5	13	**104**		
mp3 64kbs				**122**	
mp3 16kbs					**122**

(d) SVM method applied to spatialization effect detect.

ref. class	Classified as				
	stud.mix	mono	amp. pan.	phs. pan.	HRTF
stud.mix	**117**			1	4
mono		**108**		14	
amp. pan.			**487**	1	
phs. pan.		2	11	**475**	
HRTF	1	2			**485**

(e) SVM method applied to decade predict.

ref. class	Classified as				
	60s	70s	80s	90s	2000s
60s	**327**	49	8	7	5
70s	73	**123**	**136**	45	19
80s	9	74	**272**	35	6
90s	26	29	61	**216**	64
2000s	11	11	16	44	**314**

As for the supervised case, the spatialization classes are also the best predicted in the unsupervised case. These results are very promising for future unsupervised applications.

Table 6. Unsupervised classification results for each task in terms of cluster purity, optimal number of features, number of clusters K.

Task	Cluster purity	# of feat.	K
Dynamic range control	0.62	4	7
Spatialization	0.80	5	5
Lossy compression	0.78	3	5
Resampling	0.71	2	5
Noise add.	0.78	7	5
Noise add.+Resampling	0.63	6	10

Table 7. Top-10 features sorted by descending order of the Fisher score (FS) [22] for each classification task.

rank	dynamic range control		spatialization		lossy compression		content alteration	
	feat. name	FS	feat. name	FS	feat. name	FS	feat. name	FS
1	aRMS	0.80	isMono	0.71	mean AS	1	median AS	1
2	SSPS	0.70	CCCor	0.60	slope AS	0.96	mean BW	0.74
3	min DH	0.42	CD5	0.53	max BW	0.39	max BW	0.69
4	CCCor	0.23	SSPS	0.45	mean BW	0.22	min SE	0.33
5	DH pk. pos.	0.13	CD1	0.23	median AS	0.06	mean SE	0.28
6	CD1	0.05	CD4	0.07	std BW	0.06	skew. SE	0.18
7	entropy DH	0.04	aRMS	0.05	std AS	0.05	median SE	0.16
8	skew. DH	0.03	CD3	0.04	max AS	0.02	max SE	0.14
9	std. DH	0.02	Bal	0.02	skew. BW	0.02	entropy SE	0.12
10	slope DH	0.01	slope AS	0.02	iqr BW	0.02	min DH	0.10

3.2. Scenario 2: Decade prediction

Since the processing chain applied in professional studio has evolved over the years, and since our audio features allows describing this chain, we test the use our audio features to predict directly the decades during which a music track was recorded. For this, we consider the dataset of 1980 music tracks previously used in [3]. It covers the years from 1960 to 2000. We subdivide them into 5 decades considered as classes. Each musical piece is resampled at 44.1 kHz and only the 60 first seconds are analyzed. As before, we tested the following supervised classification algorithms: KNN, LDA and SVM using a 3-fold cross-validation scheme. We also applied an artist filter in order to ensure that the same artist is not present both in the training and testing set (it allows to prevent over-fitting [24]).

Results are indicated in Table 8 in terms of recall and accuracy. The best results are obtained using SVM (63%). In comparison, the results obtained in [3] were 61% without adding Mel-Frequency Cepstrum Coefficients (MFCC) to the set of audio quality features and 64% including them. It should be noted however, that the splitting between train and test used here is not the same as the one used in [3].

Table 8. Supervised classification results for decade prediction in terms of recall and of accuracy.

Method	Class name					**Accuracy**
	60s	70s	80s	90s	2000s	
KNN	0.77	0.38	0.63	0.49	0.71	0.60
LDA	0.69	**0.43**	0.62	0.52	0.77	0.60
SVM	**0.83**	0.31	**0.69**	**0.55**	**0.79**	**0.63**

4. CONCLUSION

In this paper, we proposed a set of audio features for the automatic characterization of the audio quality. During an experiment with artificially generated mixing, we showed that the proposed approach can efficiently predict the type of audio effects and alterations applied to the original audio signal. With real commercial music tracks, we also showed that the same approach can be used to predict the decade during which the track was recorded. This approach paves the way of more sophisticated systems designed for automatic mixing, playlist generation or database indexing. Future works will consist in further investigating a larger set of "realistic" signal alterations and using audio quality annotated dataset as in [25].

5. REFERENCES

[1] Pedro Duarte Pestana, Zheng Ma, Joshua D. Reiss, Alvaro Barbosa, and Dawn A. A. Black, "Spectral characteristics of popular commercial recordings 1950-2010," in *135th AES Convention*, NY, USA, Oct. 2013.

[2] Joerg Bitzer and Juergen Herre, "Coding of moving pictures and audio ISO/IEC JTC 1/SC 29/WG 11," Tech. Rep., International Organization for Standardization Organization Internationale Normalization, Shanghai, China, Oct. 2002.

[3] D. Tardieu, E. Detruty, and G. Peeters, "Production effect: Audio features for recordings techniques description and decade prediction," in *Proc. Digital Audio Effects Conf. (DAFx'11)*, Sept. 2011, pp. 441–446.

[4] Alex Wilson and Bruno M Fazenda, "A lexicon of audio quality," in *Proceedings of the 9th Triennial conference of the European Society for the Cognitive Sciences of Music (ESCOM 2015), Manchester, UK*, 2015.

[5] Paul Kendrick, Francis Li, Bruno Fazenda, Iain Jackson, and Trevor Cox, "Perceived audio quality of sounds degraded by nonlinear distortions and single-ended assessment using hasqi," *Journal of the Audio Engineering Society*, vol. 63, no. 9, pp. 698–712, 2015.

[6] Bruno Fazenda, Paul Kendrick, Trevor Cox, Francis Li, and Iain Jackson, "Perception and automated assessment of audio quality in user generated content," in *139th AES Convention*, Manchester, UK, Oct. 2015.

[7] A. Wilson and BM. Fazenda, "Variation in multitrack mixes : analysis of low-level audio signal features," *Journal of the Audio Engineering Society*, vol. 64, no. Issue 7/8, pp. 466–473, 2016.

[8] J. Reiss D. Barchiesi, "Reverse engineering of a mix," *Journal of the Audio Engineering Society*, vol. 58, pp. 563–576, 2010.

[9] S. Gorlow and S. Marchand, "Reverse engineering stereo music recordings pursuing an informed two-stage approach," in *Proc. Digital Audio Effects Conf. (DAFx'13)*, 2013.

[10] K. Bronner and H. Rainer, *Audio Branding. Brands, Sound and Communication*, Nomos, 2009.

[11] Daniele Barchiesi and Josh Reiss, "Automatic target mixing using least-squares optimization of gains and equalization settings," in *Proc. Digital Audio Effects Conf. (DAFx'09)*, 2009, pp. 7–14.

[12] Udo Zölzer, *Digital Audio Signal Processing*, John Wiley & Sons, 2005.

[13] D. Luo, W. Luo, R. Yang, and J. Huang, "Compression history identification for digital audio signal," in *Proc. IEEE ICASSP'12*, Mar. 2012, pp. 1733–1736.

[14] Carlos Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Proc. IEEE WASPAA'03*, Sept. 2003, pp. 55–58.

[15] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proc. IEEE WASPAA'01*, NY, USA, Oct. 2001, pp. 99–102.

[16] "Information technology generic coding of moving pictures and associated audio information ISO/IEC 13818-3 - part 3: Audio," Tech. Rep., 1998.

[17] G Widmer, K Seyerlehner, T Pohle, and M Schedl, "Automatic music detection in television productions," in *Proc. of the International Conference on Digital Audio Effects*, 2007.

[18] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive MIR research," in *Proc. ISMIR'14*, Taipei, Taiwan, Oct. 2014.

[19] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley-Blackwell, NY, USA, 1958.

[20] F. Lauer and Y. Guermeur, "MSVMpack: a multi-class support vector machine package," *Journal of Machine Learning Research*, vol. 12, pp. 2269–2272, 2011, `http://www.loria.fr/~lauer/MSVMpack`.

[21] G. A. F. Seber, *Multivariate Observations*, Hoboken, NJ: John Wiley & Sons, Inc.

[22] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," in *115th AES Convention*, NY, USA, Oct. 2003.

[23] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press., 2008.

[24] A. Flexer, "A closer look on artist filters for musical genre classification," in *Proc. ISMIR'07*, 2007, pp. 341–344.

[25] Alex Wilson and Bruno M Fazenda, "Perception of audio quality in productions of popular music," *Journal of the Audio Engineering Society*, vol. 64, no. 1/2, pp. 23–34, 2016.