Objective characterization of audio signal quality: Applications to music collection description

Dominique Fourer   Geoffroy Peeters
UMR STMS (IRCAM - CNRS - UPMC), Paris, France
dominique@fourer.fr, geoffroy.peeters@ircam.fr

ircam Centre Pompidou

Abstract

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 (*e.g.* mastering, signal compression, etc.). 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.

Summary of the main contributions

- **57 audio quality features** are proposed and investigated.
- **27 distinct alteration classes** for dynamic range control, spatialization, lossy compression and content alteration, are considered.
- An application for **audio signal alteration detection** is proposed and evaluated on the Medley DB [1] (122 tracks with separated stems).
- An application **to music track decade prediction** is proposed and compared with a previously proposed method [2].

Towards objective audio quality assessment

Proposed features

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

- Time series features (**DH**, **AS**, **SE** and **BW**) are summarized by 10 scalars: mean, median, IQR, standard deviation, skewness, kurtosis, minimum, maximum, entropy and slope over time.
- for **DH** and **AS**, we also compute the centroid and the position of the maximum.
- **CD** represented by a matrix $\mathcal{D}$ of size $M \times N$ (M denotes frequency bands and N time-frames) is summarized by 5 scalars: $\mathrm{CD}_1 = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}|\mathcal{D}_{m,n}|$, $\mathrm{CD}_2 = \sigma\left(\frac{1}{N}\sum_{n=0}^{N-1}|\mathcal{D}_{m,n}|\right)$, $\mathrm{CD}_3 = \frac{1}{MN}\sum_{n=0}^{N-1}\left|\sum_{m=0}^{M-1}\mathcal{D}_{m,n}\right|$ 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.

Considered audio signal alteration effects

Effect name (# of classes)	Profiles	#
Dynamic range control (7)	no compression (linear instantaneous mix)	1
	reference studio mix	1
	dynamic range compression (SoX [4])	5
Spatialization (5)	reference studio mix	1
	monophonic mix	1
	amplitude panning	4
	phase panning	4
	HRTF (CIPIC database [3])	4
Lossy compression (5)	uncompressed WAV file	1
	MP3 compression (LAME encoder [5])	4
Content alteration (10)	resampling	5
	addition of a white Gaussian noise	5

Each effect is simulated through signal transformations.

- Dynamic range is controlled through SoX compander with typical settings (5) for music, speech and streaming.
- Spatialization effects are directly implemented for (4) different directions of arrival randomly chosen in range $[-\frac{\pi}{2};\frac{\pi}{2}]$.
- Lossy compression is completed with LAME MP3 encoder at different bitrates (4).
- Content alteration is completed for (10) different configurations.

Numerical results

Audio signal alterations detection

Supervised 3-fold cross-validation:

- 27 distinct alteration effects are applied on each of the 122 Medley DB [1] tracks, for which separated stems and artist mix are available.
- Comparison of 3 supervised classification methods.
- Best accuracy results are reached by LDA or SVM: dynamic range control (71%), spatialization (98%), lossy compression (88%) and content alteration (88%).

Method	Dynamic range control class name							**Accuracy**
	no comp.	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	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	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	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

Unsupervised classification:

- Classification of the altered signal in the feature space, using k-means through the city-block (Manhattan) distance.
- Feature selection through the IRMFSP [6] algorithm (features sorted by descending order of the Fisher Score (FS))
- Performances measured in term of cluster purity for the optimal number of features for an expected number of clusters (denoted K), for each classification problem.

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

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

Decade prediction

- **Paradigm:** the processing chain applied in professional studio has evolved over the years and can thus be described by the proposed descriptors.
- **Materials:** 1980 music tracks previously used in [2].
- **Experiment:** supervised 3-fold cross-validation classification with artist filtering.
- **Results:** 63% of accuracy reached by SVM (using radial basis function kernel).

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**

Conclusion and future works

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 with consideration to the Human perception of the audio quality.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] SoX - Sound eXchange: http://sox.sourceforge.net/
- [5] LAME encoder: http://lame.sourceforge.net/
- [6] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization", in 115th AES Convention, NY, USA, Oct. 2003.

Horizon 2020 European Union funding for Research & Innovation

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement n° 688122, as part of the ABC-DJ project (http://www.abcdj.eu)

Artist-to-Business-to-Business-to-Consumer Audio Branding System