WHAT IS THE EFFECT OF AUDIO QUALITY ON THE ROBUSTNESS OF MFCCs AND CHROMA FEATURES?

Julián Urbano, Dmitry Bogdanov, Perfecto Herrera, Emilia Gómez and Xavier Serra
Music Technology Group, Universitat Pompeu Fabra Barcelona, Spain
{julian.urbano,dmitry.bogdanov,perfecto.herrera,emilia.gomez,xavier.serra}@upf.edu

ABSTRACT

Music Information Retrieval is largely based on descriptors computed from audio signals, and in many practical applications they are to be computed on music corpora containing audio files encoded in a variety of lossy formats. Such encodings distort the original signal and therefore may affect the computation of descriptors. This raises the question of the robustness of these descriptors across various audio encodings. We examine this assumption for the case of MFCCs and chroma features. In particular, we analyze their robustness to sampling rate, codec, bitrate, frame size and music genre. Using two different audio analysis tools over a diverse collection of music tracks, we compute several statistics to quantify the robustness of the resulting descriptors, and then estimate the practical effects for a sample task like genre classification.

1. INTRODUCTION

A significant amount of research in Music Information Retrieval (MIR) is based on descriptors computed from audio signals. In many cases, research corpora contain music files encoded in a lossless format. In some situations, datasets are distributed without their original music corpus, so researchers have to gather audio files themselves. In many other cases, audio descriptors are distributed instead of the audio files. In the end, MIR research is thus based on corpora that very well may use different audio encodings, all under the assumption that audio descriptors are robust to these variations and the final MIR algorithms are not affected. This possible lack of robustness poses serious questions regarding the reproducibility of MIR research and its applicability. For instance, whether algorithms trained with lossless audio files can generalize to lossy encodings; or whether a minimum audio bitrate should be required in datasets that distribute descriptors instead of audio files.

In this paper we examine the assumption of robustness of music descriptors across different audio encodings on the example of Mel-frequency cepstral coefficients (MFCCs) and chroma features. They are among the most popular music descriptors used in MIR research, as they respectively capture timbre and tonal information. Many MIR tasks such as classification, similarity, autotagging, recommendation, cover identification and audio fingerprinting, audio-to-score alignment, audio segmentation, key and chord estimation, and instrument detection are at least partially based on them. As they pervade the literature on MIR, we analyzed the effect of audio encoding and signal analysis parameters on the robustness of MFCCs and chroma. To this end, we run two different audio analysis tools over a diverse collection of 400 music tracks. We then compute several indicators that quantify the robustness and stability of the resulting features and estimate the practical implications for a general task like genre classification.

2. DESCRIPTORS

2.1 Mel-Frequency Cepstrum Coefficients

MFCCs are inherited from the speech domain [18], and they have been extensively used to summarize the spectral content of music signals within an analysis frame. MFCCs are widely used in tasks like music similarity [1,12], music classification [6] (in particular, genre), autotagging [13], preference learning for music recommendation [19, 24], cover identification and audio segmentation [17].

There is no standard algorithm to compute MFCCs, and a number of variants have been proposed [8] and adapted for MIR applications. MFCCs are commonly computed as follows. The first step consists in windowing the input signal and computing its magnitude spectrum with the Fourier transform. We then apply a filterbank with critical (mel) band spacing of the filters and bandwidths. Energy values are obtained for the output of each filter, followed by a logarithm transformation. We finally compute a discrete cosine transform to the set of log-energy values to obtain the final set of coefficients. The number of mel bands and the frequency interval on which they are computed may vary among implementations. The low order coefficients account for the slowly changing spectral envelope, while the higher order coefficients describe the fast variations of the spectrum shape, including pitch information. The first coefficient is typically discarded in MIR applications because it does not provide information about the spectral shape; it reflects the overall energy in mel bands.

2.2 Chroma

Chroma features represent the spectral energy distribution within an analysis frame, summarized into 12 semitones across octaves in equal-tempered scale. Chroma captures the pitch class distribution of an input signal, typically used

for key and chord estimation [7, 9], music similarity and cover identification [20], classification [6], segmentation and summarization [5, 17], and synchronization [16].

Several approaches exist for chroma feature extraction, including the following steps. The signal is first analyzed with a high frequency resolution in order to obtain its frequency domain representation. The main frequency components (e.g. spectral peaks) are mapped onto pitch classes according to an estimated tuning frequency. For most approaches, a frequency value partially contributes to a set of "sub-harmonic" fundamental frequency (pitch) candidates. The chroma vector is computed with a given interval resolution (number of bins per octave) and is finally post-processed to obtain the final chroma representation. Timbre invariance is achieved by different transformations such as spectral whitening [9] or cepstrum liftering [15].

3. EXPERIMENTAL DESIGN

3.1 Factors Affecting Robustness

We identified several factors that could have an effect on the robustness of audio descriptors, from the perspective of their audio encoding (codec, bitrate and sampling rate), analysis parameters (frame/hop size and audio analysis tool) and the musical characteristics of the songs (genre).

SRate. The sampling rate at which an audio signal is encoded may affect robustness when using very high frequency rates. We study standard 44100 and 22050 Hz.

Codec. Perceptual audio coders may also affect descriptors because they introduce perturbations to the original audio signal, in particular by reducing high-frequency content, blurring the attacks, and smoothing the spectral envelope. In our experiments, we chose one lossless and two lossy audio codecs: WAV, MP3 CBR and MP3 VBR.

BRate. Different audio codecs allow different bitrates depending on the sampling rate, so we can not combine all codecs with all bitrates. The following combinations are permitted and used in our study:

- WAV: 1411 Kbps.
- MP3 CBR at 22050 Hz: 64, 96, 128 and 160 Kbps.
- MP3 CBR at 44100 Hz: 64, 96, 128, 160, 192, 256 and 320 Kbps.
- MP3 VBR: 6 (100-130 Kbps), 4 (140-185 Kbps), 2 (170-210 Kbps) and 0 (220-260 Kbps).

FSize. We considered a variety of frame sizes for spectral analysis: 23.2, 46.4, 92.9, 185.8, 371.5 and 743.0 ms. That is, we used frame sizes of 1024, 2048, 4096, 8192, 16384 and 32768 samples for signals with sampling rate of 44100 Hz, and the halved values (512, 1024, 2048, 4096, 8192 and 16384 samples) in the case of 22050 Hz.

Audio analysis tool. The specific software used to compute descriptors may have an effect on their robustness due to parameterizations (e.g. frequency ranges) and other implementation details. We use two state-of-the-art and open source tools publicly available online: *Essentia 2.0.1* [1] [2] and *QM Vamp Plugins 1.7 for Sonic Annotator 0.7* [2] [3]. Since our goal here is not to compare tools, we refer to them simply as Lib1 and Lib2 throughout the paper.
Lib1 and Lib2 provide by default two different implementations of MFCCs, both of which compute cepstral coefficients on 40 mel bands, resembling the MFCC FB-40 implementation [8, 22] but on different frequency intervals. Lib1 covers a wider frequency range of 0-11000 Hz with mel bin centers being equally spaced on the mel scale in this range, while Lib2 covers a frequency range of 66-6364 Hz. We compute the first 13 MFCCs in both systems and discard the first coefficient. In the case of chroma, Lib1 analyzes a frequency range of 40-5000 Hz based on Fourier transform and estimates tuning frequency. Lib2 uses a Constant Q Transform and analyzes the frequency range 65-2093 Hz assuming tuning frequency of 440 Hz, but it does not account for harmonics of the detected peaks. We compute 12-dimensional chroma features.

Genre. Robustness may depend as well on the music genre of songs. For instance, as the most dramatic change that perceptual coders introduce is that of filtering out high-frequency spectral content, genres that make use of very high-frequency sounds (e.g. cymbals and electronic tones) should show a more detrimental effect than genres not including them (e.g. country, blues and classical).

3.2 Data

We created an ad-hoc corpus of music for this study, containing 400 different music tracks (30 seconds excerpts) by 395 different artists, uniformly covering 10 music genres (blues, classical, country, disco/funk/soul, electronic, jazz, rap/hip-hop, reggae, rock and rock'n'roll). All 400 tracks are encoded from their original CD at a 44100 Hz sampling rate using the lossless FLAC audio codec.

We converted all lossless tracks in our corpus into various audio formats in accordance with the factors identified above, taking into account all possible combinations of sampling rate, codec and bitrate. Audio conversion was done using the *FFmpeg 0.8.3* [3] converter, which includes the LAME codec for MP3 joint stereo mode (*Lavf53.21.1*). Afterwards, we analyzed the original lossless files and their lossy versions using both Lib1 and Lib2. In the case of Lib1, both MFCCs and chroma features were computed for all different frame sizes with the hop size equal to half the frame size. MFCCs were computed similarly in the case of Lib2, but chroma features only allow a fixed frame size of 16384 samples (we selected a hop size of 2048 samples). In all cases, we summarize the frame-wise feature vectors with the mean of each coefficient.

3.3 Indicators of Robustness

We computed several indicators of the robustness of MFCCs and chroma, each measuring the difference between the descriptors computed with the original lossless audio clips and the descriptors computed with their lossy versions. We blocked by tool, sampling rate and frame size under the assumption that these factors are not mixed in practice within the same application. For two arbitrary

[1] http://essentia.upf.edu

[2] http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html

[3] http://www.ffmpeg.org

vectors x and y (each containing $n = 12$ MFCC or chroma values) from a lossless and a lossy version, we compute five indicators to measure how different they are.

Relative error δ. It is computed as the average relative difference across coefficients. This indicator can be easily interpreted as the percentage error between coefficients, and it is of especial interest for tasks in which coefficients are used as features to train some model.

$$\delta(x, y) = \frac{1}{n} \sum \frac{|x_i - y_i|}{\max(|x_i|, |y_i|)}$$

Euclidean distance ε. The Euclidean distance between the two vectors, which is especially relevant for tasks that compute distances between pairs of songs, such as in music similarity or other tasks that use techniques like clustering.

Pearson's r. The common parametric correlation coefficient between the two vectors, ranging from -1 to 1.

Spearman's ρ. A non-parametric correlation coefficient, equal to the Pearson's r correlation after transforming all coefficients to their corresponding ranks in $x \cup y$.

Cosine similarity θ. The angle between both vectors. It is is similar to ε, but it is normalized between 0 and 1.

We have 400 tracks$\times$19 *BRate:Codec*$\times$6 *FSize*=45600 datapoints for MFCCs with Lib1, MFCCs with Lib2, and chroma with Lib1. For chroma with Lib2 there is just one *FSize*, which yields 7600 datapoints. This adds up to 144400 datapoints for each indicators, 722000 overall.

3.4 Analysis

For simplicity, we followed a hierarchical analysis for each combination of sampling rate, tool, feature and robustness indicator. We are first interested in the mean of the score distributions, which tells us the expected *robustness* in each case (e.g. a low ε mean score suggests that the descriptor is robust because it does not differ much between the lossless and the lossy versions). But we are also interested in the *stability* of the descriptor, that is, the variance of the distribution. For instance, a descriptor might be robust on average but not below 192 Kbps, or robust only with a frame size of 2048.

To gain a deeper understanding of the variations in the indicators, we fitted a random effects model to study the effects of codec, bitrate and frame size [14]. The specific models included the *FSize* and *Codec* main effects, and the bitrate was modeled as nested within the *Codec* effect (*BRate:Codec*); all interactions among them were also fitted. Finally, we included the *Genre* and *Track* main effects to estimate the specific variability due to inherent differences among the music pieces themselves. We did not consider any *Genre* or *Track* interactions because they can not be controlled in a real-world application, so their effects are all confounded with the *residual* effect. Note though that this residual does not account for any random error (in fact, there is no random error in this model); it accounts for high-order interactions associated with *Genre* and *Track* that are irrelevant for our purposes. This results in a Resolution V design for the factors of interest (main effects unconfounded with two- or three-factor interactions) and a Resolution III design for musical factors related to genre (main effects confounded with two-factor interactions) [14]. We ran an ANOVA analysis on these models to estimate variance components, which indicate the contribution of each factor to the total variance, that is, their impact on the robustness of the audio descriptors.

4. RESULTS

Table 1 shows the results for MFCCs. As shown by the mean scores, the descriptors computed by Lib1 and Lib2 are similarly robust (note that ε scores are not directly comparable across tools because they are not normalized; actual MFCCs in Lib1 are orders of magnitude larger than in Lib2). Both correlation coefficients r and ρ, as well as cosine similarity θ, are extremely high, indicating that the shape of the feature vectors is largely preserved. However, the average error across coefficients is as high $\delta \approx 6.1\%$ at 22050 Hz and $\delta \approx 6.7\%$ at 44100 Hz.

When focusing on the stability of the descriptors, we see that the implementation in Lib2 is generally more stable because the distributions have less variance, except for δ and ρ at 22050 Hz. The decomposition in variance components indicates that the choice of frame size is irrelevant in general (low $\hat{\sigma}^2_{FSize}$ scores), and that the largest part of the variability depends on the particular characteristics of the music pieces (very high $\hat{\sigma}^2_{Track} + \hat{\sigma}^2_{residual}$ scores). For Lib2 in particular, this means that controlling encodings or analysis parameters does not increase robustness significantly when the sampling rate is 22050 Hz; it depends almost exclusively on the specific music pieces. On the other hand, the combination of codec and bitrate has a quite large effect in Lib1. For instance, about 42% of the variability in Euclidean distances is due to the *BRate:Codec* interaction effect. This means that an appropriate selection of the codec and bitrate of the audio files leads to significantly more robust descriptors. At 44100 Hz both tools are clearly affected by the *BRate:Codec* effect as well, especially Lib1. Figure 1 compares the distributions of δ scores for each tool. We can see that Lib1 has indeed large variance across groups, but small variance within groups, as opposed to Lib2. The robustness of Lib1 seems to converge to $\delta \approx 3\%$ at 256 Kbps, and the descriptors are clearly more stable with larger bitrates (smaller within-group variance). On the other hand, the average robustness of Lib2 converges to $\delta \approx 5\%$ at 160-192 Kbps, and stabil-

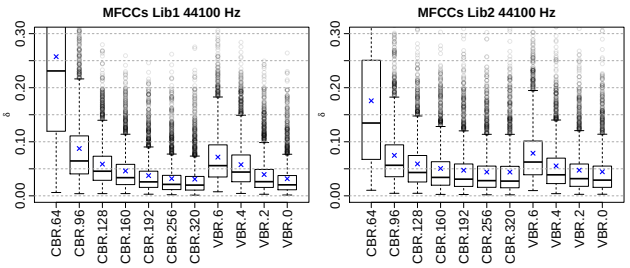

Figure 1. Distributions of δ scores for different combinations of MP3 codec and bitrate at 44100 Hz, and for both audio analysis tools. Blue crosses mark the sample means. Outliers are rather uniformly distributed across genres.

		22050 Hz					44100 Hz				
		δ	ε	r	ρ	θ	δ	ε	r	ρ	θ
Lib1	$\hat{\sigma}^2_{FSize}$	1.08	3.03	1.73	0	1.74	0.21	0.09	0.01	0	0
	$\hat{\sigma}^2_{Codec}$	0	0	0	0	0	0	0	0	0	0
	$\hat{\sigma}^2_{BRate:Codec}$	31.25	42.13	21.61	8.38	21.49	46.98	41.77	22.52	24.03	21.51
	$\hat{\sigma}^2_{FSize\times Codec}$	0	0	0	0	0	0	0.20	0.07	0.05	0.06
	$\hat{\sigma}^2_{FSize\times(BRate:Codec)}$	4.87	11.71	12.36	1.23	13.21	7.37	18.25	17.98	10.85	18.02
	$\hat{\sigma}^2_{Genre}$	0.99	4.53	3.92	0.08	3.80	1.12	0.52	0.90	0.32	0.89
	$\hat{\sigma}^2_{Track}$	19.76	5.84	6.46	11.59	5.73	10.12	3.91	2.65	5.23	2.59
	$\hat{\sigma}^2_{residual}$	42.05	32.75	53.92	78.72	54.03	34.19	35.26	55.87	59.52	56.92
	Grand mean	0.0591	1.6958	0.9999	0.9977	0.9999	0.0682	1.8820	0.9998	0.9939	0.9998
	Total variance	0.0032	3.4641	1.8e-7	3.2e-5	1.5e-7	0.0081	11.44	1.6e-6	0.0005	1.4e-6
	Standard deviation	0.0567	1.8612	0.0004	0.0056	0.0004	0.0897	3.3835	0.0013	0.0214	0.0012
Lib2	$\hat{\sigma}^2_{FSize}$	1.17	0.32	0.16	0.24	0.18	0.25	0	0	0	0
	$\hat{\sigma}^2_{Codec}$	0	0	0	0	0	0	0	0	0	0
	$\hat{\sigma}^2_{BRate:Codec}$	4.91	6.01	2.32	0.74	3.14	23.46	24.23	14.27	13.31	15.02
	$\hat{\sigma}^2_{FSize\times Codec}$	0	0	0	0	0	0	0	0	0	0
	$\hat{\sigma}^2_{FSize\times(BRate:Codec)}$	0.96	0.43	0.03	0.04	0.09	7.17	8.09	10.35	6.34	10.86
	$\hat{\sigma}^2_{Genre}$	4.21	14.68	2.84	0.61	4.41	0.37	5.37	0.50	0	0.48
	$\hat{\sigma}^2_{Track}$	52.34	61.05	32.07	66.10	41.26	27.33	14.10	6.55	13.32	5.53
	$\hat{\sigma}^2_{residual}$	36.41	17.51	62.57	32.27	50.92	41.42	48.21	68.32	67.03	68.11
	Grand mean	0.0622	0.0278	0.9999	0.9955	0.9999	0.0656	0.0342	0.9998	0.9947	0.9999
	Total variance	0.0040	0.0015	8.9e-8	0.0002	3.5e-8	0.0055	0.0034	6.4e-7	0.0002	4.8e-7
	Standard deviation	0.0631	0.0391	0.0003	0.0131	0.0002	0.0740	0.0587	0.0008	0.0150	0.0007

Table 1. Variance components in the distributions of robustness of MFCCs for Lib1 (top) and Lib2 (bottom). Each component represents the percentage of total variance due to each effect (eg. $\hat{\sigma}^2_{FSize} = 3.03$ indicates that 3.03% of the variability in the robustness indicator is due to differences across frame sizes; $\hat{\sigma}^2_x = 0$ when the effect is so extremely small that the estimate is slightly below zero). All interactions with the *Genre* and *Track* main effects are confounded with the *residual* effect. The last rows show the grand mean, total variance and standard deviation of the distributions.

ity remains virtually the same beyond 96 Kbps. These plots confirm that the MFCC implementation in Lib1 is nearly twice as robust and stable when the encoding is homogeneous in the corpus, while the implementation in Lib2 is less robust but more stable with heterogeneous encodings.

The *FSize* effect is negligible, indicating that the choice of frame size does not affect the robustness of MFCCs in general. However, in several cases we can observe large $\hat{\sigma}^2_{FSize\times(BRate:Codec)}$ scores, meaning that for some codec-bitrate combinations it does matter. An in-depth analysis shows that these differences only occur at 64 Kbps though (small frame sizes are more robust); differences are very small otherwise. Finally, the small $\hat{\sigma}^2_{Genre}$ scores indicate that robustness is similar across music genres.

A similar analysis was conducted to assess the robustness and stability of chroma features. Even though the correlation indicators are generally high as well, Table 2 shows that chroma vectors do not preserve the shape as well as MFCCs do. When looking at individual coefficients, the relative errors are similarly $\delta \approx 6\%$ in Lib1, but they are greatly reduced in Lib2, especially at 44100 Hz. In fact, the chroma implementation in Lib2 is more robust and stable according to all indicators [4]. For Lib1, virtually all the variability in the distributions is due to the *Track* and *residual* effects, meaning that chroma is similarly robust across encodings, analysis parameters and genre. For Lib2, we can similarly observe that errors in the correlation indicators depend almost entirely on the *Track* effect, but δ and ε depend mostly on the codec-bitrate combination. This indicates that, despite chroma vectors preserve their shape, the individual components vary significantly across encodings; we observed that increasing the bitrate leads to larger coefficients overall. This suggests that normalizing the chroma coefficients could dramatically improve the distributions of δ and ε. We tried the parameter `normalization=2` to have Lib2 normalize chroma vectors to unit maximum. As expected, the effects of codec and bitrate are removed after normalization, and most of the variability is due to the *Track* effect. The correlation indicators are practically unaltered after normalization.

[4] Even though these distributions include all frame sizes in Lib1 but only 16384 in Lib2, the *FSize* effect is negligible in Lib1, meaning that these indicators are still comparable across implementations

5. ROBUSTNESS IN GENRE CLASSIFICATION

The previous section provided indicators of robustness that can be easily understood. However, they can be hard to interpret because in the end we are interested in the robustness of the various algorithms that make use of these features; whether $\delta = 5\%$ is large or not depends on how MFCCs and chroma are used in practice. To investigate this question we consider a music genre classification task. For each sampling rate, codec, bitrate and tool we trained one SVM model with radial basis kernel using MFCCs and another using chroma. For MFCCs we used a standard frame size of 2048, and for chroma we set 4096 in Lib1 and the fixed 16384 in Lib2. We did random sub-sampling validation with 100 random trials for each model, using 320 tracks for training and the remaining 80 for testing.

We first investigate whether a particular choice of encoding is likely to classify better when fixed across training and test sets. Table 3 shows the results for a selection of encodings at 44100 Hz. Within the same tool and descriptor, differences across encodings are quite small, approximately 0.02. In particular, for MFCCs and Lib1 an ANOVA analysis suggests that differences are signifi-

		22050 Hz					44100 Hz				
		δ	ε	r	ρ	θ	δ	ε	r	ρ	θ
Lib1	$\hat{\sigma}^2_{FSize}$	1.68	2.77	0.20	0.15	0.38	2.37	2.42	0.24	0.34	0.50
	$\hat{\sigma}^2_{Genre}$	2.81	2.75	1.29	1.47	0.81	3.12	2.61	1.17	1.25	0.85
	$\hat{\sigma}^2_{Track}$	20.69	19.27	17.75	18.52	16.63	22.28	20.78	18.81	19.92	18.64
	$\hat{\sigma}^2_{residual}$	74.82	75.21	80.75	79.86	82.17	72.22	74.19	79.79	78.49	80.01
	Grand Mean	0.0610	0.0545	0.9554	0.9366	0.9920	0.0588	0.0521	0.9549	0.9375	0.9922
	Total variance	0.0046	0.0085	0.0276	0.0293	0.0014	0.0048	0.0082	0.0286	0.0298	0.0013
	Standard deviation	0.0682	0.0924	0.1663	0.1713	0.0373	0.0695	0.0904	0.1691	0.1725	0.0355
Lib2	$\hat{\sigma}^2_{Codec}$	63.62	34.55	0	0	0	32.32	21.59	0	0	0
	$\hat{\sigma}^2_{BRate:Codec}$	0.71	0.23	0	0	0	61.80	39.51	0.01	0.03	0.04
	$\hat{\sigma}^2_{Genre}$	0.25	15.87	2.90	4.05	7.95	0.62	9.98	3.43	1.33	3.66
	$\hat{\sigma}^2_{Track}$	19.29	32.77	96.71	92.75	91.80	3.27	13.79	94.24	93.04	77.00
	$\hat{\sigma}^2_{residual}$	16.14	16.58	0.38	3.20	0.25	1.98	15.13	2.32	5.60	19.30
	Grand mean	0.0346	0.0031	0.9915	0.9766	0.9998	2.6e-2	2.2e-3	0.9989	0.9928	1
	Total variance	0.0004	5e-6	0.0002	0.0007	6.1e-8	4.6e-4	4.8e-6	3.7e-6	0.0001	1.8e-9
	Standard deviation	0.0195	0.0022	0.0135	0.0270	0.0002	0.0213	0.0022	0.0019	0.0122	4.2e-5

Table 2. Variance components in the distributions of robustness of Chroma for Lib1 (top) and Lib2 (bottom), similar to Table 1. The *Codec* main effect and all its interactions are not shown for Lib1 because all variance components are estimated as 0. Note that the *FSize* main effect and all its interactions are omitted for Lib2 because it is fixed to 16384.

		64	96	128	160	192	256	320	WAV
Lib1	MFCCs	.383	.384	.401	.403	.395	.402	.394	.393
	Chroma	.275	.281	.288	.261	.278	.278	.284	.291
Lib2	MFCCs	.335	.329	.332	.341	.336	.336	.344	.335
	Chroma	.320	.325	.320	.323	.325	.319	.320	.313

Table 3. Mean classification accuracy over 100 trials when training and testing with the same encoding (MP3 CBR and WAV only) at 44100 Hz.

cant, $F(7, 693) = 2.34, p = 0.023$; a multiple comparisons analysis reveals that 64 Kbps is significantly worse than the best (160 Kbps). In terms of chroma, differences are again statistically significant, $F(7, 693) = 3.71, p < 0.001$; 160 Kbps is this time significantly worse that most of the others. With Lib2 differences are not significant for MFCCs, $F(7, 693) = 1.07, p = 0.378$. No difference is found for chroma either, $F(7, 693) = 0.67, p = 0.702$. Overall, despite some pairwise comparisons are significantly different, there is no particular encoding that clearly outperforms the others; the observed differences are probably just Type I errors. There is no clear correlation either between bitrate and accuracy.

We then investigate whether a particular choice of encoding for training is likely to produce better results when the target test set has a fixed encoding. For MFCCs and Lib1 there is no significant difference in any but one case (testing with 160 Kbps is worst when training with 64 Kbps). For chroma there are a few cases where 160 Kbps is again significantly worse than others, but we attribute these to Type I errors as well. Although not significantly so, the best result is always obtained when the training set has the same encoding as the target test set. With Lib2 there is no significant difference for MFCCs or chroma. Overall, we do not observe a correlation either between training and test encodings. Due to space constrains, we do not discuss results for VBR or 22050 Hz, but the same general conclusions can be drawn nonetheless.

6. DISCUSSION

Sigurdsson et al. [21] suggested that MFCCs are sensitive to the spectral perturbations that result from low bitrate compression, mostly due to distortions at high frequencies. They estimated squared Pearson's correlation between MFCCs computed on original lossless audio and its MP3 derivatives, using 4 different MFCC implementations. All implementations were found to be robust at bitrates of at least 128 Kbps, with $r^2 > 0.95$, but a significant loss in robustness was observed at 64 Kbps in some of the implementations. The most robust MFCC implementation had a highest frequency of 4600 Hz, while the least robust implementation included frequencies up to 11025 Hz. Their music corpus contained only 46 songs though, clearly limiting their results. In our experiments, all encodings show $r^2 > 0.99$. However, we note that Pearson's r is very sensible to outliers with such small samples. This is the case of the first MFCC coefficients, which are orders of magnitude larger than the last coefficients. This makes r extremely large simply because the first coefficients are remotely similar; most of the variability between feature vectors is explained because of the first coefficient. This is clear in our Table 1, where $r \approx 1$ and variance is nearly 0. To minimize this sensibility to outliers, we also included the non-parametric Spearman's ρ correlation coefficient as well as the cosine similarity. In our case, the tool with the larger frequency range was shown to be more robust under homogeneous encodings, while the shorter range was more stable under heterogeneous conditions.

Hamawaki et al. [10] analyzed differences in the distribution of MFCCs for different bitrates using a corpus of 2513 MP3 files of Japanese and Korean pop songs with bitrates between 96 and 192 Kbps. Following a music similarity task, they compared differences in the top-10 ranked results when using MFCCs derived from WAV audio, its MP3 encoded versions, and the mixture of MFCCs from different sources. They found that the correlation of the results deteriorates smoothly as the bitrate decreases, while ranking on a set of MFCCs derived from different formats revealed uncorrelated results. We similarly observed that the differences between MFCCs of the original WAV files and its MP3 versions decrease smoothly with bitrate.

Jensen et al. [12] measured the effect of audio encoding on performance of an instrument classifier using MFCCs.

They compared MFCCs computed from MP3 files at only 32-64 Kbps, observing a decrease in performance when using a different encoder for training and test sets. In contrast, performance did not change significantly when using the same encoder. For genre classification with MFCCs, our results showed no differences in either case. We note though that the bitrates we considered are much larger. Uemura et al. [23] examined the effect of bitrate on chord recognition using chroma features with an SVM classifier. They observed no obvious correlation between encoding and estimation results; the best results were even obtained with very low bitrates for some codecs. Our results on genre classification with chroma largely agree in this case as well; the best results with Lib2 were also obtained by low bitrates. Casey et al. [4] evaluated the effect of lossy encodings on genre classification tasks using audio spectrum projection features. They found a small but statistically significant decrease in accuracy for bitrates of 32 and 96 Kbps. In our experiments, we do not observe these differences, although the lowest bitrate we consider is 64 Kbps. Jacobson et al. [11] also investigated the robustness of onset detection methods to lossy MP3 encoding. They found statistically significant changes in accuracy only at bitrates lower than 32 Kbps.

Our results showed that MFCCs and chroma features, as computed by Lib1 and Lib2, are generally robust and stable within reasonable limits. Some differences have been noted between tools though, largely attributable to the different frequency ranges they employ. Nonetheless, it is evident that certain combinations of codec and bitrate may require a re-parameterization of some descriptors to improve or even maintain robustness. In practice, these parameterizations affect the performance and applicability of algorithms, so a balance between performance, robustness and generalizability should be sought. These considerations are of major importance when collecting audio files for some dataset, as a minimum audio quality might be needed for some descriptors.

7. CONCLUSIONS

In this paper we have studied the robustness of two common audio descriptors used in Music Information Retrieval, namely MFCCs and chroma, to different audio encodings and analysis parameters. Using a varied corpora of music pieces and two different audio analysis tools we have confirmed that MFFCs are robust to frame/hop sizes and lossy encoding provided that a minimum bitrate of approximately 160 Kbps is used. Chroma features were shown to be even more robust, as the codec and bitrates had virtually no effect on the computed descriptors. This is somewhat expected given that chroma does not capture information as fine-grained as MFCCs do, and that lossy compression does not alter the perceived tonality. We did find subtle differences between implementations of these audio features, which call for further research on standardizing algorithms and parameterizations to maximize their robustness while maintaining their effectiveness in the various tasks they are used in. The immediate line for future work includes the analysis of other features and tools.

8. ACKNOWLEDGMENTS

This work is partially supported by an A4U postdoctoral grant and projects SIGMUS (TIN2012-36650), CompMusic (ERC 267583), PHENICX (ICT-2011.8.2) and GiantSteps (ICT-2013-10).

9. REFERENCES

[1] J.J. Aucouturier, F. Pachet, and M. Sandler. "The way it sounds": timbre models for analysis and retrieval of music signals. *IEEE Trans. Multimedia*, 2005.

[2] D. Bogdanov, N. Wack, et al. ESSENTIA: an audio analysis library for music information retrieval. In *ISMIR*, 2013.

[3] C. Cannam, M.O. Jewell, C. Rhodes, M. Sandler, and M. d'Inverno. Linked data and you: bringing music research software into the semantic web. *J. New Music Res.*, 2010.

[4] M. Casey, B. Fields, et al. The effects of lossy audio encoding on genre classification tasks. In *AES*, 2008.

[5] W. Chai. Semantic segmentation and summarization of music: methods based on tonality and recurrent structure. *IEEE Signal Processing Magazine*, 2006.

[6] D. Ellis. Classifying music audio with timbral and chroma features. In *ISMIR*, 2007.

[7] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *ICMC*, 1999.

[8] T. Ganchev, N. Fakotakis, and G. Kokkinakis. Comparative evaluation of various MFCC implementations on the speaker verification task. In *SPECOM*, 2005.

[9] E. Gómez. *Tonal description of music audio signals*. PhD thesis, Universitat Pompeu Fabra, 2006.

[10] S. Hamawaki, S. Funasawa, et al. Feature analysis and normalization approach for robust content-based music retrieval to encoded audio with different bit rates. In *MMM*, 2008.

[11] K. Jacobson, M. Davies, and M. Sandler. The effects of lossy audio encoding on onset detection tasks. In *AES*, 2008.

[12] J.H. Jensen, M.G. Christensen, D. Ellis, and S.H. Jensen. Quantitative analysis of a common audio similarity measure. *IEEE TASLP*, 2009.

[13] B. McFee, L. Barrington, and G. Lanckriet. Learning content similarity for music recommendation. *IEEE TASLP*, 2012.

[14] D.C. Montgomery. *Design and Analysis of Experiments*. Wiley & Sons, 2009.

[15] M. Müller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE TASLP*, 2010.

[16] M. Müller, H. Mattes, and F. Kurth. An efficient multiscale approach to audio synchronization. In *ISMIR*, 2006.

[17] J. Paulus, M. Müller, and A Klapuri. Audio-based music structure analysis. In *ISMIR*, 2010.

[18] L.R. Rabiner and R.W. Schafer. *Introduction to Digital Speech Processing*. Foundations and Trends in Signal Processing. 2007.

[19] J. Reed and C. Lee. Preference music ratings prediction using tokenization and minimum classification error training. *IEEE TASLP*, 2011.

[20] J. Serrà, E. Gómez, and P. Herrera. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In Z. Raś and A.A. Wieczorkowska, editors, *Advances in Music Information Retrieval*. Springer, 2010.

[21] S. Sigurdsson, K.B. Petersen, and T. Lehn-Schiler. Mel Frequency Cepstral Coefficients: an evaluation of robustness of MP3 encoded music. In *ISMIR*, 2006.

[22] M. Slaney. Auditory toolbox. *Interval Research Corporation, Technical Report*, 1998. `http://engineering.purdue.edu/~malcolm/interval/1998-010/`.

[23] A. Uemura, K. Ishikura, and J. Katto. Effects of audio compression on chord recognition. In *MMM*, 2014.

[24] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and HG. Okuno. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE TASLP*, 2008.