A Performance Evaluation of the Timbre Toolbox and the MIRtoolbox on Calibrated Test Sounds

Savvas Kazazis,[†] *Nicholas Esterer, Philippe Depalle, Stephen McAdams*

Schulich School of Music, McGill University
[†]savvas.kazazis@mail.mcgill.ca

ABSTRACT

We evaluate the accuracy of the Timbre Toolbox (v. 1.2) and the MIRtoolbox (v. 1.6.1) on audio descriptors that are putatively related to timbre. First, we report and fix major bugs found in the current version of the Timbre Toolbox, which have gone previously unnoticed in publications that used this toolbox as an analysis tool. Then, we construct sound sets that exhibit specific spectral and temporal characteristics in relation to the descriptors being tested. The evaluation is performed by comparing the theoretical (real) values of the sound sets to the estimations of the toolboxes.

1. INTRODUCTION

The Timbre Toolbox [1] and the MIRtoolbox [2] are two of the most popular MATLAB [3] toolboxes that are used for audio feature extraction within the music information retrieval (MIR) community. They have been recently evaluated according to the number of presented features, the user interface and computational efficiency [4], but there have not been performance evaluations of the accuracy of the extracted features. The aim of this paper is: (1) to detect and summarize the bugs in the current version of the Timbre Toolbox and (2) to evaluate the robustness of audio descriptors these toolboxes have in common and that are putatively related to timbre. For this purpose, we synthesized various sound sets using additive synthesis, calculated the theoretical (real) values of each descriptor tested, and compared these values with the estimations of the toolboxes. Section 2 summarizes the bugs found in the current publically available version of the Timbre Toolbox (v. 1.2). Section 3 describes the construction of the sound sets used for evaluating the performance of the MIRtoolbox (v. 1.6.1) and a beta version of the Timbre Toolbox, which fixes the reported bugs. Section 4 presents the results of the evaluation and Section 5 summarizes our findings.

2. POINTS OF CONSIDERATION AND BUG FIXING IN THE TIMBRE TOOLBOX

In this section, we report the bugs found in the current version of the Timbre Toolbox and some issues related to user interaction. The Timbre Toolbox incorporates the following sound models of the time-domain signal for extracting audio descriptors: the temporal energy envelope; the short-term Fourier transform (STFT) on a linear amplitude scale (STFTmag) and a squared amplitude scale (STFTpow); the output of an auditory model based on the concept of the Equivalent Rectangular Bandwidth, which is either calculated using recursive gammatone filters (ERBgam), or their finite impulse response approximation using the fast Fourier transform (ERBfft); and a sinusoidal harmonic model [1].

In some cases, especially when the amplitude of the lower frequencies is lower than the upper ones, the harmonic representation using the default amplitude threshold for detecting harmonics will not analyze even strictly harmonic sounds. Furthermore, the default analysis limit of 20 harmonics could also be problematic for analyzing low-frequency sounds having spectral energy that increases with harmonic number. However, this scenario is very unlikely to occur in natural sounds, but it is still possible with synthetic sounds used in psychoacoustic experiments (e.g., [5]) or in electroacoustic music. Another conceptual bug is the estimation of inharmonicity: according to Eq. 1, which is presented in [1], a signal with a fundamental frequency of 100 Hz and a partial at 150 Hz will be less inharmonic than a signal with the same fundamental and a partial at 190 Hz even though the partial of the second signal is only detuned by 10 Hz below the next harmonic.

$$\text{inharm} = \frac{2}{f_0}\frac{\sum_{h=1}^{\mathrm{H}}(f_h - hf_0)a_h^2}{\sum_{h=1}^{\mathrm{H}} a_h^2} \tag{1}$$

In v. 1.2, the end user only had access to summary statistics, and as such it was not possible to evaluate the time-varying patterns of audio descriptors. Furthermore, the export format of the results was a text file. This did not facilitate further processing of the results especially in the case of a batch analysis where the output consists of several text files. Also, MATLAB ran out of memory when the Timbre Toolbox processed long audio files.

According to Peeters et al. [1], the window that should be used for the harmonic analysis is a Blackman window. However, in the toolbox's implementation, the window is a boxcar (i.e., no window weighting at all), but we also noticed that the removal of the window's energy contribution to the input sound was implemented incorrectly. Furthermore, some calculations on audio descriptors returned the results in normalized frequency (including the spectral centroid) without warning the user and led to misinterpretations (e.g., [6]).

Although the actual sampling rate is read directly from the file, in some sound models it was not actually used: the parameters related to the FFT analysis were specified according to a fixed sampling rate of 44.1 kHz no matter the actual sampling rate of the input file. Finally, in most of the employed sound models, the computations of spectral spread, skewness, kurtosis and spectral slope were implemented incorrectly.

The analysis results presented in this paper are based on a beta version of the Timbre Toolbox that fixes and takes into consideration all of the above-mentioned points except the calculation of inharmonicity and the threshold settings used in the harmonic representation.

3. CONSTRUCTION OF THE TEST SOUND SETS

The sounds were constructed using additive synthesis, which allows for a direct computation of the audio descriptors. Each sound set was designed to exhibit specific sound qualities that are directly related to the descriptors being tested. In this way, we are able to systematically test the performance of the toolboxes by tracking the circumstances under which certain audio descriptors are poorly calculated. All sounds were synthesized at 44.1 kHz with 16-bit resolution and peak amplitude of 6 dB relative to full scale (dBFS). To avoid spectral spread induced by an abrupt onset and offset when performing the FFT on these synthetic sounds, we applied a 10-ms raised inverse cosine ramp to all sounds except the ones used to test the attack time and the attack and decrease slopes. Durations were fixed at 600 ms and all sounds contained harmonics up to (but not including) the Nyquist frequency.

We used the following fundamental frequencies for all the sound sets except those related to the temporal energy envelope: C#1 (34.65 Hz), D2 (73.42 Hz), D#3 (155.56 Hz), C4 (261.63 Hz), E4 (329.63 Hz), F5 (698.46 Hz), A5 (880 Hz), F#6 (1479.98 Hz), G7 (3135.96 Hz) and B7 (3951.07 Hz). The C4 was slightly detuned from 261.63 Hz to 258 Hz in order to match exactly the frequency of an FFT bin and to test whether the estimations would be improved; for a sampling frequency of 44.1 kHz and an FFT size of 1024 samples (default setting of the Timbre Toolbox) the bins are spaced 43 Hz apart. We used such a wide frequency range because as the fundamental frequency increases and approaches the Nyquist limit, the number of "significant" FFT bins decreases, which may affect the accuracy of the results, especially in the presence of noise.

3.1. Attack Time, Attack Slope and Decrease Slope

The Timbre Toolbox uses the "weakest effort method" for estimating the attack time and the attack and decrease slopes [1], whereas the MIRtoolbox uses a similar method based on Gaussian curves [2]. In these adaptive methods, the threshold energy level that the signal must surpass is not fixed, but is determined as a proportion of the maximum of the signal's energy envelope. The attack and decrease slopes are then estimated as the average temporal slope during the start and end times of the attack portion. An 'effort' is defined as the time it takes for the signal to go from one threshold value to the next. It is therefore logical to assume that if the signal varies rapidly and non-linearly during the attack time, the true attack time values may be poorly estimated.

For testing the accuracy of this method, we constructed nine attack envelopes for each of ten logarithmically spaced attack times ranging from 1 to 300 ms. The shape of the envelopes was determined by:

$$y(t) = mt^{\mathrm{b}} \tag{2}$$

where m controls the slope of the attack time and b is a curvature constant which was assigned the following values: 3, 2.5, 2 and 1.5 for an exponential shape; 1 for a linear shape; and 0.67, 0.5, 0.4 and 0.33 for a logarithmic shape. The attack envelopes were then applied to a flat harmonic spectrum with a fundamental frequency of 258 Hz and a total duration of 600 ms. A similar procedure was used for testing the estimations of decrease slope.

3.2. Spectral Centroid

For this sound set, we used a flat spectrum with octave-spaced harmonics and included in the above-mentioned set a lower fundamental of C0 (16.35 Hz). In order to systematically test the accuracy of spectral centroid estimation, we iteratively removed just one harmonic from the initial spectrum up to the last one for every fundamental. This way, the sounds generated from the last fundamental just contain a single frequency component, because there is only one harmonic present due to the Nyquist limit, and therefore the spectral centroid ideally should match the value of the fundamental frequency estimation.

3.3. Spectral Spread, Skewness, Kurtosis and Roll-off

For testing the estimations of spectral spread, skewness, kurtosis, and roll-off, we designed a sound set in which the sounds vary by fundamental frequency and according to spectral slopes. By precisely controlling the spectral slopes, we directly alter in a predictable way the higher statistical moments of the spectrum and the frequency below which 95% of the signal energy is contained. In our analysis, we took into account the fact that the MIRtoolbox uses a default value of 85%. For every fundamental, we constructed a spectrum that contained both odd and even harmonics with a $1/n^2$ power decrease, where n denotes the harmonic number. Then in nine steps we altered linearly the energy distribution of the harmonics until we reached a flat spectrum. The same procedure was repeated by starting from a flat spectrum and reaching in nine steps a positive slope of the harmonics which had an n^2 power increase.

3.4. Harmonic Spectral Deviation and Spectral Irregularity

Spectral deviation (in the Timbre Toolbox) and spectral irregularity (in the MIRtoolbox) are the same descriptors but are computed slight differently with respect to a scaling factor. MIRtoolbox offers two estimation methods based on Jensen [7] and Krimphoff et al. [8]. Here, we only tested the estimation based on Krimphoff's method (Eq. 3.), which is the only option available in the Timbre Toolbox. For every fundamental, we started from a flat spectrum that only contained the fundamental with even harmonics, and we gradually increased the level of the odd ones until we reach a flat spectrum in ten steps.

$$\text{dev} = \sum_{h=2}^{\text{H-1}} \left| a_h - \frac{a_{h-1} + a_h + a_{h+1}}{3} \right| \tag{3}$$

3.5. Spectral Flatness

To evaluate the accuracy of the estimations of spectral flatness, we applied a Gaussian spectral window centered at the middle harmonic to a flat spectrum that contained both odd and even harmonics, and progressively altered its standard deviation in ten steps so that the last window resulted in an extremely peaky spectrum. For altering the width of the window we used the following coefficients, which are proportional to the reciprocal of the standard deviation: 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7 and 8. This process was done for the whole range of fundamentals.

3.6. Inharmonicity

This sound set is similar to 3.2, but here we used inharmonic spectra. The inharmonic components were kept fixed in the whole sound set, and were spaced according to an inharmonicity coefficient that controlled the amount of deviation from each harmonic, which varied linearly from 0 to 0.5 with respect to the harmonic number. Inharmonicity was increased by gradually increasing the amplitude of the inharmonic components instead of increasing their deviation from the harmonics. The inharmonic components were initially attenuated with a $1/n^2$ envelope to reach a flat spectrum in ten steps μ by gradually increasing linearly their energy distribution.

4. RESULTS

We evaluate toolbox performance by analyzing the sound sets with each toolbox and calculating the normalized root mean squared (RMS) error between their output and the theoretical values. The theoretical values were calculated using either the power or magnitude scale depending on the input representation being tested. MIRtoolbox's default input representation using 'mirspectrum' is based on a STFT with a Hamming window and a half overlapping frame length of 50 ms, which is similar to the 'STFTmag' representation used in the Timbre Toolbox. For the Timbre Toolbox, we tested all the available input representations because there is no default option. For analyzing the sounds, we used the default settings of each toolbox, and the summary statistics from the frame-by-frame analysis were derived using the median values.

4.1. Temporal Energy Descriptors

MIRtoolbox uses two estimation methods for calculating the attack and decrease slopes: 'Diff', which computes the slope as a ratio between the magnitude difference at the beginning and end of the attack period and the corresponding time difference; and 'Gauss', which is similar to Peeters' method [1]. Table 1 shows the results of the error analysis. The observed general trend for both toolboxes was that short attack times (about less than 40 ms) were significantly overestimated, whereas longer attack times were mainly underestimated. The Timbre Toolbox also systematically estimated the exponential attacks as being longer than the logarithmic attacks.

Descriptors	**Timbre Toolbox**	**MIRtoolbox** (Diff / Gauss)
Attack Time	24.40	21.57
Attack Slope	36.85	36.15 / 36.82
Decrease Slope	37.31	37.53 / 37.36

Table 1. RMS error (%) of temporal energy descriptors.

4.2. Spectral and Harmonic Descriptors

Although we tested the accuracy of extracted descriptors on all sound sets, due to space limitations, the evaluation results presented in Table 2 are based only on the designated sets for each descriptor, which were presented in the previous section. Also, we only report the most accurate results (i.e., the minimum RMS error) among the Timbre Toolbox's different input representations. In the following, we present a qualitative inspection of the errors with respect to the sound sets.

Centroid: MIRtoolbox always overestimates slightly the centroids, whereas the Timbre Toolbox returns accurate results for fundamentals of 65.4 Hz and above.

Higher-order moments of the spectrum and roll-off: the MIRtoolbox was numerically unstable returning 'Not a Number' (NaN) in the estimation of spectral centroid for the sets with fundamentals of 34.65 Hz and 73.42 Hz. Table 2 summarizes the results after removing the sounds for which MIRtoolbox returned NaNs. Timbre Toolbox's STFTpow representation provides overall the most accurate estimations even when all sounds were included in the analysis, in which case it produced a 1.37% RMS error for spectral roll-off.

Spectral Flatness: MIRtoolbox again returned Not a Number for some of the sounds with fundamentals of 34.65 Hz and 73.42 Hz, and although this sound set was not designed to test the estimation of spectral irregularity, MIRtoolbox did not provide any results for the estimation of this descriptor and exited with an error message. We also noted that in both toolboxes, as the fundamental frequency increases and spectral spread decreases, the estimations of spectral spread become more erroneous, although limited within a small margin. Although the spectral centroid and higher moments were estimated quite accurately in both toolboxes, the estimation of spectral flatness was inaccurate.

Spectral irregularity (or *deviation*): Harmonic spectral deviation is only available in the harmonic representation of the Timbre Toolbox. However, we were not able to run the analysis on the whole sound set using the default amplitude threshold setting for harmonic detection: as the fundamental frequency increases, the settings should be lowered, otherwise the sound will not be further analyzed (in the beta version tested here, the user gets warned whenever this situation occurs). The MIRtoolbox also proved to be erroneous for the estimation of this descriptor. However, both toolboxes returned quite accurate results for the spectral centroid and higher-order moments for this sound set.

Inharmonicity: The estimations of inharmonicity could not be quantitatively evaluated due to the current behavior of the Timbre Toolbox, as mentioned previously, and the unavailability of the precise equation used by MIRtoolbox. Qualitatively, and given the way this sound set was constructed (section 3.6), we expect the estimation of inharmonicity to increase for the subsets of each fundamental. Fig. 1 shows the estimations of the MIRtoolbox, which seem to be more plausible after the fifth set of fundamentals (i.e., from F5 up to B7, section 3).

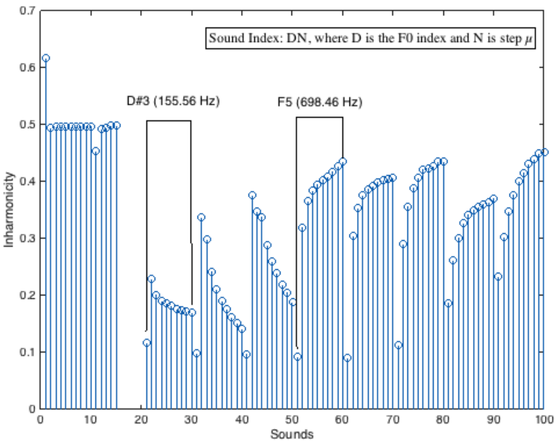

Figure 1. Inharmonicity estimation in the MIRtoolbox. The horizontal axis indicates the sound index DN, and the vertical axis the relative deviation of the partials from purely harmonic frequencies. The missing values correspond to NaN.

5. CONCLUSIONS

Before evaluating the accuracy of the toolboxes, we reported and fixed in a beta version the major bugs, configuration, and presentation issues that were encountered in the current version of the Timbre Toolbox (v. 1.2). Our evaluation on synthetic test sounds shows that for spectral descriptors, the Timbre Toolbox performs more accurately and on some sound sets outperforms the MIRtoolbox, with the short-term Fourier transform power representation (STFTpow) being overall the most robust. The estimations of spectral centroid and higher order moments of the spectrum were quite accurate with small errors except the estimation of spectral flatness, which both toolboxes estimated erroneously. The Timbre Toolbox failed to analyze some sounds using the harmonic representation with the default settings even though all sounds were strictly harmonic. In the beta version tested here, if this situation occurs, the estimation of fundamental frequency is automatically set to zero, which affects the calculation of all descriptors related to this representation. However, the user receives a warning message in order to alter the default settings appropriately. The MIRtoolbox's estimations of spectral centroid on some sounds, and spectral irregularity and inharmonicity on a specific sound set proved to be numerically unstable returning NaN, or exiting with error messages without providing any results. For descriptors that are based on the estimation of the temporal energy envelope, both toolboxes perform almost equally but poorly. We noticed that in this case the errors depend both on the attack or decay times and on the shape of the slopes. The test sound sets are available at:https://www.mcgill.ca/mpcl/resources-0/supplementary-materials

Descriptors	Timbre Toolbox	MIRtoolbox
Centroid	01.21 (STFTpow)	03.56
Spread	00.00 (STFTpow)	02.95
Skewness	02.06 (STFTmag)	03.82
Kurtosis	04.31 (STFTmag)	06.87
Roll-off	00.00 (STFTpow)	01.57
Flatness	34.87 (ERBgam)	51.82
Irregularity	N/A	31.36

Table 2. RMS error (%) of spectral energy descriptors. In the Timbre Toolbox, spectral irregularity could not be evaluated after the fifth set of fundamentals (section 3).

REFERENCES

[1] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals," *J. Acoust. Soc. Am.,* 130, pp. 2902-2916, 2011.

[2] O. Lartillot, *MIRtoolbox 1.6.1 User's Manual,* Technical report, Aalborg University, Denmark. December 2014.

[3] http://www.mathworks.com

[4] D. Moffat, D. Ronan, and J. D. Reiss, "An Evaluation of Audio Feature Extraction Toolboxes," in *Proc. DAFx,* Trondheim, Norway, 2015.

[5] A. Caclin, S. McAdams, B.K. Smith, and S. Winsberg, "Acoustic Correlates of Timbre Space Dimensions: A Confirmatory Study Using Synthetic Tones," *J. Acoust. Soc. Am.,* 118 (1), pp. 471-482, 2005.

[6] C. Douglas, *Perceived Affect of Musical Instrument Sounds,* Master's thesis, McGill University, Montreal, Canada. June 2015.

[7] K. Jensen, *Timbre Models of Musical Sound: From the model of one sound to the model of one instrument*, PhD thesis, University of Copenhagen. 1999.

[8] J. Krimphoff, S. McAdams, and S. Winsberg, "Caractérisation du timbre des sons complexes. II : Analyses acoustiques et quantification psychophysique," *Journal de Physique,* 4(C5), 625-628, 1994.