Timbre Manipulation from Audio Features Based on Fractal Additive Synthesis

Thiago Rossi Roque
Institute of Arts
University of Campinas
thiago.roque07@gmail.com

Rafael Santos Mendes
School of Electrical and Computer Engineering
University of Campinas
rafael@dca.fee.unicamp.br

ABSTRACT

The search for new sound synthesis techniques has gained considerable impulse from the advances in Music Information Retrieval (MIR). From the concept of audio features, introduced by MIR, a new idea of sound synthesis has emerged based on the manipulation of high levels parameters, more directly involved with perceptual aspects of sound. In this article, we present the initial version of a software for feature modulation based on the Fractal Additive Synthesis technique and the results on using this software as a tool to support the teaching of audio features in a music technology class for musicians at undergraduate level at the University of Campinas. Four audio features were chosen for this research: spectral centroid, even to odd harmonic energy ratio, mean harmonic band Hurst exponent and the harmonic band correlation coefficient.

1. INTRODUCTION

The act of composing music, as a form of expression through sounds, has an intrinsic relation to the understanding of the sound phenomenon. The greater the artist's mastery over his tool, the greater his creative possibilities. The German composer Helmut Lachenmann presents the composing as a result of the reflection on the means of the very act of composing [1]. According to this premise, the understanding of sound becomes something essential to the sound artist, as greater knowledge of the means allows deeper reflections and, consequently, more interesting musical compositions.

By the late 20^{th} century, the advances in electronics, computer science and signal processing made possible the development of new techniques for analysis and manipulation of sound signals. Information technology has brought new paradigms to the theory of signal processing with the emergence of machine-listening. In this same context, the multidisciplinary science of Music Information Retrieval (MIR) has been consolidated by joining several research areas focused on retrieving information from digital sounds. [2]. One of the main concepts developed by MIR is the extraction of sound features (descriptors) which aims to represent sound objects in sets of quantitative data. Through the use of these tools, it is possible to objectively describe many qualitative and perceptual aspects of sound signals.

Through the study of sound features, it is possible to analyze details of sound objects beyond traditional techniques, allowing researchers and artists to get a deeper level of comprehension from the sound phenomenon. By making possible an objective measurement of qualitative aspects of sound, the sound descriptors have become an important tool for contemporary composers.

Over the years, many different techniques for audio feature extraction have been proposed [3]. Mostly, these techniques have been used solely for content analysis and classification but, over the past decades, a few researchers begin to glimpse the use of audio features not only for the extraction of information but also for the manipulation of these contents [4, 5, 6, 7]. All of these techniques are based on the Fourier transform for spectral analysis and feature extraction. Only recently new approaches have been presented making use of *autoencoders* and machine learning techniques [8], allowing the control of timbral aspects of the sound without direct manipulation of audio features.

Unlike these aforementioned techniques, the system herein presented make use of the harmonic band wavelet transform and the fractal additive synthesis as a decomposition tool. As will be detailed in section 2, fractal additive synthesis has the capability to make much deeper analysis on the harmonic profile than more traditional techniques and provide insights over the pseudo-periodicity of the analyzed sound. Traditional signal processing techniques were preferred over machine learning approaches due to permit a better understanding of how each feature correlates to perceptual aspects of the sound.

By gaining control over each sound feature and its perceptual aspects, a composer would be able to objectively modify the timbre of his sounds. If a sufficient set of manipulable, orthogonal and unidimensional sound features is defined, as shown by many researches on musical timbre semantics [9], this methodology could possibly overcome the mapping problem between control parameters and the resulting outcome from traditional sound synthesis techniques [10].

This paper presents the development of an initial tool for feature extraction and manipulation based on the fractal additive synthesis (FAS) and the methodology used for the feature modulation. As result, a MATLAB app was created and it was used, on an experimental basis, on a music technology undergraduate class to aid the teaching of sound features, its capabilities and its relation to perceptual aspects of the sound.

1.1 Proposed Method

The technique herein proposed has an analysis/manipulation/resynthesis structure as presented in Figure 1, where a pitched sound (mainly from a musical instrument) is used for timbral manipulation.

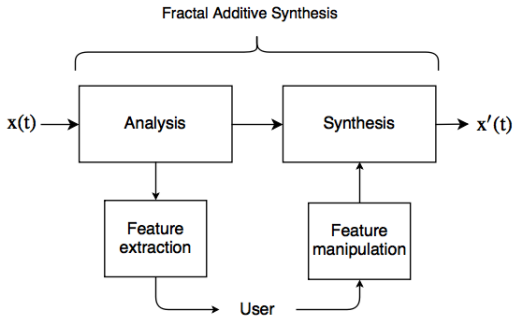

Figure 1. *System structure*

Sound features shall be extracted directly from FAS parameters during the analysis stage. Through some control parameter, the user will be able to alter these features values before the resynthesis in order to change the timbre and perceptual aspects of the resulting sound. This algorithm has been implemented as a MATLAB App with a graphical user interface to facilitate the interaction between the user and the algorithm.

For this research, the following features were selected: Spectral Centroid (SC), Spectral Spread (SS), Even to Odd Harmonic Ratio (EOR), Tristimulus (TR), Mean Harmonic Band Hurst Exponent ($\bar{H}$) and Harmonic Band Correlation Coefficient (HBCC) [1]. These descriptors were chosen based on their well-known importance on the perception of timbre [11]. Temporal features will be left for further versions.

As will be further detailed in Section 2, FAS has the capability to encode any pitched sound into a small set of parameters easily manipulable, in both spectral and temporal domains, by the separation and codification of deterministic and stochastic contents. Section 3 will cover the extraction and manipulation of the audio features in the context of FAS and section 4 presents the software designed to evaluate this technique and the results obtained.

2. FRACTAL ADDITIVE SYNTHESIS

Fractal additive synthesis (FAS) is an efficient coding technique for pitched sounds (sounds with a detectable pitch) develop by Pietro Polotti and Gianpaolo Evangelista [12] based on the harmonic band wavelet transform (HBWT) [13].

The main concept of FAS and HBWT is to encode sounds from its deterministic and stochastic contents. Concerning similar techniques, like Spectral Modelling Synthesis (SMS) [14], FAS takes a deeper analysis of each harmonic content by modelling its side-bands as a $1/f$ stochastic process.

[1] SS and TR have been used only for analysis and comparison, not yet for modulation

From Polotti's point of view, every natural sound with detectable pitch has some chaotic, but correlated, micro-fluctuation on its periodicity, making it pseudo-periodic. As a result of this pseudo-periodicity, it is possible to observe on each harmonic a $1/f$ profile along its side-bands, that is, a $1/|f - f_n|$ power spectrum centred on the n^{th} harmonic partial. Any audio coding technique that doesn't take into account this non-ideal harmonic side-band and, consequently, the pseudo-periodicity of pitched sounds, cannot be able to emulate its naturality and dynamics.

A mathematical description of FAS and HBWT can be found in the literature [15, 16] and is out of the scope of this article.

To exemplify how the separation between deterministic and stochastic contents works in FAS, two signals were analyzed and resynthesized: a first one, artificially generated, composed by five pure harmonics plus a white noise; and a second one composed by the sustain section of a cello sample (note A2, *mezzoforte*). The cello sample was obtained from The University of Iowa Electronic Music Studios [17]. Figures 2 and 3 shows the spectrum of this two signals.

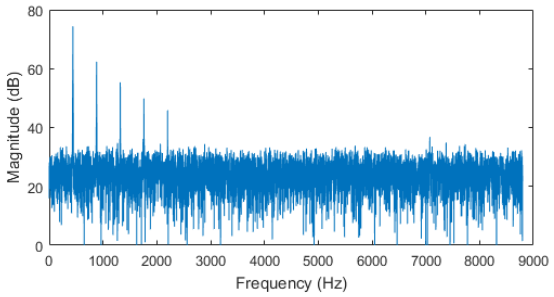

Figure 2. *Spectrum of the artificial signal composed by five pure harmonics plus white noise*

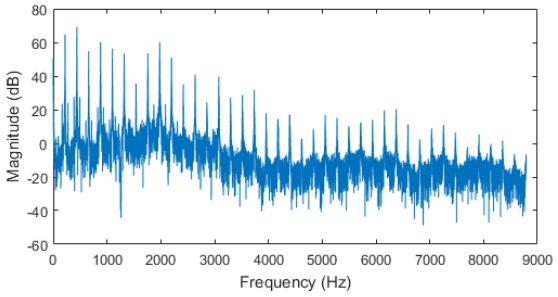

Figure 3. *Spectrum of the cello sample*

After the encoding of these signals by FAS, each deterministic and stochastic contents were synthesized separately. Figures 4 and 5 shows the spectrum solely for the stochastic content for both signals. The deterministic contents can be seen in Figures 6 and 7.

From Figures 2, 4 and 6, it can be clearly seen how this technique can separate the stochastic content from any perfect harmonic content. One thing worth noting is the apparent emergence of harmonic content beyond the original ones in Figure 6, but a closer look shows that these are part of stochastic content as its energy is as low as the noise floor presented in Figure 2.

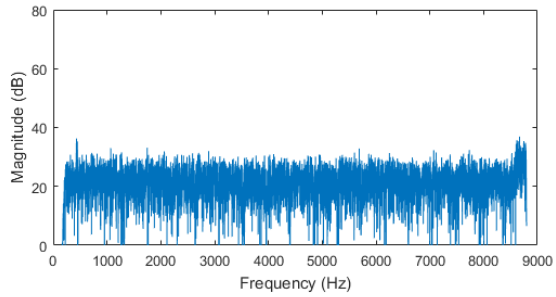

Figure 4. *Stochastic content of artificial signal*

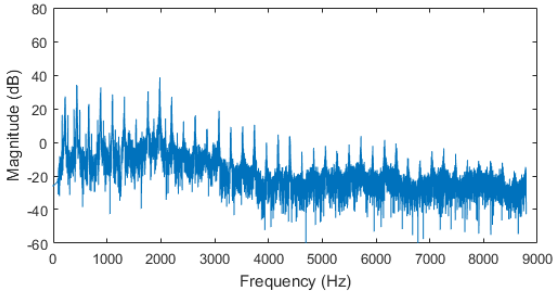

Figure 5. *Stochastic content of cello sample*

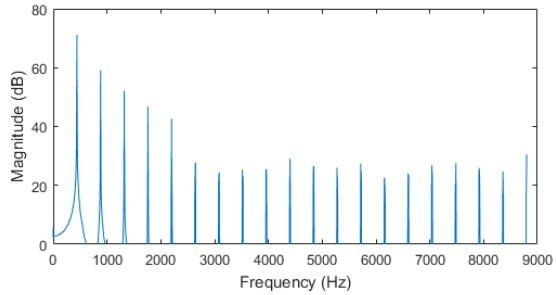

Figure 6. *Deterministic content of artificial signal*

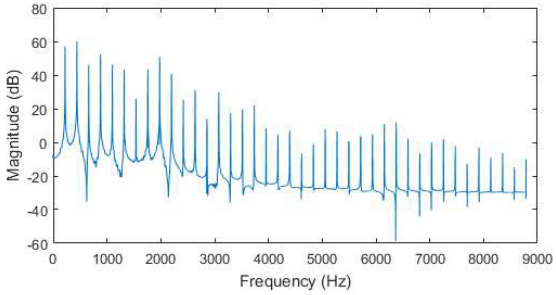

Figure 7. *Deterministic content of cello sample*

A comparison between Figures 4 and 5 shows the difference between natural and artificial, perfect, harmonics. Some amount of the harmonic profile can be seen in Figure 5 because of the non-ideal side-band that was coded as stochastic content, differently from Figure 4, where virtually no trace of the harmonic content can be seen.

These results attest to the original idea of FAS, that natural harmonics have a not negligible side-band that should be taken into account for a better codification of pitched signals.

3. FEATURE EXTRACTION AND MODULATION

As already mentioned, the chosen features for this initial analysis were: spectral centroid, even to odd harmonic energy ratio, average harmonic band Hurst exponent and harmonic band correlation coefficient. In this section, each one of these features shall be presented along with its methodology for extraction and manipulation using the FAS framework.

3.1 Spectral Centroid

Spectral centroid (SC) is one of the main audio features to characterize spectrum profile by measuring its barycentre [3]. It has a well-known correlation with the perception of brightness on sounds [18]. The extraction of the SC is done in a similar way to its original version, based on the Fourier transform. However, as we have the full spectrum split into stochastic and deterministic content, they must be merged into a single power estimation parameter $B(p)$, calculated from the variance of the stochastic coefficients $b_{n,p}$ and the deterministic coefficients $a_{N,p}$, as shown in equation (1).

$$B(p) = \sum_{n=1}^{N} Var(b_{n,p}) + Var(a_{N,p}) \tag{1}$$

From $B(p)$, the SC can be measured from equation (2), where $f(p)$ is the frequency of the harmonic partial associated with the p^{th} channel of HBWT.

$$SC_{FAS} = \frac{\sum_{p=0}^{P-1} B(p)f(p)}{\sum_{p=0}^{P-1} B(p)} \tag{2}$$

From the separation between the deterministic and stochastic content, two direct variants from SC can be easily obtained, the harmonic spectral centroid (HSC) and the stochastic spectral centroid (SSC).

The HSC is widely known in the literature and follows the same extraction procedure of the SC, although relies only on the harmonic content of the signal. From FAS, it can be measured from the equation (1) by using only $a_{N,p}$ coefficients, as shown in equation (3).

$$B(p) = Var(a_{N,p}) \tag{3}$$

The SSC is proposed here as a new possibility of sound measure and stochastic description. Together with the average harmonic band Hurst coefficient (see section 3.3), these two features may give information about the profile of the stochastic content of sounds. The SSC extraction follows the same procedure presented for the HSC, although relying on $b_{n,p}$.

Both SC, HSC and SSC modulation algorithms are based on Tae Hong Park's work [4], where a modulation curve $V(p)$, with variable angular coefficient, can change the energy ratio between each $B(p)$, changing the spectral profile, as illustrated in Figure 8.

For a greater action of $V(p)$ over the SC modulation process, instead of a straight line, $V(p)$ is made with a slight

curvature, from the polynomial shown in 5, obtained empirically. c is the angular coefficient and controls the modulation amount. A positive c will give more weight to higher frequencies, raising the SC value, negative c will lower high-frequency energy and raise low-frequency energy, lowering the SC value.

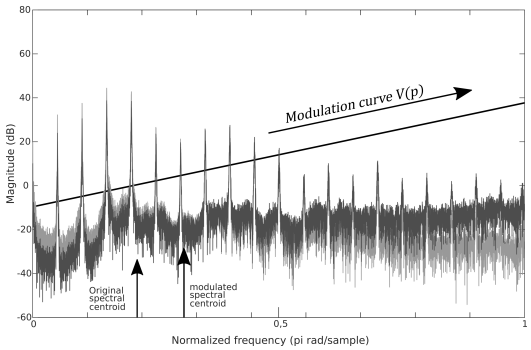

Figure 8. *Illustration of the spectral centroid modulation process*

$$B'(p) = B(p)V(p), \quad p = 1 : P \tag{4}$$

$$V(p) = (0,009|c|) * p^2 + c * p \tag{5}$$

3.2 Even to Odd Harmonic Energy Ratio

The even to odd harmonic energy ratio descriptor (EOR) measures the energy proportion between the even and the odd harmonic partials. Recent studies have shown that, after attack time and spectral centroid, EOR is the most salient timbral feature for timbre perception [19].

As a harmonic content measure, the extraction of EOR relies only on the deterministic content. From FAS, it is measured from the energy estimation of the deterministic coefficients $A_{c(k)}$.

$$EOR_{FAS} = \frac{\sum_{k=2:2:K} \log A_{c(k)}}{\sum_{k=1:2:K} \log A_{c(k)}} \tag{6}$$

EOR modulation is done from a multiplier parameter α. If $\alpha > 1$, there is an increase in the energy of odd harmonics and, in the same proportion, a decrease in the energy of even harmonics. If $0 < \alpha < 1$ the opposite happens. By raising and lowering the energy of the harmonics in the same proportion, the total energy of the signal is maintained.

$$A'_{c(k,m)} = \begin{cases} \alpha A_{c(k)} & \text{if} \quad k = 1 : 2 : K \\ \frac{1}{\alpha} A_{c(k)} & \text{if} \quad k = 2 : 2 : K \end{cases} \tag{7}$$

For better control over the modulation process, a λ parameter, related to α by equation (8) is used as a modulation index.

$$\alpha = \sqrt{\lambda} \tag{8}$$

3.3 Mean Harmonic Band Hurst Exponent

The mean harmonic band Hurst exponent $(\bar{H})$ is based on an audio feature called harmonic band Hurst exponent (H_p), developed by Aldo Díaz [20]. H_p is a measure of the harmonic side-band self-similarity for each channel p from HBWT. H_p can be extracted by the FAS parameter γ_p in the same way as the original Hurst exponent [21], as shown in equation (9).

$$\gamma_p = 2H_p + 1 \tag{9}$$

γ_p is a measure of the decay rate of the $1/f$ harmonic side-band profile and is a good measure of its stochasticity. Thus, H_p can be seen as a measure of the pseudo-periodicity of the signal and should, ideally, remain within the range $0 < H_p < 1$, corresponding to $1 < \gamma < 3$. Instead of keeping every H_p for each p side-band, it is proposed here the average harmonic band Hurst exponent, which represents the average pseudo-periodicity of the signal under analysis. $\bar{H}$ is measured from the arithmetic average of γ_p, weighted by factor 1.2, empirically defined for better value resolution, as defined by equations (10) and (11).

$$\Gamma = \frac{1}{P} \sum_{p=0}^{P} \gamma_p^{1,2}, \tag{10}$$

$$\bar{H_p} = \frac{\Gamma - 1}{2} \tag{11}$$

The modulation process is done by defining a target value $\bar{H_s}$ for $\bar{H}$ and finding the associated Γ_s from equation (12).

$$\Gamma_s = 2H_s + 1 \tag{12}$$

From the ratio between the target Γ_s and the measured Γ (Γ_a), is defined the parameter η, from which a new γ_p is calculated, as described by equation (14).

$$\eta = \frac{\Gamma_s}{\Gamma_a} \tag{13}$$

$$\gamma'_p = \eta\gamma_p, \;\; p = 1, 2, ...P \tag{14}$$

3.4 Harmonic Band Correlation Coefficient

Aiming to extract more information from the stochastic contents of FAS, the harmonic band correlation coefficient (HBCC) have been proposed based on the Pearson correlation coefficient measure of the stochastic coefficients $b_{p,n}$.

Considering $b_{p,n}$ as a time series X_n weakly stationary, the correlation coefficient of this series, ρ_i, can be obtained by the co-variance of these elements with themselves, lagged by a factor i and normalized by their variance, as presented in equation (15).

$$\rho_i = \frac{Cov(X_n, X_{n+i})}{\sigma^2} \tag{15}$$

For the HBCC, the mean of the quadratic value of ρ_i, for $1 < i < 12$ is taken from each sideband of each harmonic, as presented in equation (16), where p is the sideband index and ρ_{FAS} is the mean of the quadratic value of ρ_i.

$$HBCC = \frac{\sum_{p=0}^{P} \sqrt{\rho_{FAS}(p)^2}}{P} \tag{16}$$

The modulation of HBCC is done by applying a de-emphasis to the auto-regressive (AR) filter coefficients during the synthesis process of FAS. By reducing the effect of the AR filters during the synthesis of the stochastic coefficients $b_{p,n}$, these coefficients loses its self-similarity and gets closer to a pure uncorrelated signal. This procedure is accomplished by equation (17), where n is the AR filter coefficient index, and α is the modulation index.

$$h(n)'_{AR} = h(n)_{AR}\, e^{-\alpha n} \tag{17}$$

4. GUI FOR THE FAS FEATURE MODULATION

To facilitate the usability of the proposed algorithm for feature modulation, a MATLAB App with a graphical user interface have been developed.

This App allows the user to load any pitched sound (in .wav and .aiff formats), select which features he wants to modulate (one at a time or several at the same time), the amount of modulation for each feature and resynthesize the sound with the requested modifications. By loading the sound, the interface plots both spectral and temporal representation of the original file and presents the extracted feature values. The same representations are presented for the resynthesized sound after processing, for easy comparison. The "Partials" number controls the number of harmonic partials that will be encoded by FAS.

Figures 9 and 10 exemplifies the use of the App for the modulation of SC and $\bar{H}$, respectively.

Gradual increments on the modulation index resulted in gradual modification on both the extracted feature value and the perceptual aspects of the sound. This was a desired feature for the App due to the intention to use it as a teaching tool, allowing the user to easily understand the behavior of the modulation process from the interface parameters. Higher modulation index results in more drastic changes in the sound.

5. RESULTS

In addition to evaluating this feature modulation technique, this App has proven a good method for the analysis of pitched sounds, as it gives interesting and explicit information correlated to the timbre of the sound loaded.

Perceptual tests were performed with the team involved on this development and, although not scientifically controlled, a clear correlation between changes in the modulation index and perceptual features were noted, mainly related to the SC and the perceptiveness of brightness, and the $\bar{H}$ with the perception of noisiness.

About HBCC, subtle differences were already expected for the modulation process due to the nature of this feature,

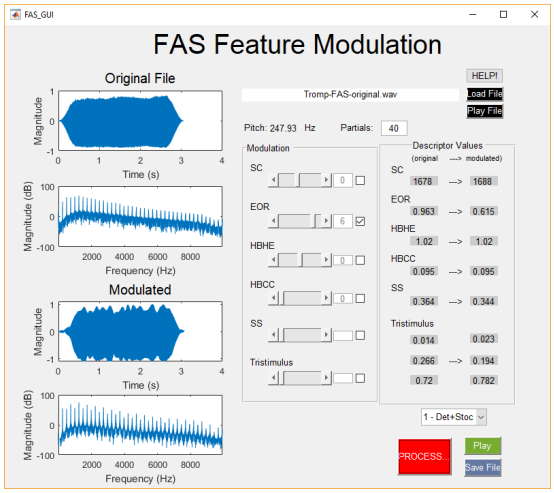

Figure 9. *Example of the App for the modulation of EOR*

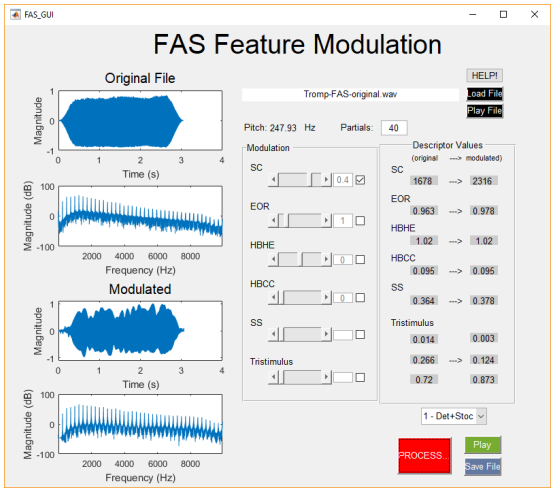

Figure 10. *Example of the App for the modulation of SC*

but for some instruments, some non-negligible differences could be noted.

Interesting results were taken from the analysis of the interaction between the modulation of a feature to the extracted values of others. Some have shown virtually no interference while others exhibit direct interaction, like EOR with TR and SC with SS.

5.1 Feature Modulation as teaching support

Recently, this application was used, on an experimental basis, to aid a lecture on audio features for music composition students at UNICAMP. Through the App, students could explore practical use of audio features and link how the values obtained from feature extraction can correlate to perceptual aspects of the sound. Although just a few audio features are present on the App, it has proven an interesting tool for teaching purposes.

The App developed, with all the source code and some audio samples, can be found under the author's github page: `https://github.com/thiago-roque07/FAS_feature_manipulation`

6. CONCLUSIONS

FAS proved to be an interesting technique for audio codification and deserve more exploration, as it presents some capabilities not found in similar techniques. The ease of working with its coefficients makes it an interesting base for many kinds of processing techniques.

The concept of feature modulation has shown a powerful tool for controlling perceptual aspects of sounds and deserves more attention from scholars and researchers. It is quite possible that in a close future many computer-aided music tools incorporate this concept at a professional level.

The Matlab App has shown itself as an interesting tool for educational, as it allows the user to experiment with different sounds and understand how some perceptual aspects of the sound correlates with audio feature values. After using the App on a lecture on audio features, all 12 students in the class mentioned a greater interest in the concept of audio features and its capabilities, showing the potential of this kind of approach to the teaching of this subject.

For future releases of the App, it is desirable to expand the number of audio features, mainly including temporal ones. The pitch detection still needs some improvements, as FAS needs a high level of precision for the pitch of the analyzed audio for the correct function of its algorithm.

7. REFERENCES

[1] H. Lachenmann, *Écrits et entretiens*. Éditions Contrechamps, 2009, ch. De la composition (1986), p. p. 129–141.

[2] M. Fingerhut, "Music Information Retrieval, or how to search for (and maybe find) music and do away with incipits," *IAML-IASA Congress, Oslo*, 2004.

[3] G. Peeters, "A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project," IRCAM, Tech. Rep., 2004.

[4] T. H. Park, J. Biguenet, Z. Li, C. Richardson, and T. Scharr, "Feature modulation synthesis (FMS)," in *Proc. ICMC*, vol. 2007, 2007.

[5] M. Hoffman and P. R. Cook, "Feature-Based Synthesis: Mapping Acoustic and Perceptual Features onto Synthesis Parameters," in *in Proceedings of the International Computer Music Conference, New Orleans*, 2006.

[6] D. Mintz, "Toward Timbral Synthesis: a New Method for Synthesizing Sound Based on Timbre Description Schemes," Master's thesis, University of California, Santa Barbara, June 2007.

[7] M. Freitas Caetano and X. Rodet, "Sound Morphing by Feature Interpolation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Czech Republic, May 2011, pp. 11–231. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00604386

[8] S. Sinclair, "Sounderfeit: Cloning a Physical Model with Conditional Adversarial Autoencoders," *Proceedings of the 16th Brazilian Symposium on Computer Music*, September 2017.

[9] A. C. Disley, D. M. Howard, and A. D. Hunt, "Timbral description of musical instruments," in *International conference on music perception and cognition*, 2006, pp. 61–68.

[10] A. Hunt and R. Kirk, "Mapping strategies for musical performance," *Trends in Gestural Control of Music*, vol. 21, pp. 231–258, 2000.

[11] M. Ilmoniemi, V. Valimaki, and M. Huotilainen, "Subjective evaluation of musical instrument timbre modifications," in *Joint Baltic-Nordic Acoustic Meeting*, 2004, pp. 8–10.

[12] P. Polotti, "Fractal Additive Synthesis: Spectral Modeling of Sound for Low Rate Coding of Quality Audio," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2003.

[13] P. Polotti and G. Evangelista, "Analysis and Synthesis of Pseudo-Periodic 1/f-Like Noise by Means of Wavelets with Applications to Digital Audio," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 1–14, 2001.

[14] X. Serra, "A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition," Ph.D. dissertation, Stanford University, 1989. [Online]. Available: files/publications/PhD-Thesis-1989-xserra.pdf

[15] P. Polotti and G. Evangelista, "Fractal additive synthesis via harmonic-band wavelets," *Computer Music Journal*, vol. 25, no. 3, pp. 22–37, 2001.

[16] T. R. Roque, "Extração e Modulação de Discritores Sonoros a Partir da Síntese Aditiva Fractal," Master's thesis, Faculdade de Engenharia Elétrica e Computação, UNICAMP, 2017.

[17] L. Fritts, "Electronic Music Studios, University of Iowa," 1997. [Online]. Available: http://theremin.music.uiowa.edu

[18] J. M. Grey and J. W. Gordon, "Perceptual Effects of Spectral Modifications on Musical Timbres," *The Journal of the Acoustical Society of America, Volume 63, Issue 5*, 05/1978.

[19] B. Wu, A. Horner, and C. Lee, "Musical timbre and emotion: The identification of salient timbral features in sustained musical instrument tones equalized in attack time and spectral centroid," in *ICMC*, 2014.

[20] A. Díaz, "Análise de Instrumentos Musicais através do Expoente Hurst de Banda Harmônica - Estudo Comparativo da Quena e de outros Instrumentos de Sopro," Master's thesis, Faculdade de Engenharia Elétrica, UNICAMP, Campinas, Brasil, 2015.

[21] B. B. Mandelbrot and J. R. Wallis, "Noah, Joseph, and Operational Hydrology," *Water Resources Research*, vol. 4, no. 5, pp. 909–918, 1968.