Audio Engineering Society

Conference Paper 10

Presented at the 6th International Conference on Audio for Games
2024 April 27–29, Tokyo, Japan



# Discerning real from synthetic: analysis and perceptual evaluation of sound effects.

Nelly Garcia[1,2], Yisu Zong[1,2], and Joshua Reiss[1,2]

[1]*Queen Mary University of London*
[2]*Centre for Digital Music*

Correspondence should be addressed to Nelly Garcia (n.v.a.garcia-sihuay@qmul.ac.uk)

**ABSTRACT**

In audio post-production, the adoption of sound synthesis offers a viable alternative for searching and recording samples in creating soundscapes. However, a central concern arises regarding the ability of synthetic sounds to match the perceived authenticity of library samples. This paper introduces an analytical approach, examining authentic and synthetic samples in five categories(burning embers, pouring water, explosions, popping bubbles and church bells) by delving into audio descriptors that distinguish both types. We focus in the utilization of machine learning classification models and a perceptual evaluation experiment. The perceptual evaluation was between five distinct synthesis techniques – granular, additive, subtractive, physically informed, and modal synthesis – revealed that subtractive synthesis is perceived as more realistic in explosion sounds, while additive synthesis works better with pouring water sounds. This study provides valuable insights into the audio descriptors that may require modification in specific synthetic models, paving the way for a deeper understanding of sound synthesis methods and facilitating their integration into the sound design process.

## 1 Introduction

Sound design significantly influences the perception of realism in the entertainment industry. Crafting immersive soundscapes is based on the use of foley.This task can be time-consuming due to the repetition of sounds and the challenge of finding suitable alternatives. In recent years, the practice of creating sounds from scratch has gained prominence across various sectors, including TV, video games, cinema, and podcasts.

Sound synthesis involves generating audio through computational processes. We utilized Nemisindo [1], a browser-based sound effect synthesis framework, as the primary tool to create our synthetic dataset. While [2] acknowledged numerous attempts to evaluate synthetic models and define sound creation methods, there is currently no established standard for problem definition, dataset, or evaluation in Foley or sound synthesis, as highlighted in [3].

In Section 2, we detail the synthesis methods used to create and compare sounds across five categories: burning embers, explosions, pouring hot water, popping

bubbles, and church bells. Section 3 explores the importance of audio descriptors in comparing Nemisindo synthetic models to real sounds. Section 4 presents our comparative analysis of various synthetic models (subtractive, additive, modal, sinusoidal, physically inspired, and granular) using a methodology similar to the Multiple Stimuli with Hidden Reference and Anchor Test ([4]). Finally, in Section 5, we provide a comprehensive discussion of our results and outline conclusions for this research.

## 2 Methods

### 2.1 Synthesis Methods

The synthesis methods in this study were derived from a code analysis of the Nemisindo models for each sound shown in Table 1, with model definitions outlined by [5].

**Additive and modal synthesis**, create intricate waveforms through the addition of sine waves with independent amplitudes, frequencies, and phases[6]. **Modulation**, involves the variation of a dynamic signal influenced by another signal, also known as the carrier and modulator signals. **Subtractive synthesis**, begins with a waveform rich in overtones, employing filters to produce sonically rich, harmonically complex outputs. The **physically informed synthesis**, has inspiration from the physical properties of sound, using known physics to guide signal-based models with enhanced accuracy. Additionally, our exploration extends to **granular synthesis**, inspired by the works of [7, 8]. Granular synthesis involves segmenting audio samples into tiny "grains" (1 to 1000 milliseconds), providing a unique manipulation to create new sounds, [7].

### 2.2 Data

We gathered sound samples from three places: BBC Sound Effects Library [1], Pro Sound Hybrid Library [2], and the dataset by Piczak [9]. At the same time, we used the Nemisindo's online procedural audio system [3] tool to make around 55 artificial sounds for each type. For fairness, we selected the sounds with the following characteristics: Explosions with just one "boom", church bells with a single hit, burning embers with noticeable crackling and hissing, popping bubbles with a sweeping sound and pouring hot water sounds with a noticeable pipe noise.

**Table 1:** Synthetic models used in Nemisindo.

| Sound Effects and Synthesis Methods | |
|---|---|
| **Sound** | **Method 1 (Nemisindo)** |
| **Explosions (1 Boom)** | Subtractive |
| **Church Bells (1 Hit)** | Additive and modal |
| **Popping Bubbles** | Additive and modal |
| **Burning Embers (Fire)** | Additive, physically informed and subtractive |
| **Pouring Hot Water** | Additive and modal |

Our dataset has 680 sounds in total. All sounds are 5 seconds long, 44.1Khz, mono-sourced and 16bps.

In section 4, we compare the synthetic samples and the real samples with granular, sinusoidal and additive synthesis from Turchet and Simoncelli's collection[4].

### 2.3 Audio Descriptors

The selection of audio descriptors in this study was influenced by [10], where a comprehensive comparison of objective evaluation metrics commonly employed in contemporary sound synthesis design was conducted. Additionally, sound descriptors from [11] helped shape our approach. These descriptions emphasized crucial audio features, including fundamental frequencies, the first four harmonics, spectral centroid, zero crossing rates, and visual aids such as spectrograms or Mel Frequency Cepstral Coefficients (MFCC) plots.

We meticulously chose a total of seventy-eight features, comprising both commonly used descriptors and global audio descriptors. Examples of these descriptors include attack time, zero crossing rates, temporal envelope, loudness, RMS, instant power, entropy, dynamic complexity, high frequency, spectral roll-off, intensity, temporal centroid to total length ratio (TCToTotal), spectral complexity, spread, skewness, kurtosis, etc.

[1] https://sound-effects.bbcrewind.co.uk/

[2] https://www.prosoundeffects.com/hybrid-library/

[3] https://nemisindo.com

[4] https://code.soundsoftware.ac.uk/projects/time-domain-probabilistic-concatenative-synthesis

For instantaneous descriptors, we computed the first 30 MFCC coefficients, spectral centroid, and Short Time Fourier Transform (STFT).

We extracted the hand-picked audio descriptors with the open-source library Essentia [12]. Following the categorization approach outlined in [13], the audio descriptors were systematically classified into temporal, spectral, and spectrotemporal domains.

## 3 Objective Evaluation

We employed a supervised approach with the sci-kit-learn library[5], utilizing both the Random Forest and XGBoost. The Random Forest aggregates the outputs of multiple decision trees[14] to generate a single result, while XGBoost[15] is an optimized distributed gradient boosting algorithm, offering parallel tree boosting. Our focus was on identifying descriptors that specifically differentiate real samples from synthetic ones.

### 3.0.1 SHAP Values and PCA's

We determined the top two audio descriptors among the initial seventy-eight using SHapley Additive exPlanations (SHAP) values, a method outlined in [16] that assigns importance values to features, revealing their impact on model predictions. A comprehensive analysis compared SHAP values from both XGBoost and Random Forest models. To interpret the results, we conducted Principal Component Analysis (PCA) specifically on the top two descriptors, a statistical technique reducing datasets and elucidating the relationship between sound components.

The description of each feature is listed below:

1. **Attack Time**: This temporal descriptor is defined as the time from the onset of a sound to its more stable phase [13]. In Essentia, this feature is computed using the LogAttackTime algorithm, which identifies the attack's onset, often estimated as the point where the signal envelope reaches 20 percent of its maximum value to account for potential noise presence.
2. **Effective Duration**: This temporal descriptor, this feature measures the time when the signal is perceptually meaningful and is computed from the signal envelope.
3. **Pitch Salience**: This spectral descriptor provides a rapid measure of the sensation of tone and is calculated as the ratio of the autocorrelation value of the spectrum to the non-shifted value.
4. **Dynamic Complexity**: Another temporal descriptor, this metric is associated with the dynamic range and the level of fluctuation within a recording. It is measured in decibels and is defined as the average absolute deviation from the global loudness level estimate.
5. **Spectral Flux**: A spectral descriptor defined by the difference between the L2-norm [17] for two consecutive frames of the magnitude spectrum.
6. **Kurtosis**: This statistical measure serves as an indicator of the shape of a distribution and quantifies how much data remains in the tail of a bell curve.
7. **High Frequency**: A spectral descriptor computed using the 'Masri' harmonic noise to ratio technique, represented in Equation 3. This descriptor provides insight into the high-frequency content of the audio signal.

$$HFC = |X(n)|^2 \cdot k \tag{1}$$

where |X(n)|, represents the magnitude of the signal $X$ at a specific point $n$ and $k$ is a constant.

Figure 1 illustrates the interaction of the top two features in the explosions, with the x-axis representing the first principal component and the y-axis denoting the second principal component.

Notably, the MaxToTotal audio descriptor, reflecting the temporal envelope, reveals the location of the maximum amplitude at the beginning, middle, or end. The lower the pitch at the begining of the temporal envelope amplitude the more realistic it's classified.

## 4 Quantitative subjective evaluation

The multistimulus test consisted of 11 audio questions. The questions included both open-ended and multiple-choice formats. The estimated test duration was 20-25 minutes. Feedback was gathered from 22 participants, all with experience in audio design, ranging in age from 21 to 63. Participants used a rating scale ranging from 0 to 100, evaluating sound samples based on realism

[5] https://scikit-learn.org/stable/

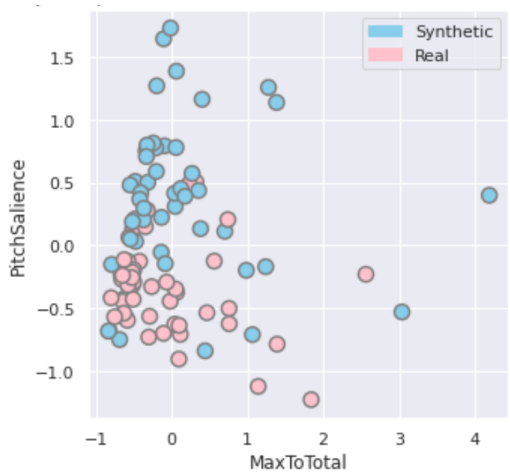


**Fig. 1:** *Top two PCA's for explosions category.*

criteria. The scoring system was defined as follows: 1-20 for completely unrealistic, 20-40 for very unrealistic, 40-60 for somewhat unrealistic, 60-80 for good, and 80-100 for realistic.

To ensure data consistency, five participants who did not rate library samples within the 80-100 range were excluded.Table 2 summarizes 69 test samples, indicating the quantity of questions and their corresponding sample count.

| Sound | Questions | Samples |
|---|---|---|
| Explosions | 3 | 7 |
| Pouring Water | 2 | 6 |
| Church Bells | 2 | 6 |
| Pouring Water | 2 | 6 |
| Popping Bubbles | 2 | 6 |

**Table 2:** Numbers of questions and samples per sound class.

For the explosion category we took a step forward comparing the standard procedural audio engine model to the optimization made by [18], where they attempted to make a more realistic sound by adding more parameters to the user controllable interface.

### 4.1 Results per sound class

The results for burning embers, pouring hot water, church bells, and bubbles after the listening test are displayed in box plots in figure 2, presenting perceptual results with a rating scale from 0 to 100 on the y-axis and synthesis methods from figure 2 on the x-axis. Median, confidence levels, outliers, and quartiles are marked. Notably, library samples didn't achieve a perfect 100 rating. Pouring hot water favors the additive method.

In bell sounds, granular scored lowest, while additive and additive+modal had similar ratings. Bubbles favored additive, contrasting with lower-rated additive+modal. Burning embers rated granular as most realistic, followed by additive, physically inspired, and subtractive as least realistic.

Figure 3 focuses on explosion; Setting 3 or the third optimization way is perceived as superior in a notched box plot adjusting the procedural audio engine model. Levene's test ($p=0.0198$) revealed unequal variances. Welch-ANOVA, due to non-normal distribution, found significant effects of synthesis methods on user perception ($p<0.00001$). Table 3 details pairwise comparisons of synthesis methods on perceptual realism ratings.

Upon analyzing the results presented in Table 3, it is evident that there are no significant differences in the ratings for Bubbles sounds across any of the methods employed. Additionally, the subtractive optimization with setting 3 exhibits no statistical difference in the perception of explosion sounds, unlike the other methods.

## 5 Qualitative Subjective Evaluation

Additionally, we surveyed participants in each category, asking the question: "Which attributes do you think can be enhanced to achieve a more authentic sound?". This inquiry sought to gather valuable insights into specific features and potential improvements that could make synthetic samples more authentic. Figure 4 illustrates the options perceived by the participants, with the most frequently mentioned enhancements highlighted in orange for each category. Participants were allowed to choose more than one option.

The current explosion lacks debris and background noise, impacting its authenticity. Suggestions for improvement include incorporating a more powerful initial impact, increasing low-frequency components, and adding echo and reverberation. In the case of church bells, the evaluation involved a single hit at different frequencies, and recommendations propose

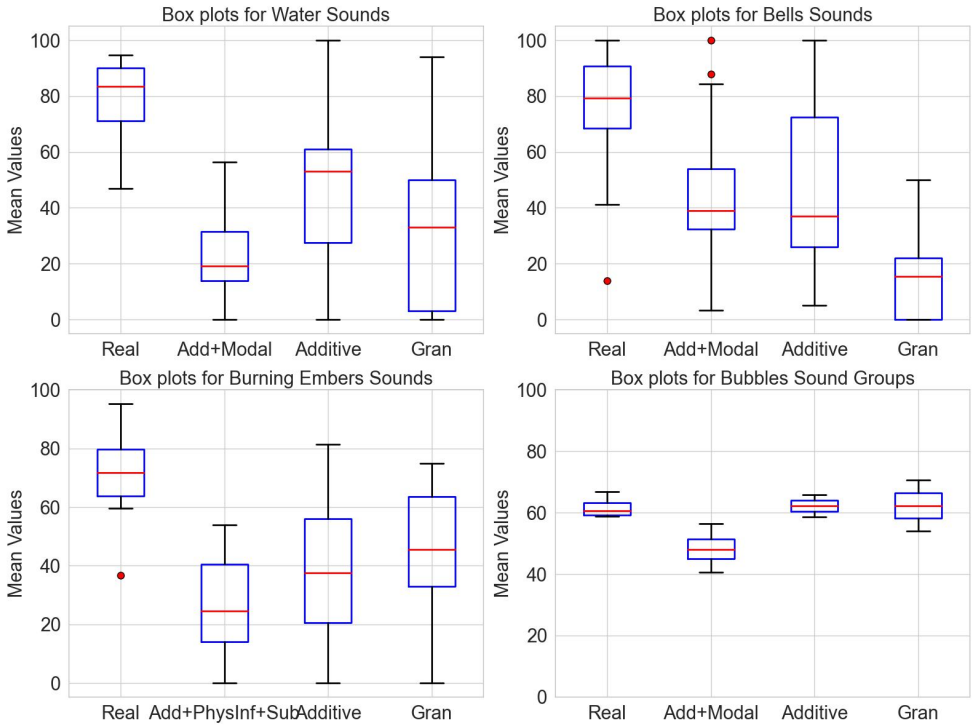


**Fig. 2:** *Burning embers,bells, water and bubbles result distribution.*

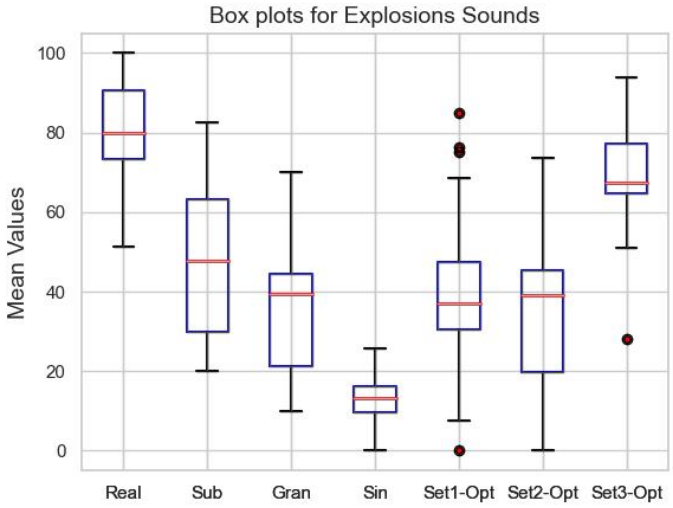


**Fig. 3:** *Explosions result distribution.Set1-Opt,set2-opt and set3-opt representations of the optimization models made by[18]*

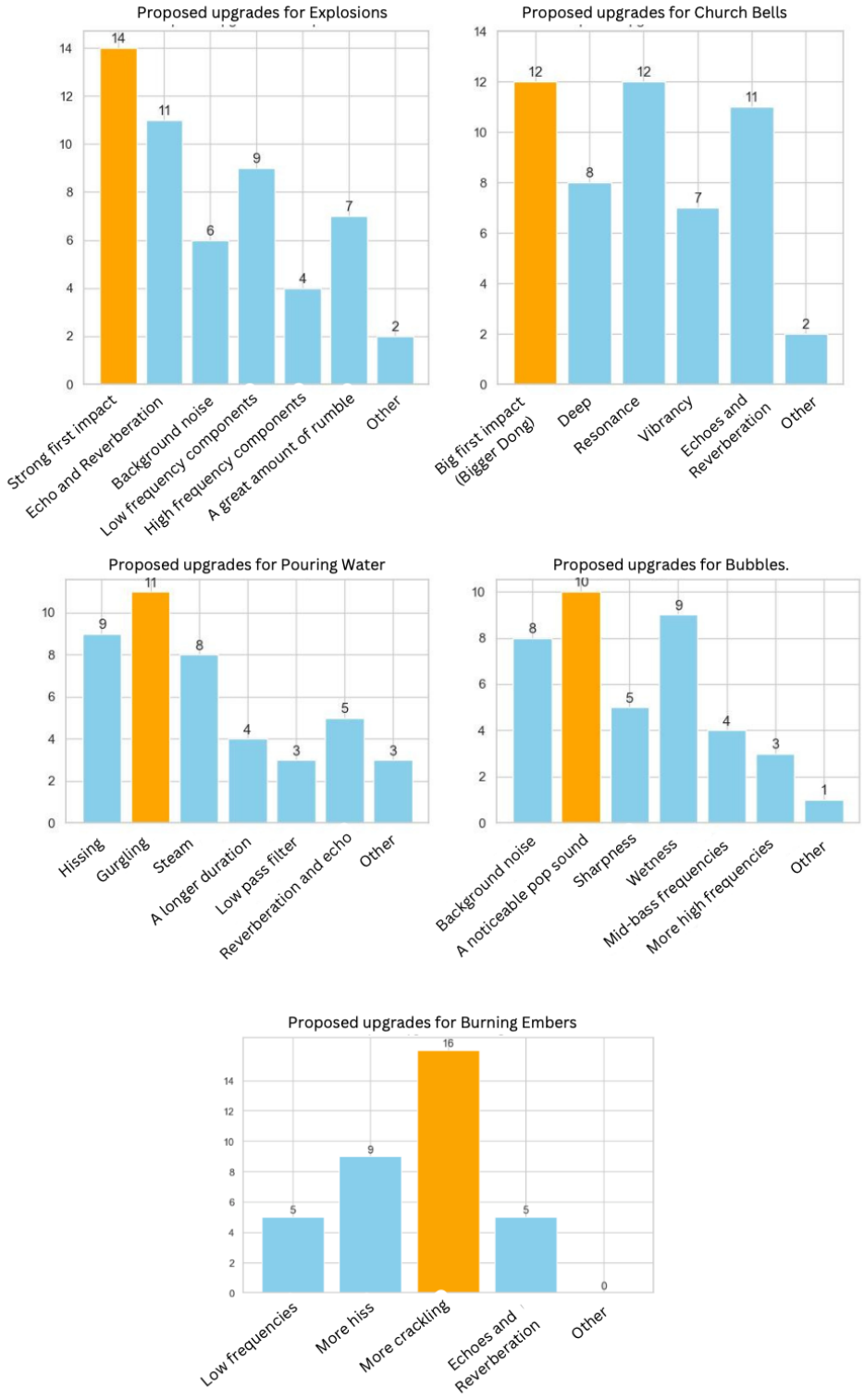


**Fig. 4:** *Burning embers, bells, water, and bubbles result distribution.*

| Group1 | Group 2 | Explosion | Bubbles | Hot Water | Church Bells | Burning Embers |
|---|---|---|---|---|---|---|
| *Granular* | *Real* | **** | o | **** | **** | **** |
| *Granular* | *SubOp 1* | o | . | . | . | . |
| *Granular* | *SubOp 2* | o | . | . | . | . |
| *Granular* | *SubOp 3* | **** | . | . | . | . |
| *Granular* | *Sinusoidal* | *** | . | . | . | . |
| *Granular* | *Subtractive* | o | . | . | . | . |
| *Granular* | *Additive* | . | o | ** | ** | o |
| *Granular* | *Physically Informed* | . | . | . | . | ** |
| *Granular* | *Additive+Modal* | . | o | o | *** | . |
| *Real* | *SubOp 1* | **** | . | . | . | . |
| *Real* | *SubOp 2* | **** | . | . | . | . |
| *Real* | *SubOp 3* | o | . | . | . | . |
| *Real* | *Sinusoidal* | **** | . | . | . | . |
| *Real* | *Subtractive* | **** | . | . | . | . |
| *Real* | *Additive* | . | o | **** | **** | **** |
| *Real* | *Physically Informed* | . | . | . | . | **** |
| *Real* | *Additive+Modal* | . | o | **** | ** | . |
| *SubOp 1* | *SubOp 2* | o | . | . | . | . |
| *SubOp 1* | *SubOp 3* | **** | . | . | . | . |
| *SubOp 1* | *Subtractive* | *** | . | . | . | . |
| *SubOp 2* | *SubOp 3* | **** | . | . | . | . |
| *SubOp 2* | *Subtractive* | *** | . | . | . | . |
| *SubOp 2* | *Sinusoidal* | *** | . | . | . | . |
| *SubOp 3* | *Sinusoidal* | **** | . | . | . | . |
| *SubOp 3* | *Subtractive* | ** | . | . | . | . |
| *Sinusoidal* | *Subtractive* | **** | . | . | . | . |
| *Physically Informed* | *Additive* | . | . | . | . | ** |
| *Additive* | *Additive+Modal* | . | o | *** | o | . |

**Table 3:** Results of pairwise comparison of synthesis methods on perceptual realism rating for each class of sound, $o > 0.05$, $*$ <0.05, $**$<0.01, $***$<0.001, $****$<0.0001, .= no comparison made. SubOp= refers to the different optimizations for the subtractive method.

a more substantial and resonant impact, along with increased echoes and reverberation for synthetic models. For burning embers, enhancements should focus on more hissing, crackling, and the inclusion of additional echoes and reverberation. Regarding pouring hot water, suggestions include introducing more gurgling and hissing sounds, along with increased reverberation, echo, and prolonged duration. The most common improvement mentioned for bubbles is a more pronounced "pop" sound, coupled with an increased sense of wetness in the environment where the bubbles are situated.

## 6 Discussion

The results reveal that subtractive models outperform sinusoidal or granular methods both objectively and subjectively when reproducing explosion sounds. Conversely, additive models, especially those employing modal synthesis, demonstrate significant effectiveness in church bell sounds but face limitations in accurately representing pouring hot water and popping bubbles sounds, where the additive method proves more realistic.

In instances where the previously analyzed sound synthesis method wasn't perceived as the most realistic, further analysis of audio descriptors and perceptual evaluation suggests potential improvements. To enhance realism in church bell sounds, it is recommended to modify the pitch. For synthesizing popping bubbles with the additive and modal methods, adjusting the loudness fluctuation over time emerges as a promising optimization step. When simulating burning embers, the initial optimization steps for the additive, physically informed, and subtractive methods include adjusting duration, loudness (Effective Duration), and pitch in

the spectrum. Regarding the pouring hot water method, participants prefer the additive/modal method, yet their ratings still fall short of the library sample. The hypotheses after the analysis focus on modifying the effective duration (temporal envelope) and spectral flux.

## 7 Conclusions

The analysis methodology presented here offers a versatile framework for evaluating the perceived realism of any sound synthesis method. By systematically examining a diverse set of audio descriptors, this approach serves as a foundational step in identifying key parameters that can be manipulated to enhance the model's realism, akin to the comparison conducted for the procedural engine models. This methodology can be adapted to assess the authenticity and quality of sound synthesis models, providing valuable insights into the factors that contribute to the perceived realism of synthesized audio. The application of this methodology to the evaluation of procedural audio models, as outlined earlier, showcases its adaptability and effectiveness in assessing diverse sound synthesis techniques.

As we delve into the variations of audio descriptors, we pave the way for improved interpretability and advancements in sound quality. The insights gained from this study not only contribute to the refinement of sound synthesis models but also lay the groundwork for future analyses and evaluations. This, in turn, propels the field of sound synthesis forward, expanding its applications and pushing the boundaries of what can be achieved in synthetic audio production.

## References

[1] P. Bahadoran, A. Benito, T. Vassallo, and J. Reiss. FXive: A Web Platform for Procedural Sound Synthesis. . In *144th Audio Engineering Society Convention (AES).*, May 2018.

[2] D. Moffat and J.D.Reiss. Perceptual Evaluation of Synthesized Sound Effects. *In ACM Transactions on Applied Perception. pp 1-19.*, April 2018.

[3] M. Kang K. Choi, S. Oh and B. McFee. A Proposal for Foley Sound Synthesis Challenge. *arXiv preprint*, July 2022.

[4] Rec. ITU-R BS.1534-3. *Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)*. January 2015.

[5] P.D. Pestana D. Menexopoulos and J. Reiss. The state of the art in procedural audio. *Journal of the Audio Engineering Society*, December 2023.

[6] A. Natsiou and S. O'Leary. Audio representations for deep learning in sound synthesis: A review. *In 18th International Conference on Computer Systems. (IEEE/ACS)*, January 2022.

[7] G. Juganaru. A procedural reflection on animation audio. Master's thesis, university of aalto., April 2017.

[8] G. Cochrane and R. Jones. *Granular synthesis and physical modeling with kaivo*. Madrona Labs, December 2016.

[9] K. Piczak. ESC: Dataset for environmental sound classification. *In 23rd ACM Multimedia Conference, pages. 1015–1018*, June 2015.

[10] R. Selfridge D. Moffat and J. D. Reiss. Sound Effect Synthesis. *Foundations in Sound Design for Interactive Media: A Multidisciplinary Approach*, June 2019.

[11] G. Peeters. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. April 2004.

[12] E. Gómez S. Gulati P. Herrera O. Mayor G. Roma J. Salamon J. Zapata D. Bogdanov, N. Wack and X.Serra. Essentia: An open-source library for sound and music analysis. In *21st ACM International Conference on Multimedia*. Association for Computing Machinery, October 2013.

[13] C. Saitis M. Caetano and S. Kai. *Timbre : Acoustics, perception, and cognition, Audio Content Descriptors of Timbre*, pages 297–333. May 2019.

[14] L. Breiman. Random forests. *Machine Learning*, pages pages:5–32, January 2001.

[15] T. Chen and C. Guestrin. *XGBoost: A Scalable Tree Boosting System*, pages 785–794. June 2016.

[16] S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS)*, December 2017.

[17] S. Dixon. Onset detection revisited. *International Conference on Digital Audio Effects(DAFx)*, September 2006.

[18] J. Reiss Y. Zong and N. Garcia. A machine learning method to evaluate and improve soundeffects synthesis model design. *6th International Conference on Audio for Games (AES)*, May 2024.