

Sound effects in media: A comparative analysis of recorded and synthetic samples in live-action and animation.

Nelly Garcia¹[0009–0007–3986–4518] and Joshua Reiss¹

Centre for Digital Music (C4DM), Queen Mary University of London, 327 Mile End Rd, Bethnal Green, London E1 4NS n.v.a.garcia-sihuay@qmul.ac.uk
<https://doi.org/0009-0007-3986-4518> joshua.reiss@qmul.ac.uk
<https://www.qmul.ac.uk/>

Abstract. Creating sound for storytelling is crucial to establishing the environment in productions such as films, TV series and video games. This process often involves repeating, layering and recording real objects or using sound libraries, which can be time-consuming and repetitive. To address these challenges, procedural audio, also known as digital foley, offers a solution by allowing sound designers to quickly generate samples. Despite its efficiency, questions remain about the believability of synthetic samples compared to real ones. In our study, we compared synthetic samples generated by an online procedural engine and integrated them with both animated and live-action visuals. Our results indicate that procedural audio is highly effective and perceived as believable in drama and sci-fi scenes, particularly for sound models such as lasers, hits, air and rockets, whereas synthetic sounds weren't as believable in cartoon productions when representing everyday actions. Finally, we identified specific models that needed optimisation and highlighted audio features that needed improvement with feedback from audio professionals.

Keywords: Sound design · Audio Art and Sonification · Digital Storytelling.

1 Introduction

Sound has two basic roles in film, tv or games. These are story telling and story supporting. Sound is usually exaggerated (‘hyper realism’) to create some reaction to the audience. [7], refers to story telling as the most important characteristic of films by using dialogues, monologues or off-narration. While story supporting relates on the effects that inhence the tension in film and suggest the audience how to feel. [24], states that sound design was properly introduced as a term in a Francis Coppola’s movie. [22], defines sound design as an approach to creating ‘new’ sounds with the intention that they will be heard in a given context of use.

Sound designers’ workflows vary depending on the sources they are working with, such as libraries, Foley, direct sound and re-recording. Foley, named after

Jack Foley, [1] refers to creating the sound of the object using different objects. (e.g. coconut shells for the horse steps).

Sound designers need to consider the period and genre of a story, as well as the venue where it takes place, previous work had interviewed professionals and stated their process [20]. They need to respond to the social, historical and cultural context of the production. This context requires deciding which sounds to include in the soundscape. [17]. Finding the right sound for an object can be time consuming and involve listening to different samples repeatedly. The hypothesis that creating sounds without spending minutes searching for the right one can help sound designers innovate their libraries, the process of sound designed explain as a scientific method [8] has confirmed this issue. For example, the sound of a monster can be created from scratch, offering a wide range of models simultaneously, rather than layering sounds. Procedural audio engines give users a wide range of parameters to modify sounds.

Though the perception of the "believability" of sound design could be subjective and depends on which story the sound designer wants to share. [2], states a division of the different ways a sound design could be created:

- **a. Realistic:** Incorporation of elements that are meant to represent real life. It's important to mention the difference between *realism* and *reality*. [15], described realism as 'the illusion of reality', as a representation, without limiting it to truthfulness or authenticity. Reality on the other hand is real life as we experience it.
- **b. Symbolist:** Interest in communicating an idea to the audience rather than representing real life. E.g a whirling sound when the scene transports you to the mind of the character.
- **c. Fantasy:** Allowance to create a new world. Use of a wider range of sound effects, music or soundscapes to create a magical atmosphere.

A concept related to the perception of reality used nowadays is *hyperrealism*, [12], which means exaggeration of the image of reality, also understood as making the sound bigger than it is perceived in reality. Overall, soundscapes set location or create a mood. Most designers will use a combination of these types of sound to create the design for a production. According to [10] the perceived soundscape of a sound effect depends on two key factors:

1. **The director's perspective:** The director's vision and storytelling are paramount. Directors often describe the desired environment to the sound designer, providing a visual and conceptual framework for how the sound should be perceived.
2. **Audience perception:** The effectiveness of a sound effect is ultimately determined by the audience's perception. The goal is to evoke the intended emotional response and enhance the overall viewing experience.

Once the story and the target audience is decided, we need to discuss the possibilities for creating the audio files that will be in the production. We considered sound synthesis as the artificial creation of sound using analogue, digital or

a combination of both [14]. Procedural audio, also known as digital foley, uses algorithms to generate audio in real time, adapting to changing inputs. Also, procedural audio involves building signal synthesis and processing chains that respond to on-screen events. However, the complexity of implementing DSP algorithms for artistic tasks, combined with the lack of evaluation methods and comparisons, has made this approach less common in the industry, leading to a preference for sample-oriented methods in sound design. Analysing and understanding the characteristics of sounds in both sound synthesis and procedural audio can improve algorithms for different sound categories, establish correlations between algorithms and perceptual concepts to provide objective explainability of the computer processes. This approach provides valuable insights for improvement e.g. [6], did an approach trying to describe the 'whoosh' sound and how it could be improved pointing out the different audio descriptors needed to create from scratch a sound of this type.

This project aims to bridge the gap between research and industry practice by comparing the 'realism' and 'believability' of synthesis methods with traditional library samples already integrated with visuals. Our research questions are:

- *How believable are synthetic sounds compared to traditional library samples when combined with visuals?*
- *In which scenarios do procedural audio models perform best in terms of audience perception?*
- *Which specific models and audio features in procedural audio need further optimisation?*

In this paper we discuss the differences in the perception of realism between synthetic and real(library) audio samples. Section 1.1 introduces the methods and previous work in the field. In Section 2, we outline the experimental design and the dataset used. Sections 3 and 4 deal with the qualitative and quantitative analysis of our results. Finally, Section 5 presents our conclusions.

1.1 Methods

To answer the first and third research questions, our primary goal is to identify low-level audio features that can help audio developers and sound designers improve algorithms for creating sounds from scratch. Audio features are defined as contextual information that can be extracted from an audio signal, they are a statistical or computational representation of an audio file. According to [13], there could be a generalisable objective set of measures for synthesised sounds. His work provided a comparative table of different audio descriptors commonly used in the evaluation of synthetic models which provided interpretability for sound improvement. We define the interpretability as the capability of understanding how the model works, i.e., the process that led to its output [25].

As mention in section 1. procedural audio can be a key aspect in the sound design process, it explores the creativity and has more areas of opportunity:

- *Optimise the design process*: Giving sound designers greater control over procedural audio algorithms can help them to produce better samples more efficiently.
- *Improve procedural audio algorithms*: The identified audio characteristics can guide the creation of sound families that closely match the characteristics of real sounds.
- *Cost-effective sound creation*: Procedural audio models can generate a wide variety of sounds, reducing the need for extensive studio time.

Through our study, we discovered several features that provide insight into specific aspects of sound, allowing sound designers to make relevant improvements. The process of this methodology is shown in figure 1.

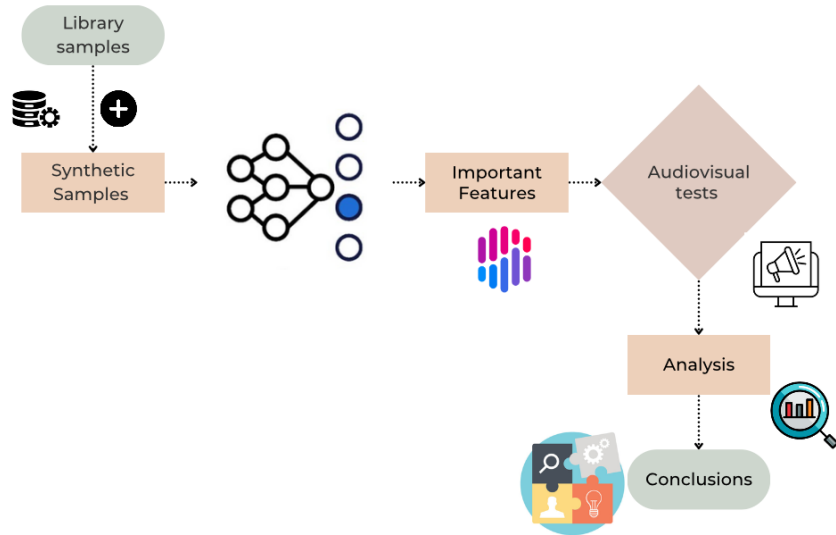


Fig. 1. Process followed to compare the objective and subjective audio descriptors

The first step in the process shown above, was to create a dataset of different videos intended for sound design, which is described in detail in Section 2. We normalised and extracted audio features from these videos, comparing each sound category using a bi-classification machine learning model.

We used XGBoost [5] and Random Forest [4] models to tackle the bi-classification problem, achieving 95% and 90% accuracy respectively. Once the models produced results, we applied SHAP (SHapley Additive exPlanations) [11] and PCA (Principal Component Analysis) [21] to interpret the results. We identified the top four audio descriptors for each sound category and conducted subjective

tests to determine whether these audio features made a noticeable difference to the audience.

For the analysis of the audio features, we used the state-of-the-art reference [23] to classify and distinguish the retrieved audio features in different sound categories.

- *Steadiness or dynamics*: Features represent values extracted from the signal at a given time, such as mean and standard deviation.
- *Temporal extent*: Descriptions apply to specific parts of the sound, such as attack and loudness.
- *Abstractness*: This refers to what the feature represents, such as cepstrum and linear prediction.
- *Extraction process*: Depending on how the data was extracted, features can be
 - Calculated directly on the waveform data.
 - Extracted after performing a signal transform (e.g., Zero Crossing Rate, FFT, Wavelet).
 - Related to the signal model (e.g., source/filter model).
 - Designed to mimic the output of the auditory system (e.g., Bark or ERB bank filter output).

The analysis of the subjective and objective ratings is discussed in sections 5 and 6.

2 Dataset and extraction

We selected eight scenes, split evenly between live action and animated formats. These scenes, can be found in YouTube ¹, had all their original sounds replaced with custom sound designs created specifically for each scene. Each video, which ranged in length from 30 to 60 seconds, covered a range of genres including action, drama, sci-fi and comedy. We used between 50 and 60 different sounds for each sound design. Figure 2 illustrates the selected scenes, with the cartoon scenes at the top and the live action scenes at the bottom.

For sound samples we accessed the BBC Sound Effects library², the Hybrid Sound Library, 50-ESC dataset [16] and the Soundsnap online library³. For the synthesis samples we used Nemisindo⁴, an online procedural audio engine. Each video had specific sound requirements (e.g. for the sci-fi videos we needed space-ship movements and lasers). Table 1. Shows the different synthetic models that were used and combined in the sound design for evaluation. The synthetic models retrieved from the procedural audio engine are shown in the third column. Approximately 16 synthetic sound models were used in all sound designs. Some

¹ <https://www.youtube.com/@nellyngz/videos>

² <https://sound-effects.bbcrewind.co.uk>

³ <https://www.soundsnap.com>

⁴ <https://nemisindo.com>

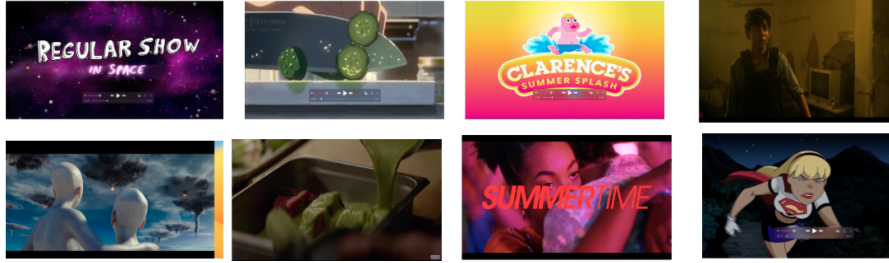


Fig. 2. Scenes for the visual evaluation

synthetic models were modified to create different sounds (e.g. the explosion model was used to create a splash).

The scenes contained different sound synthesis models. Each sound design consisted of between 30-40 sounds between real and synthetic samples. Table 2 shows which and how many synthetic models were used in each sound design.

Scene	Type-Genre	Synthetic Sounds
<i>Space War</i>	Sci-fi Live-action	Lasers, whoosh, rocket, drill, mammal sound, bells, spaceship movements
<i>Space Adventure</i>	Sci-fi Cartoon	Lasers, rocket, drill, mammal sound, bells, machinery, increasing beams, space cannons
<i>Rescue Mission</i>	Action Live-action	Gunshot, hits, impact, debris, body hits, explosions, clothes movements
<i>Superhero Fight</i>	Action Cartoon	Fire embers, crackling, hits, impact, debris, body hits, explosions, clothes movements
<i>Ramen</i>	Drama Cartoon	Hot water, whoosh, droplets, fire embers, machinery
<i>Kitchen</i>	Drama Live-action	Hot water, whoosh, droplets, fire embers, machinery
<i>Vacations</i>	Drama/Comedy Cartoon	Splash, waves, hit, air, birds, whoosh, spray, bubbles, droplets
<i>Summertime</i>	Drama/Comedy Live-action	Splash, waves, hit, air, birds, body hit, whoosh, spray, bubbles, droplets

Table 1. Scenes used in the evaluation of the sound design and synthetic sounds used in the video.

Each scene had three types of sound design, which are listed below:

- **Real sound design:** This sound design contained only library samples.

- **Synthetic sound design:** This sound design contains a mix of library and synthetic samples.
- **Bad sound design:** This sound design contained random sounds that were not synchronised with the image. The purpose of this sound design is to filter the participants, to give them the context of the story being revealed in the scene.

3 Sound designs objective analysis

For its analysis, we created a dataset of 50 samples for each sound category and compared them to the library samples we had access to. The dataset consists of a total of 1,616 sounds divided into 32 categories (16 synthetic and 16 library samples). Each category varies from 2.5% to 2.8% to achieve balance.

The costumed dataset previously built with the procedural online engine and the libraries mentioned above in section 3. All of our samples for analysis were downmixed to mono, 16-bit depth, 44.1 Khz and trimmed to 5 seconds.

Following the criteria as the [23] project and [13], we determined the type of audio descriptors to be extracted. We used the Essentia framework ⁵ [3] for feature extraction, with a total of 78 low-level features for each sound sample. These descriptors were carefully selected and combined to extract the most critical features that allow the machine to distinguish between synthetic and realistic sounds.

After evaluating the performance of the classification model, we identified the four best audio descriptors for each sound category. Table 2 shows the top audio descriptors for each sound category, divided by the feature domain where each audio descriptor can be found. The last column shows the general classification of the feature, as explained earlier in this section.

These descriptors help to understand how synthetic sounds produced by models can be compared to physical sounds. Audio descriptors belonging to the temporal and frequency domains are highlighted in green, indicating that changes in these domains could improve the perception of these synthesis models. Descriptors highlighted in yellow belong to the temporal and spectral domains, while those highlighted in purple are purely temporal features.

1. *Temporal shape:* features(global or instantaneous),computed from the waveform or the signal envelope(envelop).
2. *Spectral features:* computed by the STFT,centroid, spread, skewness, roll off frequency

The previous listed sound descriptors were extracted with Essentia. To make more sense in how this audio features were extracted, the meaning of each top feature are listed below:

- **TCToTotal:**This algorithm calculates the ratio of the temporal centroid to the total length of a signal envelope, indicating how 'balanced' the sound

⁵ <https://essentia.upf.edu>

Category	Audio Features	Feature Domain	Classification
<i>Fire</i>	High Frequency Effective Duration Flatness SFX	Temporal and Frequency	Extraction and Time Extent
<i>Rain</i>	TCToTotal, Effective Duration, Kurtosis, Flatness SFX	Temporal	Dynamicity and Time Extent
<i>Waves</i>	TCToTotal, Log Attack Time, Flatness SFX, Strong Peak	Temporal and Frequency	Extraction and Time Extent
<i>Wind</i>	Log Attack Time, TCToTotal, MaxToTotal, Pitch Salience	Temporal	Extraction and Time Extent
<i>Engine</i>	Pitch Salience, Attack Start, TCToTotal, Flatness SFX	Temporal and Frequency	Extraction and Time Extent
<i>Applause</i>	Intensity, TCToTotal, MaxToTotal, Pitch Salience	Temporal and Frequency	Extraction and Dynamicity
<i>Gunshot</i>	TCToTotal, MaxToTotal, ZCR, High Frequency	Temporal and Frequency	Extraction and Dynamicity
<i>Helicopter</i>	TCToTotal, Intensity, Effective Duration, High Frequency	Temporal and Frequency	Dynamicity and Time Extent
<i>Thunder</i>	Roll Off, TCToTotal, ZCR, MaxToTotal	Temporal and Spectral	Extraction and Time Extent
<i>Crackling</i>	Log Attack Time, TCToTotal, Attack Stop, Effective Duration	Temporal	Time Extent
<i>Fireworks</i>	ZCR, Roll Off, TCToTotal, Kurtosis	Temporal	Extraction and Time Extent
<i>Bells</i>	TCToTotal, Pitch Salience, MaxToTotal, ZCR	Temporal	Extraction and Time Extent
<i>Pouring Hot Water</i>	TCToTotal, MaxToTotal, Roll Off, ZCR	Temporal and Spectral	Extraction and Time Extent
<i>Explosion</i>	TCToTotal, MaxToTotal, High Frequency, Roll Off	Temporal and Spectral	Extraction and Time Extent
<i>Bubbles</i>	ZCR, TCToTotal, Flatness, MaxToTotal	Temporal	Extraction and Time Extent

Table 2. Top 4 Audio Features Retrieved by Machine Learning Methods and SHAP Values

is. Its value is close to 0 if most of the energy is at the beginning of the sound (e.g. decrescendo or impulsive sounds), close to 0.5 if the sound is symmetrical (e.g. 'delta unvarying' sounds), and close to 1 if most of the energy is at the end of the sound (e.g. crescendo sounds).

- **High Frequency:** Content of a spectrum, The magnitudes of the spectral bins are added together, but multiplying each magnitude by the bin "position" (proportional to the frequency). Thus if $X(k)$ is a discrete spectrum with N unique points, its high frequency content measure is represented in equation 1:

$$HFC = |X(n)|^2 \cdot k \quad (1)$$

- **Flatness SFX:** There are two thresholds defined: a lower one at 20 percent and an upper one at 95%. The thresholds yield two values: one value which

has 20% of the total values underneath, and one value which has 95% of the total values underneath. The flatness coefficient is then calculated as the ratio of these two values.

- **Effective Duration:** Effective duration of an envelope signal. The effective duration is a measure of the time the signal is perceptually meaningful. This is approximated by the time the envelope is above or equal to a given threshold and is above the -90dB noise floor. This measure allows to distinguish percussive sounds from sustained sounds but depends on the signal length.
- **Kurtosis:** Measure of the "tailedness" of the probability distribution of a real-valued random variable. Describes a particular aspect of a probability distribution.
- **LogAttackTime:** Calculates the logarithm (base 10) of the attack time of a signal envelope. The attack time is the time from when the sound becomes perceptually audible to when it reaches its maximum intensity. By default, the start of the attack is estimated at the point where the signal envelope has reached 20% of its maximum value, to account for possible noise, and the end is estimated at the point where the envelope has reached 90% of its maximum value, as in trumpet sounds.
- **PitchSalience:** This algorithm computes the pitch salience of a spectrum. The pitch salience is given by the ratio of the highest auto correlation value of the spectrum to the non-shifted auto correlation value. Pitch salience was designed as quick measure of tone sensation.
- **Zero Crossing Rate:** It is the number of sign changes between consecutive signal values divided by the total number of values. Noisy signals tend to have higher zero-crossing rate.
- **MaxToTotal:** Computes the ratio between the index of the maximum value of the envelope of a signal and the total length of the envelope. This ratio shows how much the maximum amplitude is off-center.

4 Experiment Design

Our subjective evaluation consisted of two parts. The first part was to evaluate the perception of the final videos and the second part was finding which sound effects weren't perceived as believable and need to be optimized. We created a MUSHRA-inspired test using WebMUSHRA [18], 4 sliders were used to rate the different sound designs as shown in figure 3. Our rating was from 0-100 with the following scale:

- **100 - Good sound design.** (The samples help to create a real soundscape.)
- **75 - Mostly good sound design.** Not that real(The samples do not help to create a real soundscape.)
- **50 - Average sound design** (The samples used are mostly synthetic, they do not help to create a real soundscape.)
- **25 - Somewhat bad sound design** (The samples do not help to create a real soundscape.)



Fig. 3. WebMUSHRA Test with no reference and three types of sound design

- **0 - Bad sound design** (The samples are not coherent with the visuals.)

And for the second part of the evaluation, we designed the Google form⁶ where participants could indicate which sounds did not seem believable and suggest aspects that needed to be changed or could improve perception.

5 Quantitative Evaluation

For the webmushra quantitative evaluation. We had 20 participants, 11 male and 9 female, who performed the online test. All participants had from 3-7 years of experience in the audio industry. Overall, they are sound designers(40%), audio researchers(30%), developers(15%) and musicians(15%). The answers of two participants were discarded, because the rating for bad sound design was ≥ 25 .

With the 18 remaining responses, we performed an ANOVA test, the p and f values are presented in the table 3 below, highlighted in bold are the sound designs' p -value less than 0.05. We found statistical importance difference in 5 out of the 8 sound designs. Two of these sound designs are cartoon visuals and three are live action scenes. At the same time, we performed the Bonferroni correction[19] to reduce the probability of making errors, following $\alpha = 0.05/8$, where the numerator represents the value of alpha and the denominator is the number of cases, this lead to a change of alpha to 0.00625 to better understand the

⁶ <https://docs.google.com/forms/d/e/1FAIpQLSd4IwgM0plWo2ug5Odu89mgm3yYfWCrrwrR1e75-iryGI3aw/viewform>

statistical difference. With this correction we found an statistical difference in 2 main scenes that represented daily routine sounds (Drama-kitchen scene) and the sci-fi cartoon scene where more imaginative sounds were used. As, mentioned in section 1, realism is defined as the representation of life without limiting it to truthfulness or authenticity. This results point out that even though synthetic sound could be used easily in cartoons, if the cartoons mimic a daily routine, audiences tend to spot the synthetic sound easily, whereas in sci-fi, the hyperrealism or tendency to make a sound bigger could point out why synthetic samples weren't perceived as believable.

Scene	p-value	f-value	Bonferroni correction $\alpha = 0.00625$
Action (LA)	0.00283	1.408	0.02
Action (C)	0.015	9.14	0.12
Drama (LA)	0.000092	6.05	0.0007
Drama (C)	0.00000143	7.5	0.00001
Drama/Comedy (LA)	0.0016	1.66	0.013
Drama/Comedy (C)	0.114	8.66	0.919
Space battle (LA)	0.0017	5.3	0.013
Space (C)	0.0001	5.4	0.0008

Table 3. p and f values of the ANOVA with the Bonferroni correction with $\alpha=0.00625$ tests performed on the webmushra ratings. LA: Live-action. C: Cartoon

For a better visualisation of this analysis, figure 4 represents the different rating in the corresponding scenarios. The x-axis labels represents the different sound designs used in the test, while the y-axis shows the ratings (0-100). It is easy to see that real sound designs were perceived as more believable for both cartoon and live action. Although the mean averages in cases like the summer holiday cartoon and the action live-action scenes aren't far apart. Performing the confidence levels we found out the maximum of 34.8% of difference in the summer vacation scene between the real and the synthetic, and the least difference was found in action genre in live action with 13.8%. An important observation is that the participants in the cartoon scenes rated the scenes between 50-70, which means that the models needed to be optimised or that people subconsciously think that cartoon sounds are not real because they don't associate them with a real object.

6 Qualitative Evaluation

To understand more the results shows in figure 4 we performed a more qualitative form of evaluation. Two participants' responses were discarded because they said they did not remember the sound effects at all and not feel comfortable taking them into account.

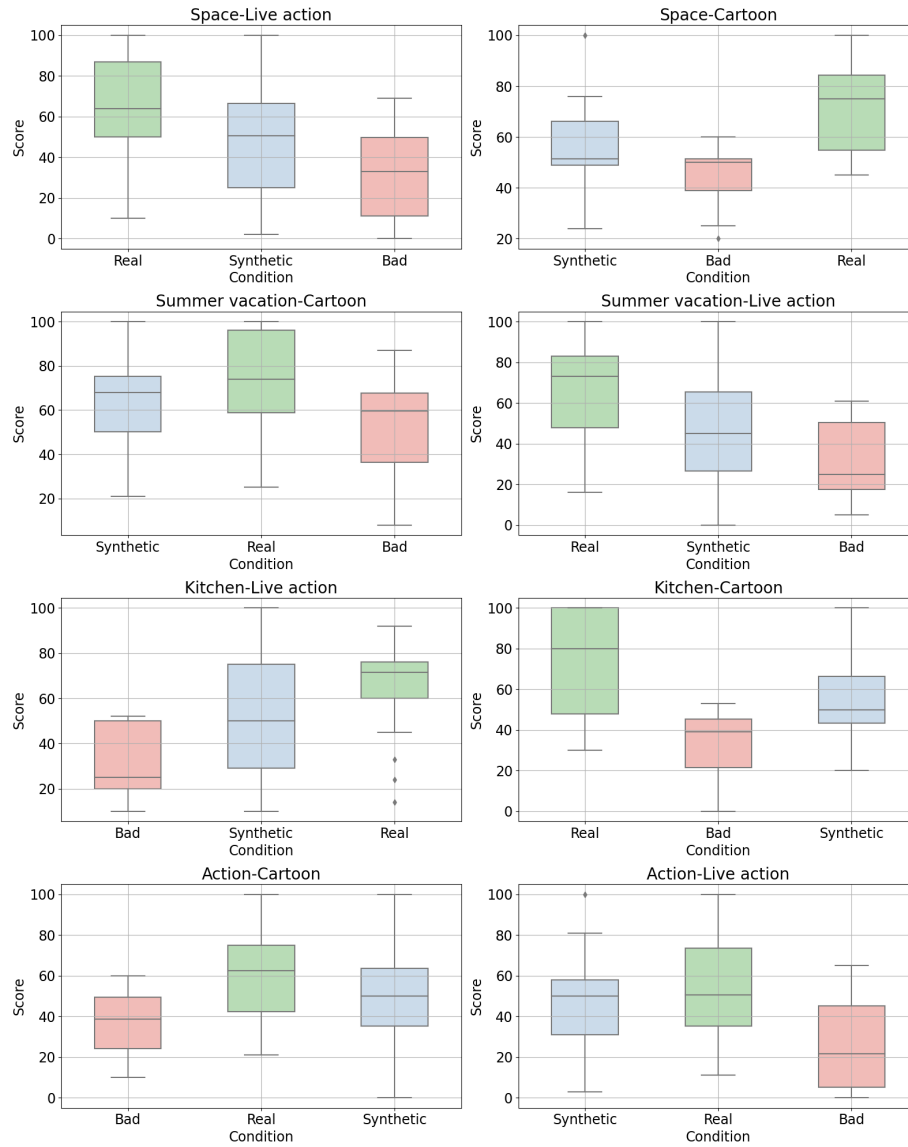


Fig. 4. Boxplot of the 3 different sound designs in the different scenes.

Participants had to choose 2 models that they did not find realistic in the sound designs, followed by a multiple choice question about which aspect could be optimised to make the sound more believable. Figure 5 shows the different sounds chosen by the participants. The orange bar represents the live action, while the blue bar represents the animated scenes. We can see that there was less choice in the sci-fi (space) scene, where only 4 sound effects were seen. We can also see that although the laser was less credible in the animation, this doesn't apply to the live-action use, where the spaceship's movements were the most selected feature. However, there was more agreement that the cartoon sounds did not quite hit the mark, one hypothesis being that they weren't big enough to achieve the hyper-realism that the audiovisual project could have had. While more options were selected in the drama and action scenes, it was more obvious that the synthetic samples were easier to identify in the cartoon, leading to the hypothesis that the object shown in the visuals and the sound effect did not manage to get the "added value".

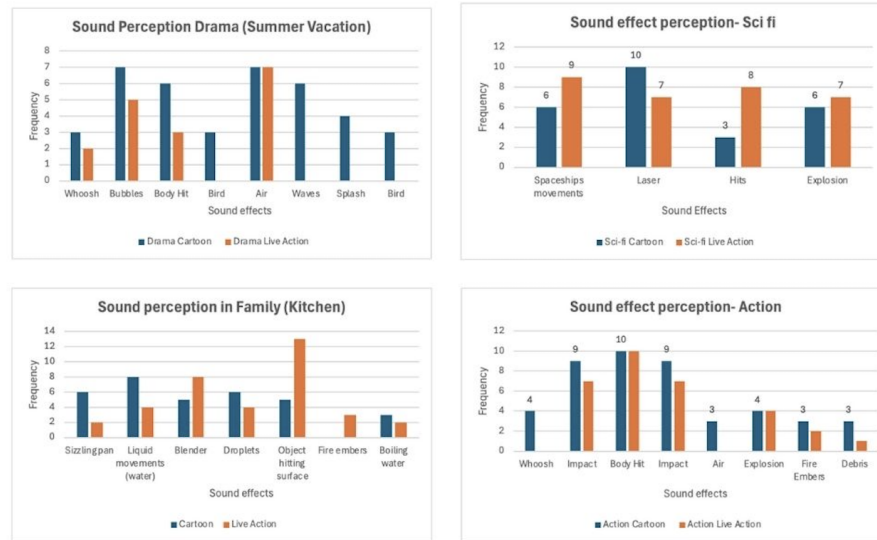


Fig. 5. Barplots of the different sound suggestions of the participants

We can see that the choice of sound supports the results of the webMUSHRA test and that there is a difference between cartoon and live action. We performed basic statistics such as mean, median, standard deviation, confidence levels, etc., and found variability in the participants' scores. The scenes with the most selected sounds were the drama/comedy and action scenes, each of which had 8 selected sound models, followed by the drama scene and finally the sci-fi scene. Our results did not come as a surprise, as people tend to match what they see

with what they hear, according to the "added value" that sound can give to an image. For example, sci-fi sounds are mostly synthetic, but the hyper-realism of Hollywood productions means that sci-fi sounds require more post-production effects. Also, when it comes to drama or action scenes, people already have an idea of how certain sounds might be heard.

However, these tests help us to highlight other aspects that have not yet been explored in the field of sound effects. Because of the lack of pure sound effects creations or tracks without music. When asked which aspects could be changed in their models, the participants selected them. We correlated the aspects with the audio features mentioned in section 2. Table 4 shows the most selected synthesis models in each genre, highlighting the participant's opinion in which aspect the model needs to be optimised and the feature domain to which this aspect correlates and which can be found in Table 2.

Scene	Sound	Aspect	Feature domain
<i>Action (LA)</i>	Body hit	More low frequencies	Frequency
	Impact	More low frequencies	
	Explosion	Pitch modulation	
<i>Action (C)</i>	Body hit	More low frequencies	Frequency
	Impact	More low frequencies	
	Debris	Pitch modulation	
<i>Drama (LA)</i>	Object hit	More layering	Frequency
	Blender	More low frequencies	
	Liquid	Pitch modulation	
<i>Drama (C)</i>	Liquid	Emphasize high frequencies	Frequency
	Sizzling pan	Emphasize high frequencies	
	Droplets	Stronger first impact	
<i>Drama/Comedy (LA)</i>	Waves	More layering	Temporal and frequency
	Bird	Equalization	
	Air	More low frequencies	
<i>Drama/Comedy (C)</i>	Bubbles	More reverberation	Spectro-temporal
	Air	Pitch modulation	
	Waves	Stronger first impact	
<i>Sci-fi (LA)</i>	Spaceship	Emphasize high frequencies	Temporal and frequency
	Hits	More low frequencies	
	Explosions	Stronger first impact	
<i>Sci-fi (C)</i>	Spaceship	More layering	Temporal and frequency
	Laser	More reverberation	
	Hits	Stronger first impact	

Table 4. Aspects to be changed in each sound effect and the classification for optimization. C: Cartoon. LA: Live-action

Table 4 helps to see how the use of the synthetic samples is perceived, and highlights the feature domain where sound designers might start looking to optimise their models. These initial experiments have provided valuable insights into the features that describe and distinguish realistic from synthetic sounds. For

example, we know that for the synthetic model of an explosion we would need to pay attention to the frequency of the first impact to create a more believable sound, and for waves we would need to pay attention to the duration and provide more than one audio file to achieve the desired realism.

7 Discussion and conclusions

Procedural audio models offer easy access and utility for sound designers. While their popularity is growing within the community, there is still a gap between industry standards and developers when it comes to optimising these models.

Our evaluation revealed different perceptions of audio realism between animated and live-action scenes. Certain synthesis models were found to be less believable, while others showed potential for optimisation.

1. **Contextual believability:** Metrics of believability vary depending on the intention and context of the audiovisual project. For example, in scenes simulating daily routines, procedural audio samples were perceived as less real than non-daily sounds. In both animated and live-action drama scenes, believability was lower than in action scenes.
2. **Genre-specific performance:** Some synthetic models performed better in live-action sci-fi films than in cartoons. This suggests that genre and context have a significant impact on the perceived realism of synthesised audio.

There is no official evaluation method for sound synthesis in audio visual projects yet, though [9] reviewed two methods used in virtual acoustic spaces, evaluating the design tools process. Sound design relies heavily on the sound designer’s ability to manipulate audio samples to fit the narrative of the story. Individual perceptions and biases play a role in determining how each on-screen object should sound. Despite these differences, our study provides a valuable database and starting point for sound designers to improve their models.

Finally we could answer our main research question in 3 points

- **Perception metrics:** We found that credibility metrics depend on the context and goals of the audiovisual project. This finding addresses our main research question.
- **Performance scenarios:** We identified scenarios where synthetic samples outperformed real ones, answering our second research question.
- **Model optimisation:** Through online listening tests and feedback correlation, we identified at least three sound effects that require optimisation. This feedback highlights the audio features that need better tuning or theoretical adjustments to improve perceived realism.

Our research highlights the importance of context and genre in the perception of procedural audio samples. The findings provide a basic understanding for audio developers, guiding them on which features to optimise for enhanced realism and believability in different audiovisual contexts.

8 Future work

By deepening our understanding of procedural audio models, we can bridge the gap between industry standards and developer capabilities, leading to more effective and persuasive sound design in different audiovisual contexts.

Currently, sound designers face challenges in understanding and manipulating procedural audio samples. The use of machine learning and neural networks for feature importance analysis could help identify the critical features of each sound category, providing sound designers with robust tools for refinement. Further psychoacoustic testing should be carried out to understand the difference in perception between synthetic and live, between live action and cartoon.

In addition, expanding the dataset to include a wider range of sound categories will improve the insights available to developers and give sound designers confidence that procedural audio models can reliably generate believable, high-quality sounds.

References

1. Ament, V.T.: The Foley grail: The art of performing sound for film, games, and animation. Routledge (2014)
2. AQA: A-level drama and theatre: Sound design guidance (2020), <https://filestore.aqa.org.uk/resources/drama/AQA-7262-SOUND-D-TG.PDF>, accessed: 2024-07-19
3. Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., Serra, X.: Essentia: an audio analysis library for music information retrieval (11 2013)
4. Breiman, L.: Random forests. Machine Learning pp. 5–32 (01 2001)
5. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System, pp. 785–794 (06 2016)
6. Cherny, E., Lilius, J., Mouromtsev, D.: A method for automatic whoosh sound description. In: Torin, A., Hamilton, B., Bilbao, S., Newton, M. (eds.) Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17), Edinburgh, UK, September 5–9, 2017. p. 459–465. University of Edinburgh (2017)
7. Dakic, V.: Sound Design for Film and Television. GRIN Verlag (2009), <https://books.google.co.uk/books?id=wSVPl1TSOm8C>
8. Hug, D., Misdariis, N.: Towards a conceptual framework to integrate designerly and scientific sound design methods (09 2011)
9. Liljedahl, M., Fagerlönn, J.: Methods for sound design: a review and implications for research and practice. Association for Computing Machinery, New York, NY, USA (2010)
10. Lopez, M., Kearney, G., Hofstädter, K.: Seeing film through sound: Sound design, spatial audio and accessibility for visually impaired audiences. British Journal of Visual Impairment **40** (05 2020)
11. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS) (2017)
12. Martusciello, F.: The reality of the reproduction. aesthetics of a "conscious" approach to sound design in the soundscape composition: a case study. Association for Computing Machinery, New York, NY, USA (2022)

13. Moffat, D., Reiss, J.: Objective evaluations of synthesised environmental sounds. (09 2018)
14. Moffat, D., Selfridge, R., Reiss, J.: Sound Effect Synthesis, pp. 274–299 (06 2019)
15. Murray, L.: Authenticity and realism in documentary sound. *The Soundtrack* **3**, 131–137 (12 2010)
16. Piczak, K.J.: Esc: Dataset for environmental sound classification. p. 1015–1018. Association for Computing Machinery, New York, NY, USA (06 2015)
17. Rocchesso, D., Bresin, R., Fernström, M.: Sounding objects. *Multimedia, IEEE* **10**, 42–52 (05 2003)
18. Schoeffler, M., Bartoschek, S., Stöter, F.R., Roess, M., Westphal, S., Edler, B., Herre, J.: webmushra — a comprehensive framework for web-based listening tests. *Journal of Open Research Software* **6** (02 2018)
19. Sedgwick, P.: Multiple significance tests: the bonferroni correction. *BMJ* (online) **344** (01 2012)
20. Selfridge, R., P.S.: Investigating the sound design process: two case studies from radio and film production. In: Design Research Society, DRS Digital Library (2022)
21. Shlens, J.: A tutorial on principal component analysis. *CoRR-Computer science bibliography* (2014)
22. Susini, P., Houix, O., Misdariis, N.: Sound design: An applied, experimental framework to study the perception of everyday sounds. *The New Soundtrack* **4**, 103–121 (09 2014)
23. Vinet, H., Herrera, P., Pachet, F.: The cuidado project. (01 2002)
24. Winters, P.: Sound design for low and no budget films. Routledge (2017)
25. Zinemanas, P., Rocamora, M., Miron, M., Font, F., Serra, X.: An interpretable deep learning model for automatic sound classification. *Electronics* **10**, 850 (04 2021)