Perceptual Evaluation of Synthesized Sound Effects

DAVID MOFFAT and JOSHUA D. REISS, Queen Mary University of London

Sound synthesis is the process of generating artificial sounds through some form of simulation or modelling. This article aims to identify which sound synthesis methods achieve the goal of producing a believable audio sample that may replace a recorded sound sample. A perceptual evaluation experiment of five different sound synthesis techniques was undertaken. Additive synthesis, statistical modelling synthesis with two different feature sets, physically inspired synthesis, concatenative synthesis, and sinusoidal modelling synthesis were all compared. Evaluation using eight different sound class stimuli and 66 different samples was undertaken. The additive synthesizer is the only synthesis method not considered significantly different from the reference sample across all sounds classes. The results demonstrate that sound synthesis can be considered as realistic as a recorded sample and makes recommendations for use of synthesis methods, given different sound class contexts.

CCS Concepts: • **Human-centered computing** → **Auditory feedback**; • **Applied computing** → **Sound and music computing**; • **Information systems** → *Multimedia and multimodal retrieval*;

Additional Key Words and Phrases: Sound synthesis, evaluation, procedural audio, perception, sound effects

ACM Reference format:
David Moffat and Joshua D. Reiss. 2018. Perceptual Evaluation of Synthesized Sound Effects. *ACM Trans. Appl. Percept.* 15, 2, Article 13 (April 2018), 19 pages.
https://doi.org/10.1145/3165287

1 INTRODUCTION

Sound synthesis is the process of computer generation of audio. Synthesized sound effects can be applied to a range of sound design fields including film, TV, video games, virtual reality, and augmented reality (Merer et al. 2013).

The field of sound synthesis has significant work in a range of areas including effective and efficient replication of existing sounds or creation of new sounds. Synthesis techniques can be divided into the following three categories

> *Sample-based models.* Audio recordings are cut and spliced together to produce new or similar sounds (Fröjd and Horner 2009).

This work was supported EPSRC grant number EP/M506394/1.

Authors' addresses: D. Moffat and J. D. Reiss, Center for Digital Music, School of Electronic Engineering & Computer Science, Queen Mary University of London, 10 Godward Square, London, E1 4FZ; emails: {d.j.moffat, joshua.reiss}@qmul.ac.uk.

ACM 1544-3558/2018/04-ART13 $15.00
https://doi.org/10.1145/3165287

> *Signal-based models.* Sounds are created, based on some analysis of real world sounds, and resynthesis of the waveform, not the underlying physical system (Pampin 2004).
>
> *Physical models.* Sounds are generated based on modeling of the physics of the system that created the sound. The more physics incorporated into the system, the better the model is considered to be Bilbao (2009).

A clear overview of synthesis research is presented in Misra and Cook (2009). The aims of current sound synthesis research include producing realistic or controllable systems for artificially replicating real-world sounds. The primary focus is on implementation efficiency (Horner and Wun 2006), interfacing and interaction control (Nordahl et al. 2010), or producing accurate models of the physical environment (Bilbao and Chick 2013).

There are many different forms that evaluation of a sound synthesis system can take. Jaffe (1995) presented 10 different methods for evaluation of synthesis techniques, and many of these methods have been implemented in the literature. For example, existing literature performs evaluation of controls and control parameters (Rocchesso et al. 2003; Merer et al. 2013; Selfridge et al. 2017b), human perception of different timbre (Merer et al. 2011; Aramaki et al. 2012), sound identification (Ballas 1993; McDermott and Simoncelli 2011), and sonic classification (Gabrielli et al. 2011; Hoffman and Cook 2006a; Moffat et al. 2017).

This work proposes to evaluate sounds produced by a synthesis system and compare them against recorded samples in the same contextual environment. This facilitates direct comparison and helps establish if a particular synthesis method can be considered indistinguishable from a recording of the intended sound. In this context, there may be instances where a synthesis method would be beneficial for use in a professional capacity, since there are typically more direct ways to control the sonic properties of a synthesis method than of a sample. The ability to produce realistic, real-time synthesized sounds is considered a challenging (Miner and Caudell 2005) and unsolved problem (Caramiaux et al. 2014).

Evaluation of physical models is a difficult task, specific to physical modelling. The complex nature and detail of some physical models makes it challenging to compare these to more general sample-based or signal-based synthesis methods. As such, evaluation of physical models is beyond the scope of this work, but physically inspired synthesis will be used for evaluation purposes.

The aim of this work is to highlight the deficits of current research and to provide insight into which synthesis methods are most effective given a specific context. Through better understanding of the perceptual realism of a range of synthesis methods, on a range of different sounds, we hope to highlight particular sound classes or contexts that would benefit from further work. The current literature on synthesis evaluation will be presented in Section 2. Section 3 will present the range of synthesis methods to be evaluated. The listening test set-up will be presented in Section 4 and the results presented in Section 5. An evaluation of the results and discussion of the impact of these results will be presented in Section 6. Section 7 will present conclusions and further work.

2 BACKGROUND

There is a wealth of synthesis evaluation research. Schwarz (2011) noted in a review of 94 published articles on sound texture synthesis that only 7 contained any perceptual evaluation of the synthesis method. Ten criteria for evaluating sound synthesis are presented by Jaffe (1995). Five of the criteria are based on the parameter control, three on computation of the synthesis method, and two on the sonic qualities of the synthesis method. This framework for evaluation was used by Tolonen et al. (1998), who also produced a rigorous review of a range of synthesis methods. Despite all this work, there is no consistently used standard process for evaluating the perceptual realism of sound synthesis.

Bonebright et al. (2005) discussed three different methods for determining perceptual qualities of audio, through identification testing, context-based rating, or attribute rating. They went on to discuss that context rating is most appropriate for sound in video games and sound synthesis. This work is discussed by Merer et al. (2011), who proposed that for synthesis of abstract sounds, an attribute rating was most appropriate.

Some work performed a range of perceptual tests. McDermott and Simoncelli (2011) ran a series of perceptual tests, where users were asked to perform an identification task where they needed to pick the right description of the sound from a set of five words. They then performed a context test, where an original sound was played and users had to select which of two options sounded most like the reference, and both options were different synthesized sounds. Participants were then asked to provide a rating for realism on a scale of 1–7 for a range of different synthesized and recorded sounds. No formal anchors were identified as the lower bounds for the sound quality.

However, most perceptual evaluation takes the form of a single test. An evaluation of concatenative synthesis methods was performed via an online MUltiple Stimulus Hidden Reference and Anchor (ITU-R BS.1534-3 2015) (MUSHRA) style listening test, in which participants rated the quality of samples and similarity to the reference sample (Schwarz et al. 2016). There was no randomization of sample order, so potential ordering bias may be an issue, and no recording of the participants' listening conditions was made. They concluded that all concatenative synthesis methods are indistinguishable from each other, in terms of the perceived quality of the sound produced and in terms of realism. A similar evaluation methodology was undertaken by Mengual et al. (2016), where different synthetic weapon sounds were evaluated with order randomization to remove bias and performed in controlled listening conditions. The conclusion was that modal sounds were synthesised convincingly, but broadband sounds needed further work to improve.

Synthesis of sword swing sounds through a similar evaluation structure, comparing to multiple different synthesis methods, recorded samples and a specific anchor, was undertaken by Selfridge et al. (2017b). Objective evaluation was also performed, through inspection and discussion of spectrogram plots. This work was extended to a range of other aeroacoustic sounds, with the same evaluation methodology (Selfridge et al. 2017a, 2017c, 2017d, 2017e). An attribute test was performed by Murphy et al. (2008), where participants were asked to rate the quality of "rollingness" of synthesized rolling sounds, in a MUSHRA style test, but no alternative synthesis methods, samples, or hidden anchors were provided for comparison. Participants were asked to browse through a range of synthesized sounds to find their preferred sound and then asked to rate the perceived realism on a 7-point Likert scale in Rocchesso and Fontana (2003). "Perceived realism" was also the evaluation criterion in Böttcher and Serafin (2009) and McDermott and Simoncelli (2011).

An alternative form of evaluation of synthesis was proposed by Gabrielli et al. (2011), an "RS Test" where participants were played a single sound only once and had to determine if it was real or synthetic. This test then iterates through a large number samples and insists on the use of an anchor or "acid sample." Hahn (2015) evaluated musical instrument sounds using the RS test. To evaluate instrument synthesis, Järveläinen et al. (2002) asked participants to match real and synthesized equivalents of samples together based on the harmonic components.

There are a range of methods for objective evaluation of synthesized sound effects, However, there is little to no consistency on objective metrics to use. Horner and Wun (2006) objectively compared different wavetable synthesis methods using "Relative Spectral Error," with no comparison to samples. In contrast, Hendry and Reiss (2010) compared a synthesis method to reference samples, through visual comparison of spectrograms, and comparison of low-level audio features, such as fundamental frequency, first four harmonic frequencies, spectral centroid, and zero crossing rate. But no comparison with other synthesis methods was undertaken.

Hamadicharef and Ifeachor (2003) proposed evaluating sound using Perceptual Evaluation of Audio Quality (PEAQ) (Thiede et al. 2000). PEAQ is an algorithm designed for determining the quality of audio compression codecs, which analyses the sound on a sample by sample basis to determine any perceptual artifacts. This work was further developed by use PEAQ to select parameters for a piano synthesizer to replicate an input audio signal (Hamadicharef and Ifeachor 2005). But the notes will never be exactly the same if played with slightly different attack or at a different sample time, thus resulting in a perceptual difference that should not be attributed to the synthesis model. Similarly, Heise et al. (2009) evaluated synthesis parameter selection using a range of low-level audio features, such as fundamental frequency, spectral shape, envelope characteristics, and overall duration. They also used the discrete cosine transform (DCT) of the Mel-Frequency Cepstral Coefficients (MFCC's) as a measure of how similar the synthesized sound was to a recorded sample.

The manner in which an individual interacts with a synthesis engine is vital to understand, and there is considerable research on this. As part of the Sounding Object project, a large body of work was undertaken in sound effect synthesis, primarily focusing on interactions with sound synthesis models (Rocchesso et al. 2003). Evaluation of the perceived quality of an interaction with a synthesis engine was performed by Böttcher and Serafin (2009) and further developed in Böttcher et al. (2013). However, this type of work is often measuring the parameter mapping more than the quality of the sound synthesis (Heinrichs and McPherson 2014). Hoffman and Cook (2006b) discussed the generalized process of synthesis parameter mapping to perceptual controls through feature vector mapping. There are other discussed methods for mapping physical controls of a synthesis engine to perceptual parameters (Aramaki et al. 2012).

Moffat et al. (2017) used feature vectors to compare the sonic similarity of different sound effects. Scavone et al. (2001) created a program for presenting sound effects on a two-dimensional plane using multi-dimensional scaling (MDS), while Lakatos et al. (1997) asked participants if they could identify the material dimensions of an impact sound in a two-alternative forced choice experiment and then employed MDS on the results. Participants were played samples and had to write free text responses in work by Ma et al. (2010).

In summary, although there is a large body of work on sound synthesis evaluation, many proposed methods have not been evaluated in terms of realism or related attributes. When evaluation has been performed, it is often not subjective, and it is even rarer for it to be comparative, where the proposed technique is compared against alternatives. Nor have standard methodologies been established. This failing of the sound synthesis community to address evaluation is a clear contributing factor to the lack of understanding of the current state-of-the-art in sound synthesis.

Evaluation of existing synthesis methods could potentially yield significant insight into the state-of-the-art in synthesis technology. Without understanding of current synthesis techniques, their benefits, and their weaknesses, it is not possible to understand where the current deficits exist. The lack of standardized evaluation methods and metrics is evident and can potentially prohibit progress in this field.

As is evident from the literature, it is never expected that a single synthesis method is effectively able to produce all possible sounds. In every case, there may be a range of synthesis approaches that are appropriate. However, this simply highlights the importance of evaluation. Identification of suitable use cases and occasions where a particular sound synthesis method is applicable is vital to having a convincing synthesis process.

3 SYNTHESIS METHODS

Six different synthesis methods were used to synthesize a range of different sound effects. The synthesis methods were selected to represent a large range of published work in the field. No completely physical models were used, due to the significant complexity in representing the full system that would represent the complex natures of the composite scenes used.

3.1 Sinusoidal Modelling

Sinusoidal Modeling Synthesis or Spectral Modeling Synthesis (SMS) (Serra and Smith 1990) is considered as a signal-based synthesis method. Sinusoidal modelling assumes that sounds can be synthesized as a summation of

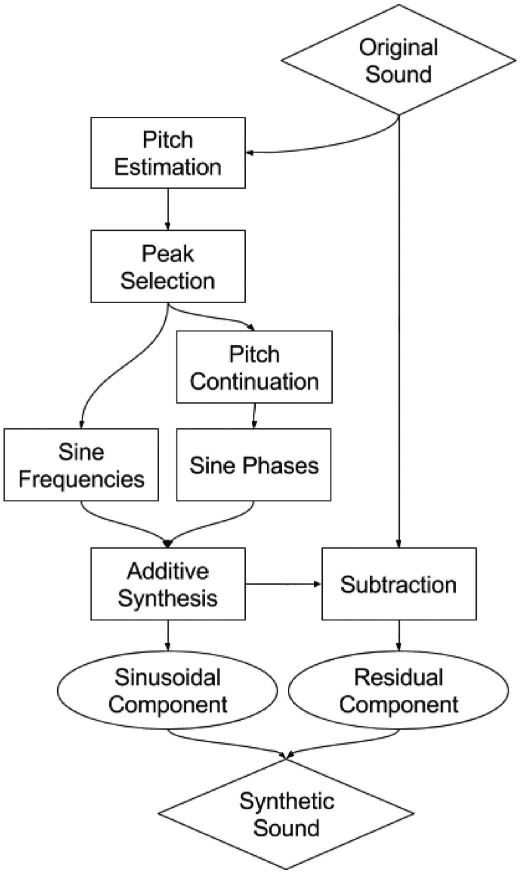


Fig. 1. Flow diagram of sinusoidal modelling, based on Serra and Smith (1990).

sine waves and a filtered noise component such that any sound $x(t)$ can be represented as

$$x(t) = \sum_{r=1}^{R} A_r(t) \sin(\theta_r(t)) + e(t), \tag{1}$$

where $x(t)$ is a summation of R sinusoids, A_r and θ_r are the amplitude and phase, respectively, of a given sinusoid at time t, and $e(t)$ is the noise component, referred to as the residual.

As presented in Figure 1, Sinusoidal modelling is performed by peak selection from the frequency spectra. These peaks are resynthesized using sine waves. The output sine waves are summed together, and the residual is calculated as the result of subtracting the summation of sine waves from the initial sound signal. The synthesis method evaluated was based on the documentation and implementation from Serra and Smith (1990) and Amatriain et al. (2002).

3.2 Additive Synthesis

Traditionally additive synthesis was a form of signal-based modelling where a series of sine waves were added together to produce complex waveform. This technique was further developed and became sinusoidal modelling,

Table 1. The Five Classes of Atoms Used for the Sound Synthesis Models, with Their Respective Synthesis Equations and Parameters, as Taken from Verron et al. (2010)

Atom	Equation	Parameters
Modal impact	$x_1(t) = \sum_{m=1}^{M} a_m \sin(2\pi f_m t) e^{-\alpha_m t}$	a_m initial amplitudes, α_m decays, f_m frequencies
Noisy impact	$x_2(t) = \sum_{n=1}^{N} a_n s_n(t) e^{-\alpha_n t}$	a_n subband amplitudes, α_n subband decays
Chirped impact	$x_3(t) = a \sin\left(2\pi \left(f_0 t + \frac{\sigma}{2} t^2\right)\right) e^{-\alpha t}$	f_0 initial frequency, σ rate of frequency shift, α decay
Band-limited noise	$X_4(f) = \begin{cases} A(t), & \text{if} \lvert f - F(t) \rvert < \frac{B(t)}{2} \\ A(t) e^{-\alpha(t)\left(\lvert f - F(t) \rvert - \frac{B(t)}{2}\right)}, & \text{otherwise} \end{cases}$	$F(t)$ center frequency, $B(t)$ bandwidth, $\alpha(t)$ filter slope, $A(t)$ amplitude
Equalized noise	$x_5(t) = \sum_{n=1}^{32} a_n(t) s_n(t)$	$[a_1(t)...a_{32}(t)]$ amplitudes

$s_n(t)$ represents subband filtered noise, band n at time step t. $x(t)$ represents a time domain signal, whereas $X(f)$ represents a frequency domain signal. $x_1 \ldots x_3$ and x_5 are calculated in the time domain, whereas X_4 is calculated in the frequency domain.

as discussed in Section 3.1. Additive synthesis has since become the process of modelling sounds as a summation of synthesized audio signals, such as noise signals, sinusoids, and chirp sounds.

For the purposes of evaluation, the Spatialized Additive Synthesizer for Environmental Sounds (SPAD) from Verron et al. (2010) was used. SPAD works on the principal of breaking every sound into one of five core sound elements, or atoms, and synthesizing each sound as one of these core elements. Elements are synthesized as per Table 1. All synthesis of atoms occurs in the time domain, apart from band-limited noise, which is synthesized in the frequency domain.

3.3 Physically Inspired Synthesis

Physically Inspired Synthesis (Cook 2007) is derived from physical modelling. It is possible to construct synthesis systems by modelling the entire physical environment in which the sound was created, but this can be incredibly complex to construct. Physically Inspired or Physically Informed Synthesis is considered as another form of signal-based modelling or as a hybrid approach between signal-based modelling and physical modelling, where the user controls represent the physics of the system, but the calculations are all approximations to allow the system to run in real time. Sounds are constructed as a combination of base units, such as filtered noise, sine, triangle, and square waves, envelope shapes, and filters.

Producing a physically inspired synthesis simulation or model of a sonic context is considered a time-consuming process. Each individual sound synthesis model needs to be manually constructed with knowledge

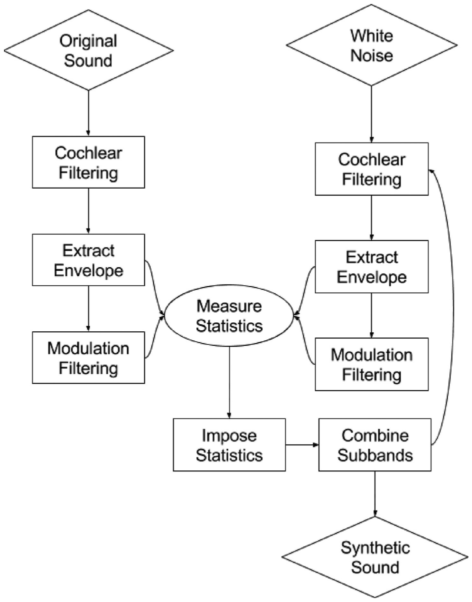


Fig. 2. Flow diagram for Statistical Synthesis, based on McDermott and Simoncelli (2011).

1. Noise signal is filtered into cochlear subbands
2. Calculate Subband envelopes
3. Sub bands of envelopes are calculated, to represent detail
4. Down sample envelopes
5. Calculate envelope statistics
6. Calculate error by comparing statistics to the original signal
7. Envelopes are modified using gradient descent method
8. Envelopes are upsampled
9. Envelopes applied to input noise signal subbands
10. Sub bands combines to create synthesis signal

Fig. 3. The iterative process undertaken to perform Statistical Synthesis.

of the physics, understanding of psychoacoustics, and experience in sound synthesis model production and workflows. Despite the labor-intensive nature, physically inspired synthesis is an effective and flexible method of sound synthesis, as once a context has been modelled, it is possible to vary a large range of parameters to create very different sounding environments, with physically and perceptually relevant interface controls.

For the purposes of evaluation, a number of synthesis models were taken from Farnell (2010) and Peltola et al. (2007).

3.4 Statistical Modelling and Marginal Statistics

Statistical Modelling is a synthesis technique where an input sound file is decomposed into a set of summary statistics. These statistics are used to shape an input noise signal and resynthesize the input audio file. The extracted statistics are based on perceptual models of audio signals. Statistics of the sound are calculated from an auditory inspired cochlear filter bank representation of the signal.

There are two different use cases presented using this algorithm, one is described as Marginal Statistics and the other as Statistical Modelling. They both take the same form but use a different set of statistics to represent the audio file. Marginal statistics are the mean, variance, skew, and kurtosis of the subband envelope and modulation power, extracted from the filtered signal representation. Statistical Modelling includes all the statistics of the marginal statistics and includes the cross-subband envelope correlation and cross-subband modulation correlations. Full mathematical descriptions are presented in McDermott and Simoncelli (2011).

Sounds were resynthesized from the set of chosen statistics, through an iterative process of shaping Gaussian white noise, as can be seen in Figures 2 and 3. For the purposes of evaluation, the synthesis method, documentation, and implementation were taken from McDermott and Simoncelli (2011).

3.5 Concatenative Synthesis

Concatenative Synthesis is a subset of granular synthesis and a form of sample-based synthesis. Segments or "grains" are made from small segments of sound samples. Grains can range from 10ms to 1s samples of audio. Concatenative synthesis is the process of selecting and recombining the grains together, in such a manner that it does not create any perceptual discontinuities.

Due to the lack of available open source implementations of concatenate synthesis, this synthesis method was implemented by the authors, based on O'Leary and Robel (2014).

A library of 46ms audio grains was constructed, selected at 1.5ms intervals from the samples. Grain selection from the library was performed using a time domain probabilistic method. Given the current grain, a subset library of grains was selected based on the Spearman correlation distance of the time domain waveform signal, such that

$$d_t = 1 - \frac{(v_r - \bar{v}_r)(v_t - \bar{v}_t)'}{\sqrt{(v_r - \bar{v}_r)(v_r - \bar{v}_r)'}\sqrt{(v_t - \bar{v}_t)(v_t - \bar{v}_t)'}}, \tag{2}$$

where v is a coordinate-wise vector of either the current grain r or query grain t, for which the distance is calculated. $\bar{v}$ denotes the mean of the vector to normalize the vector around its current mean. The Spearman distance was used, as it considers the sample vector in sequence, so small variations in sample do not result in a significant overall difference. The time domain vector to represent each grain is taken as the second half (23ms) of the current grain, and the first half of the grain within the grain library. The time domain waveform vectors of the current grain and all possible grains were selected.

From this calculated subset library of possible grains, one grain g_t is selected with probability

$$P(g_t) = \frac{1 - d_t}{\sum_{k=1}^{K} d_k}, \tag{3}$$

where d_t is the Spearman distance from the current input grain and K is the number of selected nearest neighbors, in this case 10.

The selected grain is then overlapped with the current audio grain, and the two audio samples crossfaded. The implementation is available to download.[1]

4 EXPERIMENTAL METHOD

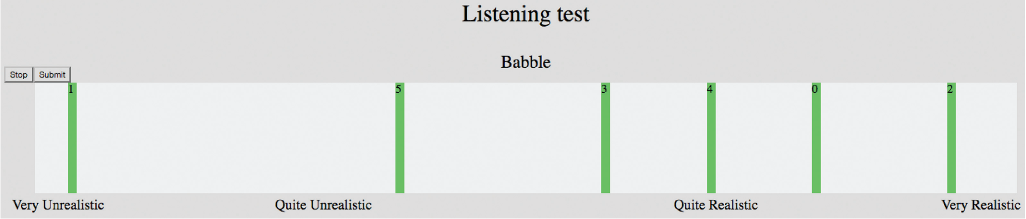
4.1 Participants

Eighteen participants between the ages of 18 and 40 took part in the experiment, of which 11 were male and 7 female. The procedure was approved by the local ethics committee. The average test duration was 17.5 min, so fatigue was not an issue.

4.2 Experimental Setup

The experiment took place in a dedicated, professionally acoustically treated listening room at Queen Mary University of London. The audio was played back over a pair of PMC AML2 loudspeakers, where the participant could adjust the volume of the audio to a comfortable level. Participants were asked to set the volume during the first test and then refrain from adjusting it during the remainder of the test. No participant moved the volume more than 3dB from its starting position, so this effect is considered negligible. The listening test was set up using the Web Audio Evaluation Tool (Jillings et al. 2015). The test was browser based so that no proprietary software had to be installed on the computer (Jillings et al. 2016). A screenshot of the user interface used for

[1] https://goo.gl/rDtIk3.


Listening test

Babble

Fig. 4. A screenshot of the user interface used by participants for inter-comparison of sound samples.

this experiment is presented in Figure 4. An online version of the listening test is available[2] with the same user interface and set of samples that were used by participants.

4.3 Materials

Participants were asked to evaluate sound textures for eight categories (applause, babble, bees, fire, rain, stream, waves, wind). These textures comprise a large range of sounds that have been used for sound synthesis evaluation in existing work (McDermott and Simoncelli 2011; Schwarz et al. 2016). They represent composite scenes containing a range of different timbres of sounds. But the long-term evolution and structure of the sound are as important contributing factors as the timbre of each individual sonic element within the complex scenes. Thus any synthesis method should model the temporal development of the sound along with the instantaneous qualities. In particular, the applause and babble sounds were selected as they are known to be challenging sounds to reproduce and may test sound synthesis methods to the limit of their capabilities.

In every category between 6 and 11 samples were provided. Sixty-six samples were evaluated in total. All samples were 44.1kHz wav files, and loudness normalized in accordance with ITU-R BS.1387-1 (1998). Each category had at least one anchor and at least one recorded sample. The recorded samples were all selected by a group of five experienced critical listeners as being realistic samples, given at least five different sample options. Each anchor was constructed from a trivial additive synthesis model, produced by deconstructing either the additive or physically inspired model to the point that it was barely perceivable as the intended sound.

The references and anchors were important within this test to encourage participants to use the entire evaluation scale, and we could review how samples were distributed within that scale, in accordance with ITU-R BS.1534-3 (2015). The reference samples allowed us to evaluate how our synthesis method compared to the genuine sound and to allow us to identify whether the samples are distinguishable from the real sample. The purpose of the anchor sample was to support evaluation of how synthesis methods compared to each other. If every synthesis method was highly realistic, then the participants may decide to use the entire evaluation scale to identify micro-differences between samples or may decide to group all samples together at the high end of the scale. The anchor ensures that there is a lower limit sample to compare against. It also performs as a confirmation that a participant has fully understood the requirements for the experiment. If a participant rated the anchor as higher than the sample, then we would infer that the participant may not have fully understood the requirements or may have some hearing defect.

A list of synthesis methods used within each sound class is presented in Table 2. To demonstrate the full range of reference sound samples, audio features were extracted from the samples (Bogdanov et al. 2013) based on recommendations (Moffat et al. 2015), and summarized attributes are presented in Table 3. All sound samples used, software implementations, and parameter settings are available online.[3]

[2] https://goo.gl/789eWl.
[3] https://code.soundsoftware.ac.uk/projects/perceptual-evaluation-of-sound-synthesis.

Table 2. Synthesis Method Used to Created Each Sound Sample

Synthesis Method	Applause	Babble	Bees	Fire	Rain	Stream	Waves	Wind
Physically Inspired	N	N	Y	Y	Y	Y	N	Y
Marginal Statistics	Y	Y	N	Y	Y	Y	N	Y
Sinusoidal Modelling	Y	Y	Y	Y	Y	Y	N	N
Additive	N	N	N	Y	Y	N	Y	Y
Statistical Modelling	Y	Y	Y	Y	Y	Y	Y	Y
Concatenative	Y	Y	Y	Y	Y	Y	Y	Y

Table 3. Summary of Attributes of Different Sounds Classes Used for Evaluation

	Environmental	Animal/Human	Synchronized	Noisy	Harmonic	Granular
Applause	N	Y	Y	Y	N	Y
Babble	N	Y	N	N	Y	Y
Bees	N	Y	N	N	Y	N
Fire	Y	N	N	Y	Y	Y
Rain	Y	N	N	Y	Y	Y
Stream	Y	N	Y	N	Y	Y
Waves	Y	N	Y	Y	N	N
Wind	Y	N	Y	Y	N	N

4.4 Procedure

Participants were provided with instructions as to the experiment they were to undertake and were asked to provide their native spoken language, whether they had previous experience of listening tests and whether they would consider themselves as accomplished musicians or audio engineers.

Participants were then asked to rate how realistic they perceived the samples within a given category, relative to all the other samples within that category. Participants were provided with a continuous linear scale on which to rate all sounds, labeled from “very unrealistic” to “very realistic.” All sounds were rated on a single horizontal scale to encourage inter-sample comparison. Participants were provided with the sound category name but, other than that, did not have any information regarding the samples. Both the ordering of categories and the initial ordering of samples within a category were randomized.

5 RESULTS

The overall results for the experiment are presented in Figure 5 using a notched box plot. In all plots the red line represents the median. The end of the notches, where the angled lines become parallel within the box plot, represents the 95% confidence intervals, and the end of the boxes represent the first and third quartiles. The end of the whiskers represent the data range not considered as an outlier. Red crosses are outliers. The anchor and reference have very low and very high medians, respectively, with small confidence ranges. This informs us that the anchor and reference function as intended.

The null hypothesis is that the perceptual evaluation scores are from the same distribution. A one-way ANOVA, with Bonferroni correction, shows that for all sound classes, the effect of each synthesis method on user perception was statistically significant F(7,946) = 176.51, p < 0.0001. Table 5 shows the statistical significance of the difference in ratings between synthesis methods for all sound samples. A post-hoc Tukey pairwise comparison, with Bonferroni correction to reduce the chance of type I errors, was used. It can be seen, for example, that concatenative synthesis is significantly different from the reference sample, marginal statistics, additive synthesis, and statistical modelling all with a $p < 0.0001$. However, concatenative synthesis

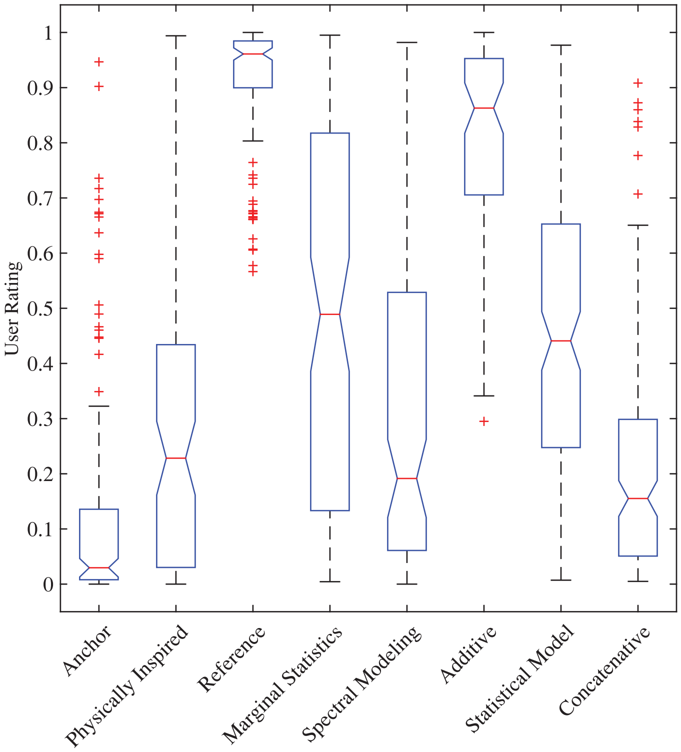


Fig. 5. Plot of the median, standard deviation and 95% confidence intervals of all synthesis results.

Table 4. Mean and Standard Deviation of Each Sound Class

Sound Class	Mean	Standard Deviation
Applause	0.40	0.37
Babble	0.35	0.33
Bees	0.44	0.35
Fire	0.38	0.32
Rain	0.40	0.37
Stream	0.39	0.36
Waves	0.49	0.37
Wind	0.57	0.32

is not significantly different from the anchor, physically inspired synthesis or the sinusoidal modelling. These results are then presented in more detail, broken down by sound class in Table 6.

The additive method performed best overall, and was the only synthesis method where the results were not significantly different from the reference. It was also significantly different from all other synthesis methods. However, this method was not used in all tests, as only a subset of sounds (fire, rain, wind, and waves) could be synthesized using additive synthesis. Table 4 shows the mean and standard deviation of each sound class. With the exception of wind, there is little variation between the means of each sound class. This suggests that the

Table 5. Results of Pairwise Comparison of Synthesis Method on Perceptual Realism Rating, with Bonferroni Correction, o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no Comparison Made

		Physically		Marginal	Sinusoidal		Statistical	
	Anchor	Inspired	Reference	Statistics	Modelling	Additive	Modelling	Concatenative
Anchor	.	**	****	****	***	****	****	o
Physically Inspired	**	.	****	****	o	****	***	o
Reference	****	****	.	****	****	o	****	****
Marginal Statistics	****	****	****	.	****	****	o	****
Sinusoidal Modelling	***	o	****	****	.	****	***	o
Additive	****	****	o	****	****	.	****	****
Statistical Modelling	****	***	****	o	***	****	.	****
Concatenative	o	o	****	****	o	****	****	.

Table 6. Results of Pairwise Comparison of Synthesis Method on Perceptual Realism Rating for Each Class of Sound, with Bonferroni Correction, o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no Comparison Made

Group 1	Group 2	Applause	Babble	Bees	Fire	Rain	Stream	Waves	Wind
Anchor	Reference	****	****	****	****	****	****	****	o
Anchor	Physically Inspired	.	.	o	*	o	o	.	o
Anchor	Marginal Statistics	o	o	.	*	****	****	.	o
Anchor	Sinusoidal Modelling	****	****	****	o	o	o	.	.
Anchor	Additive	.	.	.	****	****	.	****	**
Anchor	Statistical Modelling	****	***	****	o	****	****	*	***
Anchor	Concatenative	o	o	****	o	o	o	o	****
Reference	Physically Inspired	.	.	****	****	****	****	.	***
Reference	Marginal Statistics	****	****	.	****	****	o	.	o
Reference	Sinusoidal Modelling	****	****	*	****	****	****	.	.
Reference	Additive	.	.	.	o	o	.	**	o
Reference	Statistical Modelling	****	****	****	****	***	***	****	****
Reference	Concatenative	****	****	****	****	****	****	****	****
Physically Inspired	Marginal Statistics	.	.	.	o	****	****	.	*
Physically Inspired	Sinusoidal Modelling	.	.	****	o	o	o	.	.
Physically Inspired	Additive	.	.	.	****	****	.	.	****
Physically Inspired	Statistical Modelling	.	.	****	o	****	****	.	***
Physically Inspired	Concatenative	.	.	****	o	o	o	.	****
Marginal Statistics	Sinusoidal Modelling	***	o	.	o	****	****	.	.
Marginal Statistics	Additive	.	.	.	****	o	.	.	o
Marginal Statistics	Statistical Modelling	****	o	.	o	o	o	.	****
Marginal Statistics	Concatenative	o	o	.	o	****	****	.	****
Sinusoidal Modelling	Additive	.	.	.	****	****	****	.	.
Sinusoidal Modelling	Statistical Modelling	o	o	o	o	****	.	.	.
Sinusoidal Modelling	Concatenative	o	o	o	o	o	o	.	.
Additive	Statistical Modelling	.	.	.	****	o	.	****	****
Additive	Concatenative	.	.	.	****	****	.	****	****
Statistical Modelling	Concatenative	***	o	o	o	****	****	o	o

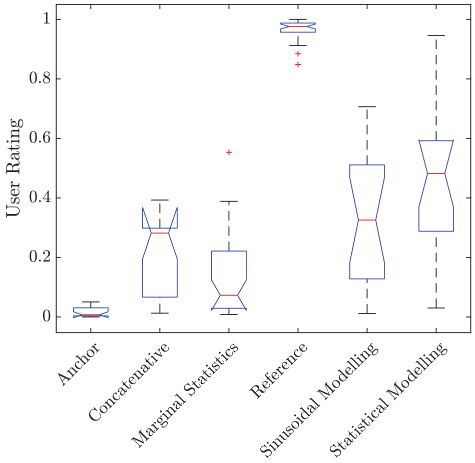


Fig. 6. Applause result distribution.

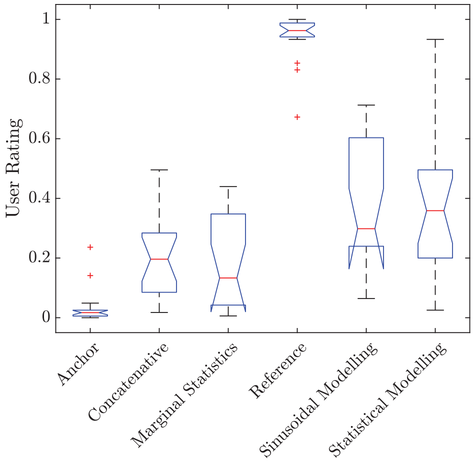


Fig. 7. Babble result distribution.

superior performance of additive synthesis was not due to higher ratings for these sound classes, but, instead, the synthesis method itself must have performed well.

Concatenative synthesis is the only method not significantly different from the provided anchor sounds. Table 5 shows that the different synthesis techniques can be broken down into three perceptual groupings, where Sinusoidal Modelling, Physically Inspired, and Concatenative are all grouped together with the Anchor.

Statistical Modelling and Marginal Statistics can also be grouped together. This is to be expected, as they are based on the same implementation with different sets of synthesis statistics.

5.1 Results Per Sound Class

Analysis of variance (ANOVA) with Bonferroni correction showed that for a given sound class the effect of each synthesis method on user perception was significant and in all cases $p < 0.0001$. A post-hoc Tukey pairwise comparisons shows the statistical significance of the differences between each synthesis methods, given each sound class, seen in Table 6. For all sound classes the anchor, the physically inspired model, the sinusoidal modelling, the statistical modelling, and the concatenative synthesis all had perceptual rating distributions that were significantly different from the reference. Marginal statistics and additive were the only two synthesis methods under which there is sometimes no clear difference between their perceptual rating distributions, given a specific sound class.

The median, standard deviation, and 95% confidence intervals for each synthesis method, for each sound class, are reported in Figures 6–13. For all sounds except wind and stream sounds, the anchor had the lowest median rating, with small confidence intervals. For wind sounds, though sinusoidal modelling had a lower median rating, there is statistically no discernible difference in their distributions, and, as such, it can be said they are equally poor. Wind is the only case where the anchor is not one of the worst samples selected. This suggests that the anchor may not have been ideal. But the concatenative synthesis method produced a very low perceptual rating with small confidence intervals, so the concatenative synthesis method can be considered as the anchor in this case. This is confirmed by the fact there is no significant difference between the anchor and the reference for wind.

In the case of synthesizing wind, additive performed better than the reference sound. This is the only case where a synthesis method outperformed the reference recorded signal. The difference in distributions between additive and the reference is not significant. The null hypothesis was not rejected, and thus additive might be considered as realistic as a recording of wind and possibly more realistic. In the case of fire and rain synthesis,

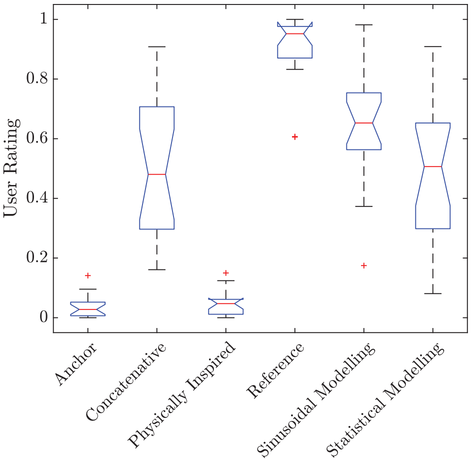


Fig. 8. Bees result distribution.

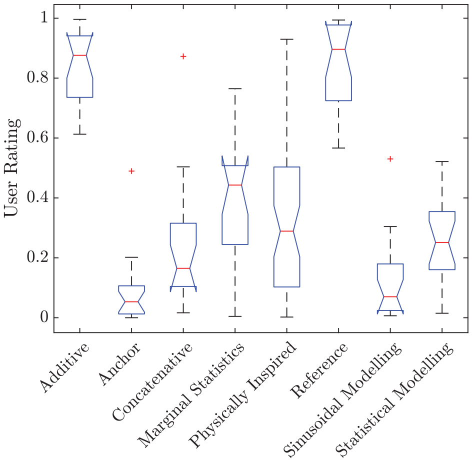


Fig. 9. Fire result distribution.

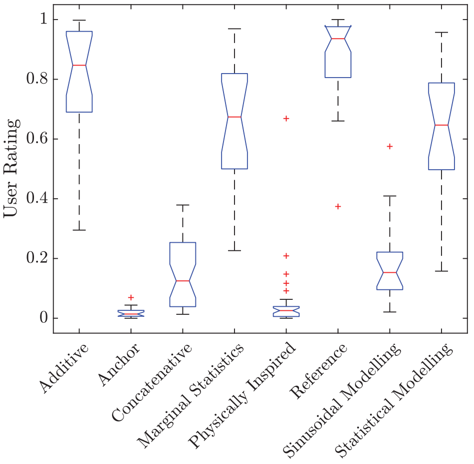


Fig. 10. Rain result distribution.

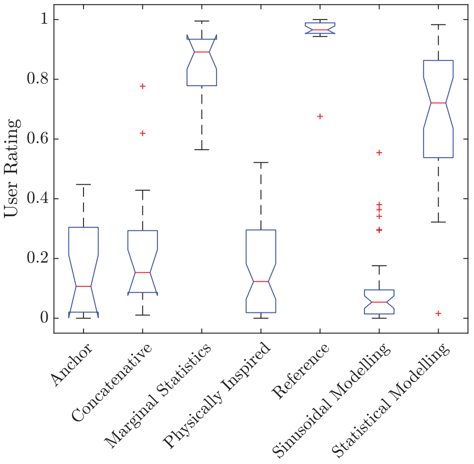


Fig. 11. Stream result distribution.

additive could also be considered as realistic as a recorded reference sample, since the null hypothesis could not be rejected, and the confidence intervals are significantly overlapping.

We present a summary of the results in Table 7, where we summarize the effectiveness of each synthesis method at producing the relevant sounds.

6 DISCUSSION

The results suggest that additive synthesis is an effective approach for environmental sounds such as fire, water, and wind sounds. These sounds can be considered as sounds constructed from band-pass filtered noise. Marginal Statistics are effective for synthesizing wind- and stream-type sounds. For applause and babble sounds, which are are more dynamic and impulsive, the statistical modelling synthesis proved to be the most effective approach in synthesizing these types of sounds. As can be seen in Table 6, wind and stream sound synthesis can effectively be produced with marginal statistical synthesis, in such a manner that the realism rating distribution is not

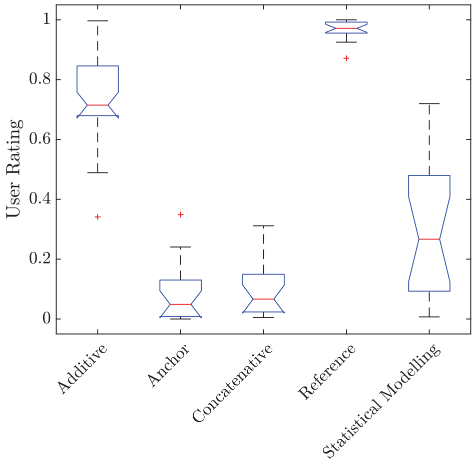


Fig. 12. Waves result distribution.

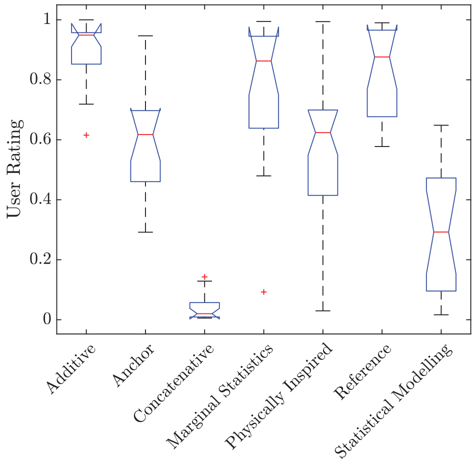


Fig. 13. Wind result distribution.

Table 7. Rating of Synthesis Method per Sound Class. 1 = Best Method, Comparable with Reference, 5 = Worst Method, Comparable with Anchor, . = No Comparison Made

Synthesis Method	Applause	Babble	Bees	Fire	Rain	Stream	Waves	Wind
Physically Inspired	.	.	5	3	5	4	.	3
Marginal Statistics	4	5	.	3	3	1	.	1
Sinusoidal Modelling	3	3	3	5	4	5	.	.
Additive	.	.	.	1	1	.	2	1
Statistical Modelling	3	3	4	4	3	2	3	4
Concatenative	3	3	4	4	4	4	4	5

significantly different from that of the reference sounds. Despite this, it is noted that marginal statistics are “sufficient to produce compelling synthetic examples of many water textures (rain, streams etc.), but not much else” (McDermott et al. 2009). McDermott et al. (2009) suggests this applies to all sounds that are based around filtered noise signals, where sounds are primarily made up of noisy audio signals with little harmonic component. As such, water and wind sounds are all effectively synthesized using the Marginal Statistical method for synthesis.

In the case of the wind sounds, the additive synthesis method performed better than the reference sample. However, the difference was not considered to be significant, so this may be a statistical abnormality. But it may also be an indication of hyper-realism. The idea of hyper-realism is simply that an unreal sound can sound “more real” than a real sound. This is particularly prominent in weapon and explosion sounds (Mengual et al. 2016), where a listener may never have heard a real gunshot sound but will have a strong opinion of a gun sound based on TV, film, and video games (Puronas 2014).

Concatenative synthesis created some noticeable artifacts in some of the samples. The artifacts seem to be caused by non-smooth transitions between frames but are only perceivable in a small number of sound contexts. This caused this synthesis method to underperform in certain cases, particularly for rain and fire sounds and, to a lesser extent, for babble. These are impulsive sounds, where the individual sonic elements may be smaller than an individual grain of sound, with variable size.

The sinusoidal modelling method also caused some audible artifacts, particularly in the fire and babble sounds. It is suspected that this was caused by spectral peaks being modelled as harmonic components, when they are actually noisy spectral peaks. There also appears to have been issues with phase recognition, which again is due to noisy signal components being modelled as harmonic components resulting in an audible vocoder-like effect.

Many of the physically inspired models were taken from Farnell (2010), which is designed as a textbook for teaching the principles of procedural audio. Thus, the focus of these designed sounds are to relate the sonic interaction rather than the exact replication of realistic sounds. These sound synthesis models did not produce convincing sounds. Despite this, we felt it important to evaluate this range of algorithms, as they are popular, well-known synthesis methods.

The results show that additive synthesis is an effective synthesis method for both slow moving and impulsive sounds. But additive synthesis allows for a very large range of possible parameters, and the individual parameter ranges were slowly selected and hand crafted by the original authors. Thus this sound synthesis method cannot easily be generalized to a large number of sounds.

Particularly for slow changing sounds, statistical synthesis is effective, either using a reduced feature set or the full feature set. It is speculated that a granular synthesis method may be most effective for impulse sounds, due to the fact that these sound textures are generally made up of a large number of small sound atoms, e.g., individual plosives in babble, claps within applause, or raindrops in rain.

No synthesis technique was capable of producing convincing applause or babble sound. This was expected, as these synthesis methods are known for being challenging sounds to synthesize. However, the sinusoidal modelling and statistical modelling performed well on these sounds. This suggests that noise components are important in the reproduction of a realistic applause or babble sounds, since statistical modelling and sinusoidal modelling involve careful noise shaping. Additive synthesis produced realistic sounding examples of fire and rain sounds. This may be because the method focuses on synthesizing individual sonic elements separately and then constructing a scene from these elements, rather than alternative methods, such as statistical modelling, which models the statistics of the entire sounds. In particularly composite scenes, such as fire and rain, the individual sonic element synthesis is more important than overall sonic structure.

7 CONCLUSIONS

We described an experiment in which participants were asked to rate 66 examples of synthesized sounds from eight different sound classes and five different synthesis methods in terms of their perceived "realism." The results demonstrate that sound synthesis methods can be as convincing as a recorded audio sample. In the case of wind, the users consistently rated the sound as more realistic than the recorded sample. In five of the eight sound classes tested, there exist synthesis techniques where synthesized sounds were indistinguishable, in terms of realism, from recorded samples.

This experiment presents a method for evaluation of synthesized sounds in a range of different sound classes and provides recommendations for synthesizing different types of sounds. It is clear that although sound synthesis can effectively synthesize a range of realistic sounds, there are many potential future directions for development of plausible sound synthesis across the full sonic range.

Despite this, there are limitations of the work presented. We evaluated a relatively small number of sound synthesis methods, and as such we cannot make any substantial claims about entire areas of synthesis research. There is a requirement for further work in comparing and evaluating more synthesis techniques.

This article identifies the need for evaluation and further development of sound synthesis. Evaluation of sound synthesis can assist in improving on the state of the art and developing future sound synthesis. This article described a clear and rigorous method for evaluation of sound synthesis, through a double blind multiple comparison evaluation test. This test methodology can be used to evaluate any sounds synthesis method to determine the perceived realism of the synthesized sound, given a single word or phrase context.

7.1 Future Work

There are clear opportunities to develop a better understanding of the current state of the art within sound synthesis. Globally comparing sound synthesis methods and looking within sound groupings can both yield meaningful results. Identification of sonic groupings would be beneficial and would encourage bespoke grouping-based synthesis research rather than global synthesis approaches. This could be developed further to a structured taxonomy or ontology of sounds based around the auditory perceptual system. Further to this, the development of an objective measure that could be used to evaluate synthesis, without the requirement for listening tests, and participants would be significantly beneficial to the synthesis community. This could be adopted by all researchers within the field and would certainly improve the standard and consistency of evaluation within the sound synthesis community.

A single process for evaluating synthesis would never be able to encapsulate everything that is required to evaluate such a multidimensional problem as sound synthesis. It is also the case that measuring the effectiveness of a synthesis method designed to synthesize a real-world sound is one of a range of important evaluation metrics, and one that is not often investigated in the literature. This type of evaluation does not negate the need for other evaluation forms but merely adds to the understanding of the utility of existing work.

REFERENCES

Xavier Amatriain, Jordi Bonada, Alex Loscos, and Xavier Serra. 2002. Spectral processing. In *DAFx: Digital Audio Effects*, Udo Zölzer (Ed.). John Wiley and Sons, Ltd., Chichester, UK, Chapter 10, 373–438.

Mitsuko Aramaki, Richard Kronland-Martinet, and Sølvi Ystad. 2012. Perceptual control of environmental sound synthesis. In *Speech, Sound and Music Processing: Embracing Research in India*. Springer, Berlin, 172–186.

James A. Ballas. 1993. Common factors in the identification of an assortment of brief everyday sounds. *J. Exp. Psychol. Hum. Percept. Perf.* 19, 2 (1993), 250.

Stefan Bilbao. 2009. *Numerical Sound Synthesis: Finite Difference Schemes and Simulations in Musical Acoustics*. Wiley Online Library.

Stefan Bilbao and John Chick. 2013. Finite difference time domain simulation for the brass instrument bore. *J. Acoust. Soc. Am.* 134, 5 (2013), 3860–3871.

Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R. Zapata, and Xavier Serra. 2013. Essentia: An audio analysis library for music information retrieval. In *Proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR'13)*. 493–498.

Terri L. Bonebright, Nadine E. Miner, Timothy E. Goldsmith, and Thomas P. Caudell. 2005. Data collection and analysis techniques for evaluating the perceptual qualities of auditory stimuli. *ACM Trans. Appl. Percept.* 2, 4 (2005), 505–516.

Niels Böttcher, Héctor P. Martínez, and Stefania Serafin. 2013. Procedural audio in computer games using motion controllers: An evaluation on the effect and perception. *International Journal of Computer Games Technology* 2013 (2013), Article ID 371374, 16 pages. DOI: 10.1155/2013/371374

Niels Böttcher and Stefania Serafin. 2009. Design and evaluation of physically inspired models of sound effects in computer games. In *Proceedings of the 35th International Conference of the Audio Engineering Society Conference: Audio for Games*. AES, London.

B. Caramiaux, F. Bevilacqua, T. Bianco, N. Schnell, O. Houix, and P. Susini. 2014. The role of sound source perception in gestural sound description. *ACM Trans. Appl. Percept.* 11, 1 (Apr. 2014), 1:1–1:19.

Perry R. Cook. 2007. Real sound synthesis for interactive applications.

Andy Farnell. 2010. *Designing Sound*. MIT Press Cambridge, UK.

Martin Fröjd and Andrew Horner. 2009. Sound texture synthesis using an overlap–add/granular synthesis approach. *J. Audio Eng. Soc.* 57, 1/2 (2009), 29–37.

Leonardo Gabrielli, Stefano Squartini, and Vesa Välimäki. 2011. A subjective validation method for musical instrument emulation. In *Proceedings of the 131st Audio Engineering Society Convention*.

Henrik Hahn. 2015. *Expressive Sampling Synthesis-Learning Extended Source–Filter Models from Instrument Sound Databases for Expressive Sample Manipulations*. Ph.D. Dissertation. UPMC Université Paris VI.

Brahim Hamadicharef and Emmanuel Ifeachor. 2003. Objective prediction of sound synthesis quality. In *Proceedings of the 115th Audio Engineering Society Convention*.

Brahim Hamadicharef and Emmanuel Ifeachor. 2005. Perceptual modeling of piano tones. In *Proceedings of the Audio Engineering Society Convention 119*.

Christian Heinrichs and Andrew McPherson. 2014. Mapping and interaction strategies for performing environmental sound. In *Proceedings of the 1st Workshop on Sonic Interactions for Virtual Environments at IEEE VR 2014*.

Sebastian Heise, Michael Hlatky, and Jörn Loviscach. 2009. Automatic cloning of recorded sounds by software synthesizers. In *Proceedings of the Audio Engineering Society Convention 127*. AES, New York, NY.

Simon Hendry and Joshua D. Reiss. 2010. Physical modeling and synthesis of motor noise for replication of a sound effects library. In *Proceedings of the Audio Engineering Society Convention 129*.

Matthew D. Hoffman and Perry R. Cook. 2006a. Feature-based synthesis: A tool for evaluating, designing, and interacting with music IR systems. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR'06)*. 361–362.

Matthew D. Hoffman and Perry R. Cook. 2006b. Feature-based synthesis: Mapping acoustic and perceptual features onto synthesis parameters. In *Proceedings of the International Computer Music Conference (ICMC'06)*.

Andrew Horner and Simon Wun. 2006. Evaluation of iterative matching for scalable wavetable synthesis. In *Proceedings of the 29th International Conference of the Audio Engineering Society : Audio for Mobile and Handheld Devices*.

ITU-R BS.1387-1. 1998. *BS. 1387, Method for Objective Measurements of Perceived Audio Quality*. Technical Report. ITU-R.

ITU-R BS.1534-3. 2015. *BS. 1534, Method for Subjective Assessment of Intermediate Quality Level of Audio Systems*. Technical Report. ITU-R.

David A. Jaffe. 1995. Ten criteria for evaluating synthesis techniques. *Comput. Music J.* 19, 1 (1995), 76–87.

Hanna Järveläinen, Tony Verma, and Vesa Välimäki. 2002. Perception and adjustment of pitch in inharmonic string instrument tones. *J. New Music Res.* 31, 4 (2002), 311–319.

Nicholas Jillings, Brecht De Man, David Moffat, and Joshua D. Reiss. 2015. Web audio evaluation tool: A browser-based listening test environment. In *Proceedings of the Conference on Sound and Music Computing 2015*.

Nicholas Jillings, Brecht De Man, David Moffat, and Joshua D. Reiss. 2016. Web audio evaluation tool: A framework for subjective assessment of audio. In *Proceedings of the 2nd Web Audio Conference*.

Stephen Lakatos, Stephen McAdams, and René Caussé. 1997. The representation of auditory source characteristics: Simple geometric form. *Attention Percept. Psychophys.* 59, 8 (1997), 1180–1190.

Xiaojuan Ma, Christiane Fellbaum, and Perry R. Cook. 2010. SoundNet: Investigating a language composed of environmental sounds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1945–1954.

Josh H. McDermott, Andrew J. Oxenham, and Eero P. Simoncelli. 2009. Sound texture synthesis via filter statistics. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009 (WASPAA'09)*. New Paltz, NY, 297–300.

Josh H. McDermott and Eero P. Simoncelli. 2011. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron* 71, 5 (2011), 926–940.

Lucas Mengual, David Moffat, and Joshua D. Reiss. 2016. Modal synthesis of weapon sounds. In *Proceedings of the 61st International Conference of the Audio Engineering Society: Audio for Games*. Audio Engineering Society, London.

Adrien Merer, Mitsuko Aramaki, Sølvi Ystad, and Richard Kronland-Martinet. 2013. Perceptual characterization of motion evoked by sounds for synthesis control purposes. *ACM Trans. Appl. Percept.* 10, 1 (Mar. 2013), 1–24.

Adrien Merer, Sølvi Ystad, Richard Kronland-Martinet, and Mitsuko Aramaki. 2011. Abstract sounds and their applications in audio and perception research. *International Symposium on Computer Music Modeling and Retrieval CMMR 2010: Exploring Music Contents* (2011), 176–187.

Nadine E. Miner and Thomas P. Caudell. 2005. Using wavelets to synthesize stochastic-based sounds for immersive virtual environments. *ACM Trans. Appl. Percept.* 2, 4 (Oct. 2005), 521–528.

A. Misra and P. R. Cook. 2009. Toward synthesized environments: A survey of analysis and synthesis methods for sound designers and composers. In *Proceedings of the International Computer Music Conference (ICMC'09)*.

David Moffat, David Ronan, and Joshusa D. Reiss. 2015. An evaluation of audio feature extraction toolboxes. In *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx'15)*.

David Moffat, David Ronan, and Joshusa D. Reiss. 2017. Unsupervised taxonomy of sound effects. In *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx'17)*.

Emma Murphy, Mathieu Lagrange, Gary Scavone, Philippe Depalle, and Catherine Guastavino. 2008. Perceptual evaluation of a real-time synthesis technique for rolling sounds. In *Proceedings of the Conference on Enactive Interfaces*. Interactive Design Foundation, Pisa, Italy.

Rolf Nordahl, Stefania Serafin, and Luca Turchet. 2010. Sound synthesis and evaluation of interactive footsteps for virtual reality applications. In *Proceedings of the IEEE Virtual Reality Conference*. IEEE, 147–153.

Sean O'Leary and Axel Robel. 2014. A montage approach to sound texture synthesis. In *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO'14)*. IEEE, 939–943.

Juan Pampin. 2004. ATS: A system for sound analysis transformation and synthesis based on a sinusoidal plus critical-band noise model and psychoacoustics. In *Proceedings of the International Computer Music Conference*, Vol. 1001. 402–405.

Leevi Peltola, Cumhur Erkut, P. R. Cook, and Vesa Valimaki. 2007. Synthesis of hand clapping sounds. *IEEE Trans. Audio Speech Lang. Process.* 15, 3 (2007), 1021–1029.

Vytis Puronas. 2014. Sonic hyperrealism: Illusions of a non-existent aural reality. *New Soundtr.* 4, 2 (2014), 181–194.

Davide Rocchesso, Roberto Bresin, and Mikael Fernstrom. 2003. Sounding objects. *IEEE MultiMedia* 10, 2 (2003), 42–52.

Davide Rocchesso and Federico Fontana. 2003. *The Sounding Object*. Mondo estremo.

G. Scavone, Stephen Lakatos, P. Cook, and Colin Harbke. 2001. Perceptual spaces for sound effects obtained with an interactive similarity rating program. In *Proceedings of International Symposium on Musical Acoustics*.

Diemo Schwarz. 2011. State of the art in sound texture synthesis. In *Proceedings of the 14th International Conference Digital Audio Effects (DAFx'11)*. 221–231.

Diemo Schwarz, Axel Roebel, Hengchin Yeh, and Amaury La Burthe. 2016. Concatenative sound texture synthesis methods and evaluation. In *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx'16)*.

Rod Selfridge, David Moffat, Eldad J. Avital, and Joshua D. Reiss. 2017d. Creating real-time aeroacoustic sound effects using physically derived models. (Unpublished).

Rod Selfridge, David Moffat, and Joshua D. Reiss. 2017a. Physically derived sound synthesis model of a propeller. In *Proceedings of the 12th International Audio Mostly Conference*. ACM.

Rod Selfridge, David Moffat, and Joshua D. Reiss. 2017b. Real-time physical model for synthesis of sword swing sounds. In *Proceedings of the International Conference on Sound and Music Computing (SMC'17)*. Espoo, Finland.

Rod Selfridge, David Moffat, and Joshua D. Reiss. 2017c. Sound synthesis of objects swinging through air using physical models. *Applied Sciences*.

Rod Selfridge, David Moffat, Joshua D. Reiss, and Eldad J. Avital. 2017e. Real-time physical model for an aeolian harp. In *Proceedings of the International Congress on Sound and Vibration*. London, UK.

Xavier Serra and Julius Smith. 1990. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Comput. Music J.* 14, 4 (1990), 12–24.

Thilo Thiede, William C. Treurniet, and others. 2000. PEAQ-The ITU standard for objective measurement of perceived audio quality. *J. Audio Eng. Soc.* 48, 1/2 (2000), 3–29.

Tero Tolonen, Vesa Välimäki, and Matti Karjalainen. 1998. *Evaluation of Modern Sound Synthesis Methods*. Technical Report. Helsinki University of Technology.

Charles Verron, Mitsuko Aramaki, and others. 2010. A 3D immersive synthesizer for environmental sounds. *IEEE Trans. Audio. Speech Lang. Process.* 18, 6 (2010), 1550–1561.

Received July 2017; revised October 2017; accepted November 2017