OPEN ACCESS Freely available online

J. R. R. Lee and J. D. Reiss "Real-Time Sound Synthesis of Audience Applause"
J. Audio Eng. Soc., vol. 68, no. 4, pp. 261–272, (2020 April.).
DOI: https://doi.org/10.17743/jaes.2020.0006

Real-Time Sound Synthesis of Audience Applause

JAKE RYAN RAJJAYABUN LEE AND JOSHUA D. REISS, *AES Fellow*
(joshua.reiss@qmul.ac.uk)

Queen Mary University of London, London, UK

We investigate a procedural model for synthesizing applause sounds that contains novel aspects to ensure high quality and usability. Synthesis of a single clap is generated as a result of filtering a noise source and applying an envelope with exponential decay, based on prior art and existing experimental data. An ensemble approach is introduced to simulate many clappers in a spatially distributed environment. This renders how applause interacts with the space in which it is hosted, including the room impulse response, and where each clap is situated relative to the listener's position. The applause features realistic build-up and fade-out based on natural audience response. The implementation contains meaningful parameters that allow a user to configure and change the sound to achieve a multitude of different types of applause, such as an "enthusiasm parameter" to simulate the greater perceived intensity from an enthusiastic audience. Subjective evaluation was performed to compare our method against recorded samples and four other popular sound synthesis techniques. It showed that the proposed implementation produced significantly more realistic results than other forms of applause synthesis, and it was almost indistinguishable from real-life recordings.

1 INTRODUCTION

The gaming, film, and virtual reality industries rely heavily on sample-based audio for the sound design. While triggering event-based samples leads to easy implementation, there are natural limitations. The sound is fixed from the point of recording, leading to drawbacks such as repetition, storage, and lack of perceptually relevant controls.

Procedural audio is nonlinear sound created in real time according to a set of programmatic rules and live input [1]. It offers a more flexible approach by allowing the parameters of a sound to be altered and sound to be generated from first principles. With procedural audio, only a model of the sound needs to be included, much smaller than an audio sample, which responds to live input data, dynamically changing how the sound is created. This reduces repetition and memory requirements, enables sound designers to achieve very specific sounds, and allows these sounds to interact with the physics of the environment [2]. However, procedural audio tends to be computationally expensive compared to sample-based techniques as the sound is continuously generated and some sounds are inherently complex.

At the moment procedural audio is rarely used in industry, partly due to current procedural audio models not sounding as realistic or high quality as sample-based audio. By creating more realistic models for sound and increasing the number of sounds that can be modeled, procedural audio may become more appealing to industry.

A popular sound effect in creative content is applause sounds. Hand clapping is common to almost all cultures, and applause is the sound of many people gathered in one place clapping their hands. Applause is most often heard when an audience shows appreciation for a performance and hence is frequently used in sound design to give a sense of realism and immersion.

The aim of this paper is to research and analyze applause sounds in order to implement and evaluate a controllable applause synthesis model that addresses the needs of the sound design community for high-quality procedural audio. The approach taken herein is based in part on the applause synthesis approach described by Peltola et al. [3,4]. However, we introduce a novel ensemble approach to render many clappers in a spatially distributed environment, with accurate reverberation. The applause also features realistic build-up and fade-out based on natural audience response, as well as an "enthusiasm parameter" to simulate the effect of greater perceived noise levels for the same number of clappers in an enthusiastic audience. Our implementation is real-time, interactive, and entirely browser-based and hence may be accessed by anyone, without dependence on specialist software or skills. Finally, we provide subjective evaluation, comparing our method against both recorded samples and the results of four other popular sound synthesis techniques.

The paper is structured as follows. Sec. 2 gives the background, including insights from the literature that were used to construct our model. Sec. 3 describes the full implemen-

tation, from individual hand clap generation to applause construction to a full web application. Sec. 4 contains all the results of subjective evaluation from listening tests. Finally, Sec. 5 provides the conclusions, including both discussion and critique of the work, and directions for further research.

2 BACKGROUND

2.1 Clapping Sounds

Hand clapping is often used as a percussion-based musical instrument or during a live musical performance as a way of engaging an audience and for an audience to show appreciation. Clapping is also one of the simplest ways to approximate an impulse, without the need for any equipment. Seetharaman and Tarzia [5] found that with a small amount of additional but automated signal processing, the claps can approximate impulsive sources to produce reliable acoustical measurements. However, hand claps do not have completely flat impulse responses, are not completely omnidirectional, have significant duration, and are not very high energy.

In [6], the impulse-like nature of hand claps was exploited to devise a means of Human-Robot Communication. The hand claps and their timing are relatively easy for a robot to decode and not that difficult for a human to encode. Yet many other forms of binary encoding exist, and as voice recognition and related technologies continue to advance, the need for hand clap-based communication diminishes.

The seminal paper in the field of clapping studies is Repp's 1987 study [7]. Experiments therein mainly involved 20 test subjects who were each asked to clap at their normal rate for 10 seconds in a quiet room. The spectra of individual claps varied widely, but there was no evidence of influence of sex or hand size on the clap spectrum. The author also tried to determine whether the subjects were able to extract information about the clapper from listening to the signal. Subjects generally assumed that slow, loud, and low-pitched hand claps were from male clappers, and fast, soft, and high-pitched hand claps were from female clappers. But this was not the case. Speed, intensity, and pitch were uncorrelated with sex, and test subjects could identify gender only slightly better than chance. Perceived differences were attributed mainly to hand configurations rather than hand size.

The fact that sonic aspects of hand claps can differ so significantly and can often be identified by listeners suggests that it may be possible to tell a lot about the source by signal analysis. Such was the case in work by Jylhä and colleagues, who proposed methods to identify a person by their hand claps [8] or identify the configuration (à la [7]) of the hand clap [9].

2.2 Synthesis of Hand Claps

Perceptual experiments in [7] were used to classify the types of hand claps used by participants in an experiment. Acoustical analysis found that the spectral shape of a clap related to the position of the hands. Palm contact was associated with low frequency content in a clap and finger-based contact was linked to higher frequency content. Some clapping types were more common than others. How we perceive applause could possibly be linked to the most common clapping type as it dominates the frequency content of the overall applause. However as a result of the small sample size (20 participants for 8 clapping types), these results are not an accurate representation of the most popular types of clapping but instead can be used as a guideline.

Spectral analysis of hand claps recorded in an anechoic chamber was performed in [4]. For each clap type, [4] gave mean attack and decay times, and [3,4] provided both mean and standard deviations for center frequency, bandwidth, and gain. Cupped palm-based contact resulted in longer decay and attack rates. The attack could be related to a small amount of air trapped between the palms of the hand that is pushed out as they come together, creating a small burst of air before impact. Additional attack could also be because it is unlikely that all fingers will make contact with the other hand simultaneously but instead occur one after the other.

Summarized results from [7,3,4] describing the characteristics of each hand clap type are given in Table 1.

2.3 Characterizing Applause

In [10], Néda et al. recorded applause from several theatre and opera performances. They observed that the applause begins with incoherent random clapping, but then synchronization and periodic behavior develops after a few seconds. This transition can be quite sudden and very strong and is an unusual example of self-organization in a large coupled system.

Uhle [11] considered the more general question of identifying applause in an audio stream. Understanding of applause is also useful for encoding the applause [12–14] that so often accompanies musical recordings, and the important spatial and temporal aspects of applause signals are known to make them particularly challenging signals to encode and decode [15]. As noted in [16], the more standard perceptual features like pitch or loudness do not do a good job of characterizing grainy sound textures like applause. Adami et al. [17] introduced a new feature, applause density, which is loosely related to the overall clapping rate but derived from perceptual experiments. In [18,19], density and other characteristics were used to investigate the realism of upmixed (mono to stereo) applause signals. Upmixing is an important problem in its own right. But the placement and processing of sounds for a stereo or multichannel environment can be considered part of the general problem of sound synthesis.

Often applause is a random selection of people clapping at a semi regular-rate, but applause may suddenly synchronize and then fall out of synchronization. In [3], a simple algorithm was given for modeling the spontaneous synchronization of applause. Each clapper refers to a master rate. Upon synchronization each clap attempts to get closer to this master rate by scheduling the next clap event $C[n]$ to be

$$C[n] = E[n] + A(M[n] - E[n]) \quad (1)$$

Table 1. The eight types of hand clap, their probability of occurrence, and signal characteristics. Derived from [3,17,4].

Clap type	Hand placement	Occurrence (%)	Center frequency (Hz)	Bandwidth (Hz)	Attack time (ms)	Decay time (ms)
A1	Angled, palm-to-palm	10	776	167	4.0	9.5
A1–	A1, very cupped	0	1,037	246	1.3	7.3
A1+	A1, flat	0	701	105	5.0	11.3
A2	Angled, intermediate	45	1,056	209	3.9	6.0
A3	Angled, fingers-to-palm	30	1,397	181	3.2	5.2
P1	Parallel, flat, palm-to-palm	0	1,101	181	4.7	10.9
P2	Parallel, flat, intermediate	10	846	195	3.0	8.7
P3	Parallel, flat, fingers-to-palm	5	1,505	280	3.5	4.9

where $E[n]$ is the next event for this clapper, $M[n]$ is the Master next event, and A is the affinity parameter (range 0 - 1). This enables the applause to go from a random distribution, $A = 0$, to fully synchronized, $A = 1$.

Characterizing aspects and features of applause is important in order for a user to be able to modify the sound to exaggerate a particular trait. Adami et al. confirmed that applause density is an appropriate perceptual attribute [16]. However, applause density metrics have not been established, though it is strongly linked to audience size. Applause can also be described in terms of the collective enthusiasm [7,9,3], which is associated with the synchronization of clappers in the audience [20], and the hand clapping rate [21–23].

2.4 Synthesis of Applause

Previous attempts of implementing applause synthesis models faced a variety of issues. Peltola and colleagues [3,4] presented physics-based analysis, synthesis, and control systems intended to produce individual hand-claps or mimic the applause of a group of clappers. They succeeded in generating isolated claps yet struggled with multiple clappers as the model made use of just one envelope generator that would be interrupted every time a new clap was generated. Peltola also made use of the standard Schroeder-Moorer algorithmic reverb [24,25], which is known to lack many aspects of realistic reverberation. There was no subjective evaluation of the results or comparison of synthesized sounds against recorded sounds. Applause was also synthesized in [17] and [26], but both were forms of reconstruction, not fully procedural or physical approaches. Many recorded individual clap signals were layered in [17], and [26] used a dictionary of sound grains, also constructed from individual clapping recordings. Neither approach was formally evaluated against other methods or recordings in terms of plausibility or a similar quality measure.

Perceptual evaluation of a range of synthesized sound effects including applause was performed in [27]. When rating four synthesis models, a reference (recorded sample) and an anchor in terms of realism, the recorded sample was rated far more realistic than all synthesized applause sounds ($p < 0.0001$ in all cases). As noted by the authors, "No synthesis technique was capable of producing convincing applause." The models were instead managing to reproduce some recognizable elements of applause such as the general noise heavy content with a few distinct claps. Yet for other sounds (bees, rain, wind. . .), the realism of these models was often rated quite close to actual recordings.

2.5 Spatial Representation of Applause

The acoustics of a space plays a large part both in how we generate and how we perceive the sound of applause [28,29]. Applause also varies in its sound based on where the listener is positioned relative to where people are clapping. If one is part of the applause it is common to distinctly hear claps that are nearby and identify their location relative to the listener. Watching the applause, on the other hand, it can be hard to pick out the detail of claps; instead the listener perceives a source with diffuse directionality and less intimate detail.

3 DESIGN AND IMPLEMENTATION

Our approach involves firstly being able to model single, isolated claps realistically with each using their own envelope. Multiple clappers can then be generated and applied to a space modeling the environment the claps are hosted in. For generating a single clap a noise source is passed through a set of filters and amplitude envelope. For applause, multiple claps can be used with claps further away from the listener attenuated appropriately to simulate distance. Claps closer to the user are then heard in greater detail and have greater distinction from the background noise.

When applause is taking place but not synchronized, the audience members tend to be clapping at a similar rate but not in sync with one another, creating a random distribution of claps in the time domain. Claps can therefore be assumed to be clapping at a general rate with some variance as opposed to completely random rates of applause, which results in an in-cohesive applause.

The implementation is compartmentalized in order to aid with testing and debugging. Distinct elements that make up the models are distinguished so that each carries out a well-defined role within the model. Row refers to the number of rows in the audience, and column refers to the stereo position of the clapper in a row. In actual implementation there are always 5 clappers per row, and the number of rows can range from 1–10. Fig. 1 shows a block diagram of the synthesis model, where it is assumed that there is only one row, with 5 clappers.

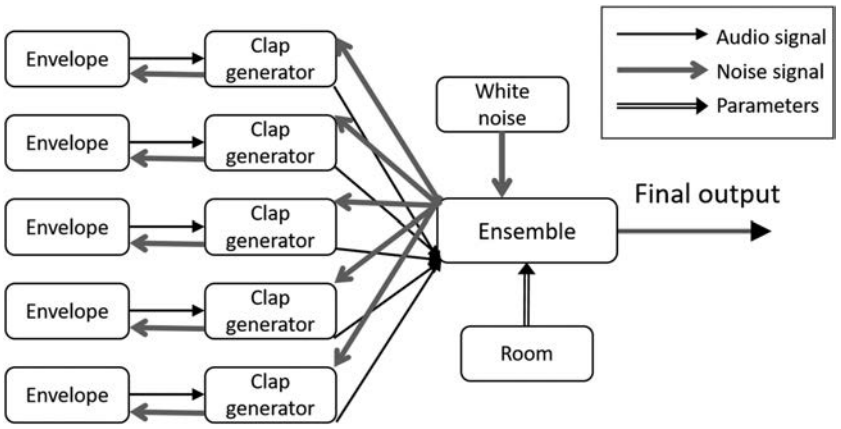

Fig. 1. Block diagram of the applause synthesis model. Noise is passed to the Ensemble block and down to Envelopes to feed an array of clappers. Audio is returned from the Envelope to clappers then to the Ensemble. The Ensemble returns an array of panner nodes and convolutional reverb parameters related to the environment in which the applause occurs.

3.1 Single Clap Generator

Subtractive synthesis often starts with a rich source that then passes through an envelope and resonator filters to give the desired spectral and dynamic attributes [30]. As in [4], we will use this approach to extract the desired spectrum by filtering out unwanted aspects of the incoming signal.

Continuous white noise was used here as the excitation signal. Both linear and exponential ramp envelopes are used. They take two arguments, delay and maxGain. Delay allows the initiation of the trigger to occur "delay" ms after the current time and maxGain defines the maximum value to which the envelope reaches. When creating an envelope, one needs to pass the current audio context, audio source, and output destination. To ensure variation between claps, attack, decay, and release values are set in a separate method.

The exponential envelope is used since it achieved the most realistic synthesis of a single clap. Amplitude is reduced exponentially down to 0.03 and then linearly to 0. The linear decay in the latter stage is due to the Web Audio API's ramp function not allowing an exponential ramp that decreases to 0. Exposing the linear decay rate also provides a release parameter to be used to add a tail on the end of each clap. It acts as an enthusiasm control where the longer the release, the more enthusiastic the applause sounds. This mimics the effect of people clapping faster and/or harder when clapping enthusiastically and there being a greater number of reflections that occur in the room, resulting in a denser sound.

The clap class incorporates the envelope and noise classes above to produce a single clap of a specified type. When creating an instance of a clap, one needs to specify the type of clap, one of the eight types of clapping identified by Repp [7]. Given the clap type, the parameters for that specific clap, such as frequencies of band pass filters, attack and decay rates, and variation in bandwidth and frequency based on [4], are set.

The filters were all implemented using the Web Audio API specification.[1] They are atypical second order IIR filters, as given in Eq. (2), for sampling frequency Fs, center frequency f_0, and quality factor Q. As such, they differ from the implementations in [4].

$$\begin{aligned}
&\text{Lowpass}: \quad H(z) = \frac{1-\cos\omega_0}{2}\frac{1+2z^{-1}+z^{-2}}{1+\alpha_{Q_{dB}}-2\cos(\omega_0)z^{-1}+(1-\alpha_{Q_{dB}})z^{-2}}\\
&\text{Highpass}: \quad H(z) = \frac{1+\cos\omega_0}{2}\frac{1-2z^{-1}+z^{-2}}{1+\alpha_{Q_{dB}}-2\cos(\omega_0)z^{-1}+(1-\alpha_{Q_{dB}})z^{-2}}\\
&\text{Bandpass}: \quad H(z) = \frac{1-z^{-2}}{1+\alpha_Q-2\cos(\omega_0)z^{-1}+(1-\alpha_Q)z^{-2}}\\
&\text{where} \quad \omega_0 = 2\pi f_0/F_s,\ \alpha_Q = \frac{sin\omega_0}{2Q},\ \alpha_{Q_{dB}} = \frac{sin\omega_0}{2\cdot 10^{Q/20}},
\end{aligned} \tag{2}$$

Two band pass filters were used, one for the common peak around 1 kHz and another for the upper peak around 2.5 kHz. This enables control over the high and mid content of the clap to produce more accurate results. Each band pass filter was connected to a low pass and high pass, respectively, to limit the frequency range, as frequencies above 2.5 kHz and below 500 Hz are not an essential component of the clap. The low pass filter's Q value was set high to 20, in order to exaggerate the high frequency peaks found in claps around 2.5 kHz. A flow diagram of the clap generator is depicted in Fig. 2.

To produce a single clap, it first checks that there is not a clap already in process by checking the state. It then sources the values for attack, decay, gain deviation, and frequency deviation. The attack and decay values are used to set the envelope parameters before producing the clap.

3.2 Room

Making use of the data found by Repp [7] and Peltola [4] regarding clap types, frequencies, and attack and decay rates will enable the synthesis of the eight clap types. How-

[1] https://webaudio.github.io/web-audio-api/#filters-characteristics

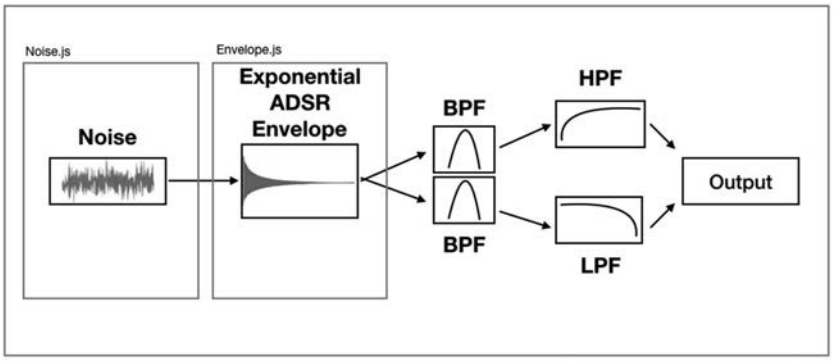

Fig. 2. Flow diagram showing how a single clap is generated with characteristics based on the type selected. The approach is subtractive synthesis, similar to the implementation by Peltola et al. [3]. ADSR represents Attack Decay Sustain Release, and BPF, HPF, and LPF are band pass, high pass, and low pass filters, respectively.

ever some of these measurements are based on an anechoic chamber environment and will sound unusual as we do not often hear claps in a space without reverb. Thus, these short impulses of each clap are placed into an "environment" to provide ambient sound.

To model the effects of reverb on applause, each clap would have to have its own unique reverb specific to its position relative to the listener, position within the room, and objects obstructing the listening path. This is intensive and would require a great deal of memory to compute.

Convolution-based reverb is applied to each clap, which is then positioned in the stereo space. Impulse response audio files may be loaded to simulate a variety of room sizes. This implementation includes room sizes generalized to small, medium, and large. Users may also upload their own impulse response for more specific control over the sound of the reverb of the applause.

Panning position values may be changed by setting the selected width of the mix. Each panner also has an associated gain so that the position of the listener may also be altered by adjusting the output volume of each stereo position.

3.3 Ensemble

The applause is stored as a multi-dimensional array where the x-direction refers to the position in the stereo space, y-direction refers to the row number (higher row number implies further from the listener position), and each element represents an individual clapper. As the distance increases, that row of clappers have their output volume set lower to mimic the greater distance between them and the listener. The maximum gain for the envelope of each clap is set to 0.25 and the minimum gain is set to 0.19. These values were chosen as a result of trial and error to achieve the correct representation of a single hand clap. The output gain, controlling volume within the applause, of each clap is set as $0.25-0.2y/Y_{max}$, where Y_{max} is the maximum row, which is by default 10. As the row number increases, the output volume becomes quieter to portray distant clappers. This implementation suits the sounds of small and larger audiences, as commonly with smaller audiences one can hear all the claps distinctly. This is modeled correctly here since a smaller number of rows does not reduce the volume too much. Only with larger audiences does one stop hearing the claps further away but instead hears a sound texture of pseudo-randomly distributed clapping sounds.

By defining all the intervals during the initialization it allows the audience size to be altered as a parameter. The parameter can be set within the range of 1–10 and only clappers up to the chosen value are created. The clap type is randomly selected according to a probability distribution based on [7].

A flow diagram of the ensemble generator is depicted in Fig. 3.

3.4 A Web Application

The FXive website[2] hosts a range of fully procedural, real-time sound synthesis models (fire, wind, whistling, creaking, ocean waves, alarms...) that run directly in the browser. The aim was for the applause synthesis model to be included as part of FXive [31]. Thus it was implemented in JavaScript while making use of and integrating with HTML5, CSS, and several technologies specific to web-based audio.

The Web Audio API[3] supports complex audio and musical web browser-based processes and applications [32]. It provides an interface for real-time audio processing and synthesis from within the web browser. Nodes of different types are defined within an audio context and then connected together to form audio chains.

Nexus UI is a library that assists in creating interfaces for web audio instruments [33]. It includes a range of UI components designed for spatialization, visualization, and other general components. It gives a simple way to define an interface for a real-time sound model and to then control the parameters of that model in real time.

JSAP is a JavaScript framework for web audio applications, using the Web Audio API to define a plug-in structure

[2] https://fxive.com/#

[3] https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API

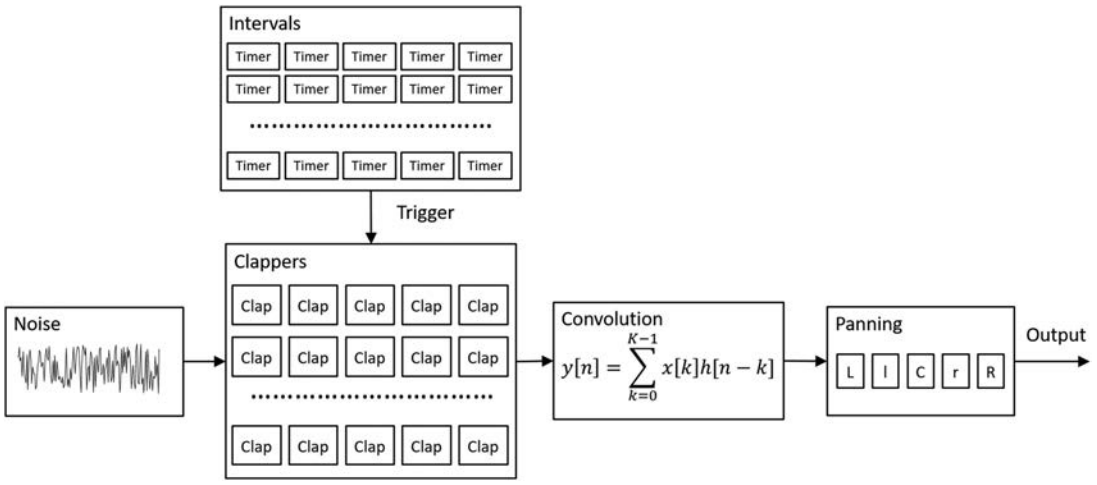

Fig. 3. Flow diagram of the ensemble approach to generating applause from hand claps. Panning is to 5 positions, far left (L), left (l), center (C), right (r), and far right (R).

and host integration for a plug-in. JSAP enables cross adaptive processing, which prevents multiple calculations taking place on features. Semantic interaction allows parameters of JSAP plug-ins to be changed based on textual information [34]. FXive uses JSAP to create audio plug-ins for each of the models available on the website. By using JSAP instead of the traditional Web Audio API integration it will enable future proofing to a certain extent. The clap plug-in could be used in other sound models, with additional audio effects, or in a larger platform for applause analysis.

We bring together all elements to implement a JSAP plug-in with controllable JSAP parameters. It handles creation of all clappers and connecting them to a Room to create an applause in a space. All timing related to the applause including overall clapping rate, synchronization, getting out of synchronization, start duration, stop duration, starting the applause, stopping the applause, and enthusiasm of the audience is managed by this class. This class provides an interface for JSAP parameters to interact with the Room class without making Room another JSAP plug-in.

3.5 Modifiable Parameters

What separates real-time synthesis models from sample-based systems, apart from being generated in real time, is that they can be modified to achieve a wide range of different sounds in real time. This model contains a set of relevant and useful JSAP controllable parameters that enable a user to manipulate the sound of the applause to what is required. The interface is made up of Nexus UI components that change the values of these JSAP parameter values. While the parameters are controlling the Ensemble the interface components are defined in the controls.js file and the handlers are implemented in the main.js file.

The Start/Stop parameter refers to the ability to start and stop the applause, the option to set the duration it takes before all the clappers are active, and the duration it takes all the clappers to deactivate. This is key for creating realistic applause, as quite often these timings can have an impact on the overall mood presented by the sound. Having a shorter start time tends to imply that the audience is overall quite enthused about what they are applauding. The start/stop parameter is implemented as a Nexus button that toggles the state between starting and stopping. The duration parameters are Nexus number boxes, which are limited between 100–20,000 ms.

The Clapping Rate refers to the average speed of all the clappers when unsynchronized. This parameter allows a user to alter the average speed to create different tones. Generally, the faster the rate the more enthusiastic the audience is compared to slower, more disinterested applause. A general clapping interval of around 250–270 ms sounds relatively enthused whereas a 330–400 ms interval sounds normal verging on disinterested. The available range to choose from is 190–500 ms.

Synchronization and variation: Each clapper needs to be given an interval in order to be triggered on a periodic basis. In real applause, everyone claps at a similar rate. However, this creates a more distinct background noise from the distant clappers and sometimes results in clappers hitting at the same time, causing large sudden peaks in the audio. Thus, the implementation has a clap tempo randomizer, dependent on the number of rows, to make further clappers clap at a faster rate than ones closer to the listener. Claps in the distance can be up to 50% faster, while claps closer to the listener can be up to 30% slower than the general clapping rate. This variation and randomness creates a very realistic sound of applause of a general constant clapping rate.

The applause model attempts to implement affinity characteristics for synchronization. As the affinity value is increased toward 1 the clapping rate would adjust to the synchronized clapping rate. Eq. (3) determines the intervals for each clapper in order to trigger claps quasiperiodically.

$$\begin{aligned} &R_{i,j} \sim U([1 + {}^{i}/_{2Y_{max}}, 0.7 + {}^{0.3i}/_{Y_{max}})] \\ &T_{i,j} = C(1-A)R_{i,j} + A \cdot M \\ &i = 0, 1, ...Y_{max} - 1,\ j = 0, 1, ...N_C - 1 \end{aligned} \tag{3}$$

where $R_{i,j}$ is a random number in the uniform distribution from $1+i/2Y_{max}$ to $0.7+0.3/Y_{max}$, $T_{i,j}$ is the interval time, C is the Clapping Rate of the current clapper, A is the affinity, and M is the master synchronization rate. i and j range over the number of rows of clappers Y_{max} and number of clappers in a row N_C, respectively.

While this works with relative success and replicates some of the natural behavior of synchronized applause, where some claps are not completely synchronized with the majority, there could be a stronger sense of synchronicity between the claps. Desynchronization is performed by decrementing the affinity parameter until the tempo is reset to the current general clapping rate.

To implement variation of clapping rate, the Nexus Interval class was used, which utilizes the W3 clock for timing. This provides each clap with greater timing accuracy compared to JavaScript's internal clock. Each interval is provided with a callback function for triggering its associated clap.

The perceived Enthusiasm of a crowd depends on many factors. It can be difficult to mimic in just applause since it tends to be accompanied with the sounds of cheering, whistling, and stomping, as well as increase in both clapping rate and overall volume for each clap. Here, the enthusiasm parameters control just the release of each clap. A longer release results in each clap contributing a bit more noise to the overall sound, which creates the illusion that more people are clapping harder and faster. This parameter ranges from 0–200 ms. A value of 60 sounds relatively normal in terms of enthusiasm whereas anything above 70 sounds quite enthusiastic.

With any sound source, the environment in which it is hosted can greatly affect the sound that we recognize. All the clap sounds have been implemented from information obtained from claps in anechoic conditions. This sound is not what we associate with applause as more commonly there are reflections of sound occurring within the space. The Room Size parameter allows a user to select an impulse response for the room the applause is hosted in. This impulse response is loaded into a buffer and applied to each stereo position using convolution to create a wet (with reverb) signal that is then mixed with the dry signal. The user can select from different room sizes and contexts:

- Small – Office-like space or classroom
- Medium – Small/medium music venues or halls
- Large – Church-like reverb

In order to model a wide variety of applause situations, a user needs to be able to alter the perceived audience size. As the Number of Rows increases, it introduces a new layer of distant clappers at each point in the stereo field, resulting in a denser sound that is perceived as more people being present during the applause. Larger crowds tend to be placed in larger rooms so claps further away are relatively quiet, but in smaller rooms with smaller crowds each clap tends to be quite distinct and easy to pick out.

With procedural audio focused toward game applications, being able to place a listener within the applause is a useful feature to include. Room Position makes use of the Nexus Pan2D to get the position of a user relative to a set of speakers. Nexus Pan2D is an interface and JavaScript helper function, included in the NexusUI API [33], for moving a sound around an array of speakers. We defined four nodes to represent the four stereo positions and one for the listener. The interface provides the Ensemble plug-in with the distances of each speaker from the user and uses these values to set the gain values for each stereo position.

The interface provides buttons that correspond to each of the eight **Single Clap** types, enabling the user to clap as part of the applause. They are implemented using an array of radio buttons that trigger one element in an array defined by the Ensemble plug-in.

Being able to alter the stereo **Mix Width** allows a user to create a sense of where the applause is taking place. With maximum width the sound is immersive and places the listener within the crowd whereas little stereo width makes the applause directional as if it is in front of the listener. The range is from 0.2–2, where 2 is complete stereo width and 0.2 is very directional audio. 0.2 is used as a minimum value as having all the claps in the same place generated a poor output result.

This model has the ability to load **Presets** to create common applause sounds. When selected the preset handler sets all of the parameters of the Ensemble class to predefined values. The current implementation defines a set of six presets of common applause sounds. The "small room" preset, for instance, showcases office-based applause or small part speech applause, which are both scenarios used often in TV and film. The medium room example is attuned to concert applause from enthusiastic fans. There is also a dramatic applause preset which involves people joining in over an extended period of time before coming together in a crescendo of enthusiastic applause. This kind of sound is often used in films and TV shows usually to show the success of a performance. Others include a restrained "Golf Applause" and enthusiastic "TV Studio" applause.

4 SUBJECTIVE EVALUATION

The subjective evaluation of applause synthesis was based on the approach in [27], whereby different applause synthesis models were evaluated against each other and a recorded sample (the reference) in order to determine which synthesis method produces the most realistic result. Evaluation was performed using the Web Audio Evaluation Toolkit [35], which provides a platform for perceptual audio evaluation experiments.

Two recorded samples were used as references for two applause situations: a small room with few people in the audience [36] and a larger room with more people [27]. The samples from our model were produced from the presets "small room, medium attendance, enthusiastic" and "medium room, large attendance, enthusiastic" to replicate the situations of the reference samples. Evaluation also included samples from the synthesis techniques used in the evaluation in [27]; sinusoidal modeling, concatenative synthesis, marginal statistics, and statistical modeling.

The audio perceptual evaluation (APE) method [37] was applied. This is a multistimulus paradigm to present a user with a continuous scale (Very unrealistic - Quite unrealistic - Quite realistic - Very realistic) where samples can be played and dragged across the scale to rate them. Whether the participant had audio experience was confirmed in order to compare how the model performs for audio professionals and inexperienced participants. For consistency, all samples were set to the same loudness and a 44.1 kHz sample rate.

For each participant, two tests were presented in randomized order, one for the Small room applause scenario and one for the Medium-sized room scenario.

A sample of 10 people was asked to rate a set of samples in terms of realism, using the interface depicted in Fig. 4. 60% of participants had experience with audio by either playing an instrument or working with audio-related technology. Each participant completed the evaluation while wearing Bose QC35 headphones to ensure consistency in terms of the listening conditions.

The reference and synthesis models used for each test are shown below, along with mean scores from the evaluation in [23]:

- Reference (Small room [36] or Medium size room [27]).
- Sinusoidal Modeling [38], Mean 0.32. Sinusoidal modeling is an advanced version of additive synthesis where any periodic waveform can be modeled as a sum of sinusoids at different amplitudes and phases.
- Concatenative Synthesis [39], Mean 0.29. A form of granular synthesis that involves a sound being constructed from small segments (grains) of audio around 10–1,000 ms in duration. The element of synthesis comes from the selecting and composition of these segments such that there are no perceptual discontinuities in the output sound.
- Marginal Statistics [40], Mean 0.09. The aim of marginal analysis is to be able to shape an input source based on some mathematical data obtained from analyzing a target sound. The specific characteristics that marginal analysis involves are the mean, variance, and skew.
- Statistical Modeling [40], Mean 0.5. A statistical model represents some data by a range of statistical tools. In the case of audio it aims to represent a sound as a set of mathematical characteristics. This modeling technique includes the parameters used in marginal statistics, cross-sub-band envelope correction, and cross-sub-band modulation corrections [23]. These mathematical characteristics are used to shape some form of noise, commonly Gaussian white noise, to synthesize a resultant output signal.

4.1.1 Results

Fig. 5 gives the results of the multi-stimulus test, and Table 2 summarizes these results in terms of mean and standard deviation for perceived realism of the samples, including the reference and each synthesis method. It is clear that the model is producing a highly realistic applause sound that comes close to that of a real-life recording, whereas the other synthesis methods do not meet the same standard. In addition, the ratings of some of the other synthesis models vary quite dramatically from participant to participant. An example is sinusoidal modeling, where participant ratings vary for the large audience from a value of 0.021 up to 0.81. This could be due to people focusing carefully on the ordering at the more realistic end and just putting the others in a position relatively lower than the top samples.

With the synthesis of a smaller audience our model produces a result very close to the reference, with a difference of only 4% in their mean scores (0.92 compared to 0.96); see Fig. 5 (top). In fact 60% of people put our model ahead of the reference sample, which shows how realistic a representation the model produces. The next highest scoring model was concatenative, which scored a mean of 0.12.

Results for the Medium Room, Large Audience are shown in Fig. 5 (bottom). Here, the reference achieved the highest score in terms of its realism, 0.98, the proposed model achieved a mean score of 0.92, and statistical synthesis was the next best performing model with a mean of 0.35.

Aggregating results, the overall mean score for the reference was 0.97, for the proposed method was 0.92, and for the next best method, Statistical Synthesis, was 0.20. Realism of the recorded sample may also be associated with attributes not related to clapping. The small room recording has artifacts that might be due to cheers, and there was a quiet piano tone at the start of the medium-sized room reference sample. Both of these may give away the fact that they were real recordings.

As the model produces such a realistic sound it was mostly compared to the reference sample rather than the other models, which get pushed to the bottom of the scale. This made it hard to draw meaningful conclusions regarding relative rankings of the other synthesis methods. Nor was it possible to find any statistically significant differences between participants with and without professional audio experience.

5 DISCUSSION AND CONCLUSION

We implemented a synthesis model for the real-time generation of the sound of applause. It first synthesizes the eight types of clapping by applying a bank of filters and envelopes to a noise source, following the characteristics analyzed by [7] and [4]. The sound of applause is made up of multiple claps occurring within the same space. Our model maps a 2D array of clappers to a 2D array of intervals that control the timing of the corresponding clapper. All the claps are then placed within a room using an impulse response and stereo positioning. There are also a variety of user-modifiable parameters in order to obtain a wide range of different applause sounds. Parameters include audience size, room size, enthusiasm, position in room, general clapping rate, synchronization clapping rate, and synchronization on/off.

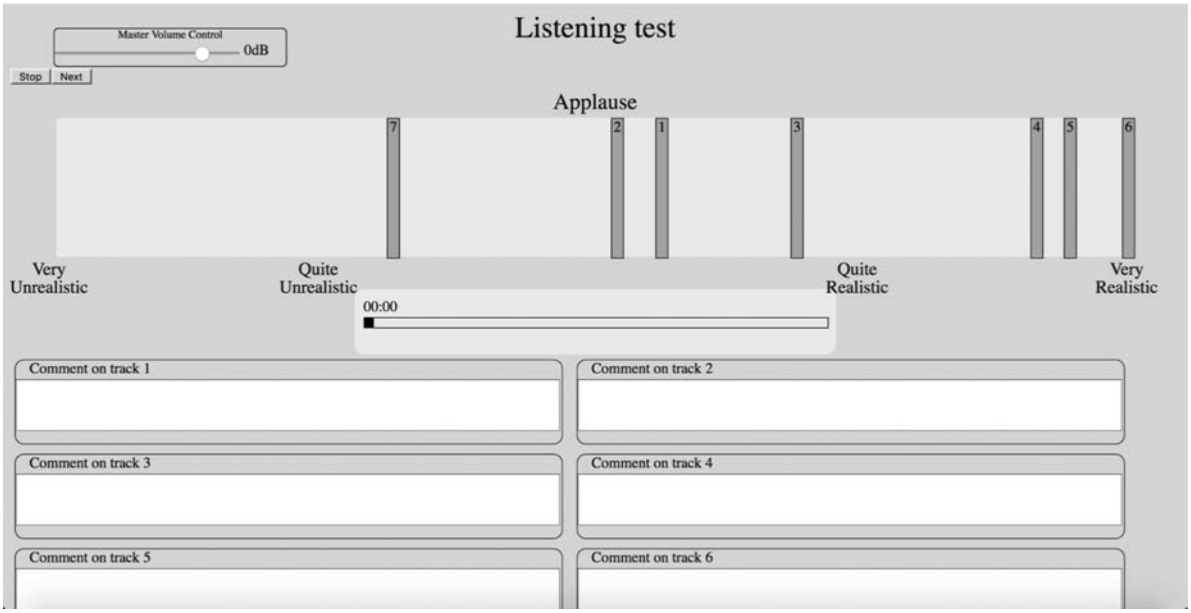

Fig. 4. A screenshot of the User Interface used for subjective evaluation.

Models already exist for applause but our model produces realistic results that are comparable to real-life recordings. In subjective evaluation of realism, our model scored a mean of 92% and was only beaten by the reference overall by 5%. The small room situation produced the most realistic representation of applause as 60% of people scored our model higher than the reference sample.

Though our model produced accurate and realistic results, improvements could be made to ensure a complete model of audience behavior and applause characteristics. Previous models included synchronization algorithms to create the sense of synchronized, quasiperiodic behavior. These tend to be implemented using the affinity-based model to gradually achieve synchronicity when affinity is 1. Though our model uses this algorithm, the synchronization lacks some realism since the strength is diminished when all the claps are triggered together. The reason for this is that the Nexus interval class uses the W3 clock and time stretching functionality to change the interval time smoothly by interpolating between tempos. This time stretch will naturally vary from one tempo change to another, causing unwanted delays. A possible way to resolve this issue is to use the W3 clock directly instead of the Nexus interface for it. This would result in greater control of the timing and eliminate the need for time stretches, which cause issues in regards to affinity-based synchronization.

Our model included meaningful parameters for controlling and manipulating how the applause sounds, one example being the enthusiasm parameter, which adds a further release to each clap, resulting in an increase of perceived average noise/volume and thus simulating a more enthusiastic sound. However, this is not the only factor that is affected by an audience showing greater enthusiasm. Applause density [17] can refer to the size of the audience and enthusiasm of the audience and result in an increased average noise level. People also tend to clap slightly faster when enthusiastic in order to increase the overall noise level.

There are a few common sounds that tend to come alongside applause. Enthusiasm can be demonstrated with clapping but also with "whoops," laughing, cheers, foot-stamping, and whistling. This would add to a truly real representation of applause from an audience.

The sound of applause is strongly related to the environment in which it is hosted and the location of the listener. Being able to model both being within the audience and potentially looking at the audience from a stage could be useful features to include in future models. Specifically modeling the listener as being part of the audience could

Table 2. Summary statistics for perceived realism of all sound samples.

	Small room, medium audience		Medium room, large audience	
Model	Mean	Standard deviation	Mean	Standard deviation
Reference	0.96	0.04	0.98	0.02
Proposed	0.92	0.09	0.92	0.09
Concatenative	0.12	0.1	0.09	0.12
Marginal	0.06	0.05	0.12	0.11
Sinusoidal	0.07	0.07	0.3	0.3
Statistical	0.06	0.05	0.35	0.26

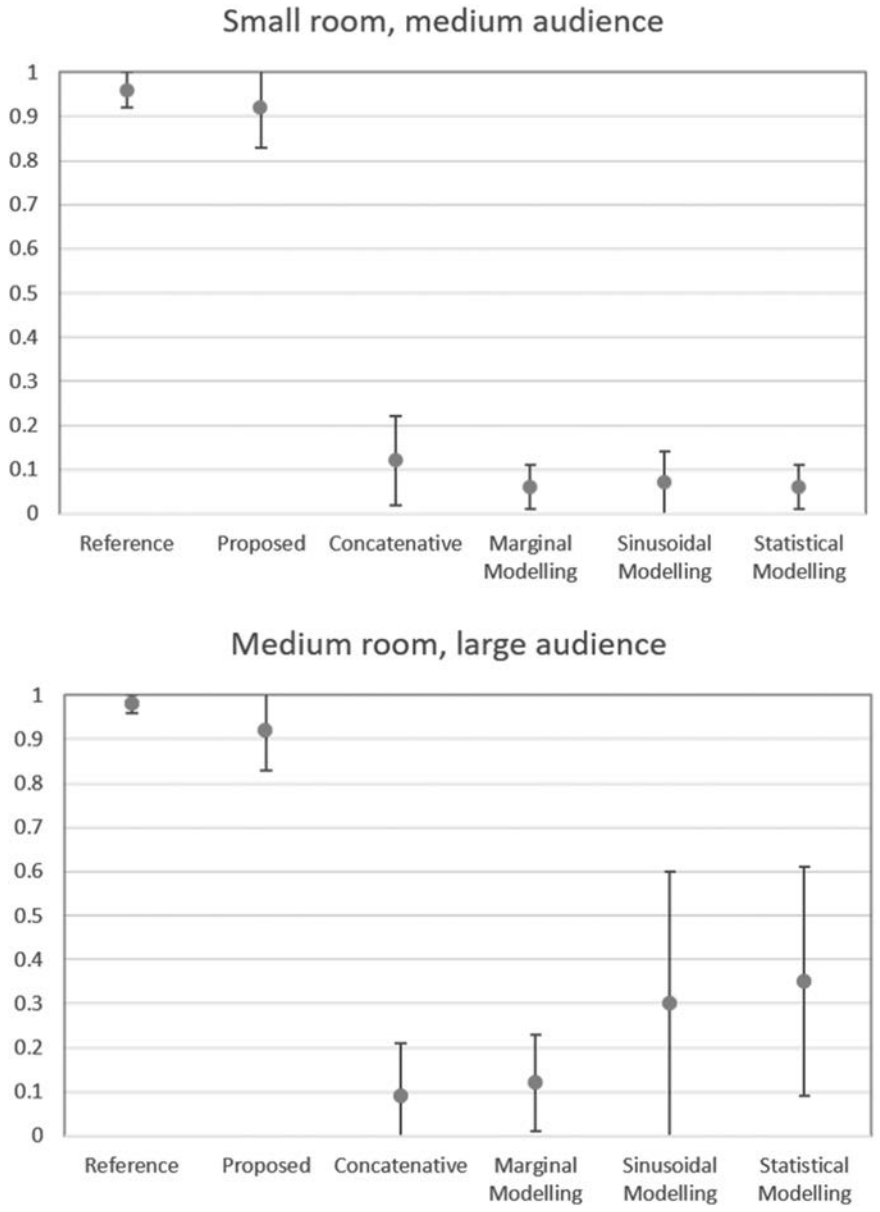

Fig. 5. Multistimulus testing results. Mean score and standard deviation error bars of the reference sample and each model from both small and medium audiences, when rating in terms of realism.

be enhanced using spatial rendering techniques to create a more immersive experience.

Finally, the evaluation might not be considered fair as our model is tailored for generating applause whereas the others are general synthesis methods. However, they are popular and represent the state of the art. These synthesis methods have in the past performed better than dedicated models for sounds, such as a concatenative model of wind, which was rated more "believable" than a physically inspired model [27]. Authors have also claimed that similar methods produce strong results for modeling sound textures, e.g. [41,42], which could include applause. That said, further evaluation should include the Peltola method [3] among the compared algorithms since it is the closest equivalent and unlike other methods uses some knowledge of the nature of applause, not just the statistics or features of the applause signal. Recent work also showed very little correlation between proposed objective audio similarity measures and the results of subjective testing for a variety of sound synthesis models [43]. So further research is needed to establish relevant features that can be used in objective evaluation of realism and audio quality in sound synthesis.

6 ACKNOWLEDGMENT

This work has been supported by the EPSRC/Innovate UK grant EP/R005435/1 - Autonomous System for Sound Integration and Generation (ASSIGN), and the EPSRC grant EP/L019981/1 - Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption.

REFERENCES

[1] A. Farnell, *Designing Sound* (Westchester Book Composition, 2016).

[2] R. Selfridge, D. Moffat, E. Avital and J. D. Reiss, "Creating Real-Time Aeroacoustic Sound Effects Using Physically Derived Models," *J. Audio Eng. Soc.*, vol. 66, no. 7/8, pp. 594–607 (2018 July/Aug.), https://doi.org/10.17743/jaes.2018.0033.

[3] L. Peltola, C. Erkut, P. R. Cook, and V. Välimäki, "Synthesis of Hand Clapping Sounds," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 3, pp. 1021–1029 (2007). https://doi.org/10.1109/TASL.2006.885924.

[4] L. Peltola *Analysis,, Analysis,, Parametric Synthesis, and Control of Hand Clapping Sounds, M.Sc. thesis,* Helsinki University of Technology (2004).

[5] P. Seetharaman and S. P. Tarzia, "The Hand Clap as an Impulse Source for Measuring Room Acoustics," presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), convention paper 8585.

[6] K. Hanahara, Y. Tada, and T. Muroi, "Human-Robot Communication by Means of Hand-Clapping (Preliminary Experiment with Hand-Clapping Language)," *IEEE Int. Conf. Syst. Man Cybernet. (ISIC)*, pp. 2995–3000 (2007 Oct.).

[7] B. H. Repp, "The Sound of Two Hands Clapping: an Exploratory Study," *J. Acoust. Soc. Amer.*, vol. 81, no. 4, pp. 1100–1109 (1987 Apr.), https://doi.org/10.1121/1.394630.

[8] A. Jylhä, C. Erkut, U. Simsekli, and A. T. Cemgil, "Sonic Handprints: Person Identification with Hand Clapping Sounds by a Model-Based Method," presented at the *AES 45th International Conference: Applications of Time-Frequency Processing in Audio* (2012 Mar.), conference paper 1-4.

[9] A. Jylhä and C. Erkut, "Inferring the Hand Configuration from Hand Clapping Sounds," presented at the *11th International Conference on Digital Audio Effects (DAFx)* (2008).

[10] Z. Néda, E. Ravasz, Y. Brechet, T. Vicsek, and A. -L. Barabási, "The Sound of Many Hands Clapping: Tumultuous Applause Can Transform Itself into Waves of Synchronized Clapping," *Nature*, vol. 403, no. 6772, pp. 849–850 (2000), https://doi.org/10.1038/35002660.

[11] C. Uhle, "Applause Sound Detection," *J. Audio Eng. Soc.*, vol. 59, no. 4, pp. 213–224 (2011 Apr.).

[12] A. Adami, A. Herzog, S. Disch, and J. Herre, "Transient-to-Noise Ratio Restoration of Coded Applause-like Signals," presented at the *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2017), https://doi.org/10.1109/WASPAA.2017.8170053.

[13] S. Disch and A. Kuntz, "A Dedicated Decorrelator for Parametric Spatial Coding of Applause-like Audio Signals," in A. Heuberger, G. Elst, and R. Hanke (Eds.), *Microelectronic Systems*, pp. 363–371 (Springer-Verlag, Berlin Heidelberg, 2011), https://doi.org/10.1007/978-3-642-23071-4_34.

[14] G. Hotho, S. van de Par, and J. Breebaart, "Multichannel Coding of Applause Signals," *EURASIP J. Adv. Signal Process.*, vol. 2008 (2008), https://doi.org/10.1155/2008/531693.

[15] M. -V. Laitinen, F. Kuech, S. Disch, and V. Pulkki, "Reproducing Applause-Type Signals with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 59, no. 1/2, pp. 29–43 (2011 Jan.).

[16] A. Adami, A. Taghipour, and J. Herre "On Similarity and Density of Applause Sounds," *J. Audio Eng. Soc.*, vol. 65, no. 11, pp. 897–913 (2017), https://doi.org/10.17743/jaes.2017.0034.

[17] A. Adami, S. Disch, G. Steba, and J. Herre, "Assessing Applause Density Perception Using Synthesized Layered Applause Signals," presented at the *19th International Conference on Digital Audio Effects (DAFx)* (2016).

[18] A. Adami, L. Brand, and J. Herre, "Investigations Towards Plausible Blind Upmixing of Applause Signals," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), convention paper 9750.

[19] A. Adami, L. Brand, S. Disch, and J. Herre, "Blind Upmix for Applause-Like Signals Based on Perceptual Plausibility Criteria," presented at the *20th International Conference on Digital Audio Effects* (2017).

[20] Z. Néda, A. Nikitin, and T. Vicsek, "Synchronization of Two-mode Stochastic Oscillators: A New Model for Rhythmic Applause and Much More," *Physica A: Statist. Mech. Appl.*, vol. 321, no. 1–2, pp. 238–247 (2003), https://doi.org/10.1016/S0378-4371(02)01779-X.

[21] K. Kawahara, Y. Kamamoto, A. Omoto, and T. Moriya, "Evaluation of the Low-Delay Coding of Applause and Hand-Clapping Sounds Caused by Music Appreciation," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9225.

[22] K. Kawahara, A. Fujimori, Y. Kamamoto, A. Omoto, and T. Moriya, "Implementation and Demonstration of Applause and Hand-Clapping Feedback System for Live Viewing," presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), convention paper 299.

[23] Z. Néda, E. Ravasz, T. Vicsek, Y. Brechet, and A. -L. Barabási, "Physics of the Rhythmic Applause," *Phys. Rev. E*, vol. 61, no. 6, pp. 6987–6992 (2000), https://doi.org/10.1103/physreve.61.6987.

[24] M. R. Schroeder, "Natural Sounding Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 10, no. 3, pp. 219–224 (1962).

[25] J. A. Moorer, "About This Reverberation Business," *Comp. Music J.*, vol. 3, no. 2, pp. 13–28 (1979), https://doi.org/10.2307/3680280.

[26] W. Ahmad and A. M. Kondoz, "Analysis and Synthesis of Hand Clapping Sounds Based on Adaptive Dictionary," presented at the *International Computer Music Conference (ICMC)* (2011).

[27] D. Moffat and J. D. Reiss, "Perceptual Evaluation of Synthesized Sound Effects," *ACM Trans. Appl. Percep.*, vol. 15, no. 2, pp. 1–19 (2018 Apr.), https://doi.org/10.1145/3165287.

[28] S. Farner, A. Solvang, A. Sæbo, and U. P. Svensson, "Ensemble Hand-Clapping Experiments Under the Influence of Delay and Various Acoustic Environments," *J. Audio Eng. Soc.*, vol. 57, no. 12, pp. 1028–1041 (2009 Dec.).

[29] S. Ntalampiras, "Generalized Sound Recognition in Reverberant Environments," *J. Audio Eng. Soc., vol. 67*, no. 10, pp. 772–781 (2019), https://doi.org/10.17743/jaes.2019.0030.

[30] R. Mignot and V. Välimäki, "Extended Subtractive Synthesis of Harmonic Musical Tones," presented at the *136th Convention of the Audio Engineering Society* (2014 Apr.), convention paper 9038.

[31] P. Bahadoran, A. Benito, T. Vassallo, and J. D. Reiss, "FXive: A Web Platform for Procedural Sound Synthesis," presented at the *144th Convention of the Audio Engineering Society* (2018 May), convention paper 416.

[32] P. Adenot and C. Wilson, "Web Audio API," https://www.w3.org/TR/webaudio/#introduction (2015).

[33] B. Taylor, J. T. Allison, W. Conlin, Y. Oh, and D. Holmes, "*Simplified Expressive Mobile Development With NexusUI, NexusUp, and NexusDrop*," *New Int. Music. Expr. (NIME)*, pp. 257–262 (2014).

[34] N. Jillings, Y. Wang, R. Stables and J. D. Reiss, "Intelligent Audio Plugin Framework for the Web Audio API," presented at the *Web Audio Conference* (2017).

[35] N. Jillings, B. De Man, D. Moffat and J. D. Reiss, "*Web Audio Evaluation Tool: A Browser-Based Listening Test Environment*," presented at the *Sound and Music Computing (SMC)* (2015).

[36] A. Cap, "applause small room.wav," https://freesound.org/people/ascap/sounds/242581/ (2014).

[37] B. De Man and J. D. Reiss, "APE: Audio Perceptual Evaluation Toolbox for MATLAB," presented at the *136th Convention of the Audio Engineering Society* (2014 Apr.), convention paper 151.

[38] X. Amatriain, J. Bonada, A. Loscos and X. Serra "Spectral Processing," in U. Zölzer (Ed.), *DAFx: Digital Audio Effects*, pp. 373–438 (John Wiley and Sons, Ltd., Chichester, UK, 2002), https://doi.org/10.1002/047085863X.ch10.

[39] S. O'Leary and A. Robel, "A Montage Approach to Sound Texture Synthesis," *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP), vol.* 24, no. 6, pp. 1094–1105 (2016), https://doi.org/10.1109/TASLP.2016.2536481.

[40] J. H. McDermott and E. P. Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," *Neuron,* vol. 71, no. 5, pp. 926–940 (2011), https://doi.org/10.1016/j.neuron.2011.06.032.

[41] J. -A. Adrian, T. Gerkmann, S. van de Par, and J. Bitzer "Synthesis of Perceptually Plausible Multichannel Noise Signals Controlled by Real World Statistical Noise Properties," *J. Audio Eng. Soc., vol. 65*, no. 11, pp. 914–928 (2017), https://doi.org/10.17743/jaes.2017.0035.

[42] M. Fröjd and A. Horner, "Sound Texture Synthesis Using an Overlap–Add/Granular Synthesis Approach," *J. Audio Eng. Soc., vol. 57*, no. 1/2, pp. 29–37 (2009).

[43] D. Moffat and J. D. Reiss, "Objective Evaluations of Synthesised Environmental Sounds," presented at the *Digital Audio Effects (DAFx)* (2018).

THE AUTHORS

Jake Ryan Rajjayabun Lee

Joshua D. Reiss

Jake Lee graduated from Queen Mary University of London with a first class honors degree in Electronic Engineering with a specialty in Music and Audio Systems. This research paper is based mainly on his final year project and brought together three areas of fascination for Jake: audio synthetics, sound design modeling, and programming audio technology. Jake continued to work on the project after graduating while also embarking on a role as a research assistant on FXive, a web-based procedural audio project, and a backend Java developer working for Wunderman Thompson Commerce providing e-commerce solutions for major retailers.

•

Josh Reiss is Professor of Audio Engineering with the Centre for Digital Music at Queen Mary University of London. He has published more than 200 scientific papers (including over 50 in premier journals and 5 best paper awards) and co-authored two books. His research has been featured in dozens of original articles and interviews on TV, radio, and in the press. He is a Fellow and former Governor of the Audio Engineering Society (AES), chair of their Publications Policy Committee, and co-chair of the Technical Committee on High-resolution Audio. He co-founded the highly successful spin-out company, LandR, and recently formed a second start-up, FXive. He maintains a popular blog, YouTube channel, and Twitter feed for scientific education and dissemination of research activities.