SUBJECTIVE EVALUATION OF SOUND QUALITY AND CONTROL OF DRUM SYNTHESIS WITH STYLEWAVEGAN

Antoine Lavault, Axel Roebel, Matthieu Voiry

SUBJECTIVE EVALUATION OF SOUND QUALITY AND CONTROL OF DRUM SYNTHESIS WITH STYLEWAVEGAN

Antoine Lavault
Apeira Technologies
Le Creusot, France
a.lavault@
apeira-technologies.fr

Axel Roebel
UMR 9912 STMS
IRCAM, Sorbonne Université
Paris, France
roebel@ircam.fr

Matthieu Voiry
Apeira Technologies
Le Creusot, France
m.voiry@
apeira-technologies.fr

ABSTRACT

In this paper we investigate into perceptual properties of StyleWaveGAN, a drum synthesis method proposed in a previous publication. For both, the sound quality as well as the control precision StyleWaveGAN has been shown to deliver state of the art performance for quantitative metrics (FAD and MSE of the control parameters). The present paper aims to provide insight into the perceptual relevance of these results. Accordingly, we performed a subjective evaluation of the sound quality as well as a subjective evaluation of the precision of the control using timbre descriptors from the AudioCommons toolbox. We evaluate the sound quality with mean opinion score and make measurements of psychophysical response to the variations of the control. By means of the perceptual tests, we demonstrate that StyleWaveGAN produces better sound quality than state-of-the-art model DrumGAN and that the mean control error is lower than the absolute threshold of perception at every point of measurement used in the experiment.

1. INTRODUCTION

In the 1980s the first drum machines and drum synthesizers using analogue and digital synthesis techniques appeared. While these drum machines are still used nowadays for their unique sonic fingerprint, they did not provide an extensive set of controls over said synthesis. Recently, triggered by the successful application of deep neural networks to other signal generation tasks, several data driven drum sound synthesis models have been proposed. A drum generator based on an U-Net architecture has been proposed in [1]. That generator learns a deterministic mapping from perceptual features. Subsequently many drum synthesis models based on generative adversarial networks (GAN) [2] have been proposed. GAN operating in the waveform domain have been studied in [3, 4], while [5] has proposed a GAN operating in the frequency domain generating real and imaginary part of the short time Fourier transform (STFT) of the generated sound. Another method working in the spectral domain using a variational auto encoder (VAE) has been presented in [6]. The authors report some blurring in the generated spectra. The Controllable Raw Audio Synthesis with High-resolution (CRASH) proposed in [7] is a score based generative model that supports a large variety of applications (class conditional synthesis, inpainting, interpolation) unfortunately suffers from rather long inference times.

A distinguishing feature of the drum sound generators are the possible controls. Most of the methods presented above [6, 4, 5, 7] support conditioning the generator on the drum class. Another musically interesting feature is the control by means of perceptual features. [1, 5] proposed to use timbral features from the AudioCommons project [8].

Our first contribution was in [9] where our goal was to create an algorithm for drum sound synthesis suitable for professional music production. The system should provide high sound quality, real-time generation and musically relevant controls. As a first result we presented drum synthesis with StyleWaveGAN [9], a GAN based synthesis model adapted from StyleGAN [10, 11] for the task of drum synthesis with control using differentiable perceptual features.

The present paper extends the results presented in [9] with the following contributions. First, while [9] used a quantitative measure, the Frechet Audio Distance (FAD) [12], to evaluate the quality and diversity of the generated samples. The present investigation extends the results obtained in [9] by means of a subjective evaluation of the sounds generated with StyleWaveGAN. Second, with respect to the perceptual controls, we present a refined discussion of the choice of descriptors that have been selected, provide an extended discussion about the implementation of the differentiable feature estimators, and provide a subjective study that allows characterizing the perceptual relevance of the remaining errors in the perceptual features of the generated sounds.

The reminder of the paper is organized as follows: Section 2 motivates the control strategies that were elected for StyleWaveGAN and briefly summarizes the essential elements of the model architecture. Section 3 describes the training data set, training parameters and the subjective evaluations, section 4 discusses the results of the perceptual tests, and section 5 summarizes the results and gives an outlook on further research.

Table 1: *Comparison of state of the art neural drum synthesizers*

Reference	Sample Rate	Duration
WaveGAN [3]	16kHz	1.1s
NeuroDrum [1]	16kHz	1s
DrumGAN [5]	16kHz	1.1s
Neural Drum Machine.[6]	22.05kHz	1s
Drysdale et al.[4]	44.1kHz	0.4s
CRASH [7]	44.1kHz	0.5s
Ours	**44.1kHz**	**1.5 s**

2. MODEL AND MOTIVATIONS

As introduction into the discussion, the present section will provide the necessary context introducing the StyleWaveGAN and its control strategies presented in [9]. Because an important focus of the present paper is the perceptual evaluation of the control the following section puts more emphasis on the control strategies and remains concise on the other details of the model and its training.

More precisely, the section 2.1 provides a significantly extended motivation of the control strategy (including controllable features and implementation) of the StyleWaveGAN. Section 2.2 extends the description of the differentiable timbre descriptors with a discussion of the modifications required to implement differentiable versions of the timbre descriptors. The following sections summarize the basic ideas of th StyleWaveGAN model introduced in [9], and in section 2.6 we explain the objective evaluation of the control strategy produced in [9] and motivate the perceptual evaluation of the control error that is proposed in the present study.

2.1. Selection of synthesizer controls

Since StyleWaveGAN is intended for use in professional audio production, we wanted its control scheme to remain intuitive. Accordingly we selected two levels of control.

The most fundamental control is the drum class. Selection of the drum class is a classical control for drum synthesizers and seems required for any professional use of the model for music production. Accordingly this control is available in most of the neural drum synthesizers [6, 4, 5, 7]. In our experiments, we are using 5 labels (kick, snare, tom, closed hi-hat and open hi-hat). We chose these labels as they represent the most common drum and cymbal types in modern pop and rock genre.

Second there is a perceptual control of the sound generated for each instrument. This kind of control is not provided by classical drum synthesizers and is therefore a key distinguishing feature of neural drum synthesizers.

A common choice in the literature [1, 5] for adding perceptual controls to drum synthesis are the timbral descriptors from the AudioCommons project [8]. The high-level nature of these descriptors makes them extremely powerful as they will cover a lot of possibilities while limiting the number of varying parameters. Such descriptors allow for intuitive control hence keeping a user in a state of creative flow. The full set of timbral models available in the AudioCommons toolbox are listed in Table 2. These descriptors were specially crafted from the study of popular timbre designations given to a collection of sounds from the Freesound data set. The perceptual models were built by combining existing low-level features found in the literature [13][14], which correlate with the chosen timbral designation.

The StyleWaveGAN uses only 3 out of the 8 descriptors found in Table 2. These descriptors were selected following an informal survey among drummers and musicians. The survey consisted of the following question "Among the following features (cf. Table 2), which one would you like to have in a drum synthesizer ?". Brightness and warmth were deemed important as they represent opposite ends of the frequency spectrum as well as being common terms among the music production jargon. Depth was of interest since it allowed for temporal manipulation of the lower frequencies and especially their decay. While the other AudioCommons features are also of interest, we focused on these three as they represent the preferred choices of potential future users.

Table 2: *Summary of AudioCommons models, the descriptors marked in bold are those selected for controlling the synthesis.*

Model	Underlying processing methods
Booming	Loudness + RMS
Brightness	Spectral centroid
Depth	Centroid + envelope
Hardness	Onset + Bark + Centroid
Reverb	Decay estimation
Roughness	Peak detection
Sharpness	Loudness
Warmth	Envelope + centroid

2.2. Differentiable AudioCommons Timbre Models

A first, rather straightforward, approach for learning perceptual controls was proposed in [1]. The model being a deterministic generator conditioned besides others on the perceptual controls these controls are learned as part of the reconstruction error. Later, using a GAN, [5] proposed using learned estimators of the respective descriptors to add the control error as an objective to the overall loss. For StyleWaveGAN [9] we have proposed to re-implement the three selected descriptors in Table 2 in order to make them fit directly into the training process as differentiable functions. The main motivation here is the fact that learning a differentiable proxy by means of training a neural network cannot guarantee the correct evaluation of the features to the same degree than implementing the features following the reference implementation. This is especially important for signals that do not have the same signal properties (e.g. the balance between resonances and noise) as the target signals and will therefore require very strong generalization of the estimator. Such signals will necessarily appear in the early states of the training process, and wrong evaluation may lead to undesirable local minima. Extending controls to values outside of the range of values that were available for training the proxy will pose problems as well.

In order to make the three timbral descriptors differentiable, some adjustments and modifications had to be done. Most notably, we approximated the time domain representation of the IIR filters used in the AudioCommons toolbox by means of a spectral domain representation. This approximation facilitated the automatic differentiation in Tensorflow. Even with these modifications, the difference between the original and the reimplemented version remained negligible (less than 1% of mean absolute difference on ENST-Drums [15]). The differentiable descriptors provided in [9] can be found at `https://github.com/ALavault/tf_timbral_models`.

2.3. Generative Adversarial Networks and StyleGAN

Generative Adversarial Networks (GAN) are a family of models consisting of two competing networks [2]. These networks are called generator and discriminator respectively. The goal of the discriminator is to distinguish whether a sample at its input is from the training data set or not while the generator aims to generate samples deemed real by the discriminator while having a random latent vector as its input.

Instead of using the original GAN structure, we used an evolution called StyleGAN [10, 11]. StyleGAN attempts to mitigate

the entangled representation when using noise as latent vector and input of the generator. The key idea here is to use a *style encoding*, a vector which is obtained through a mapping network and is then used to control (through an affine transform) every layer of a synthesis network.

2.4. Proposed architecture

Since StyleGAN was originally used for image generation, modifications for direct waveform generation were required. This meant flattening 2D convolutions and adapting the upsampling method, as well as changing minor parts of the training of StyleGAN.

[9] uses the same number of filters, with respect to the depth as StyleGAN2[11]. Just like StyleGAN2, the synthesis network uses input/output skips and the discriminator is a residual network.

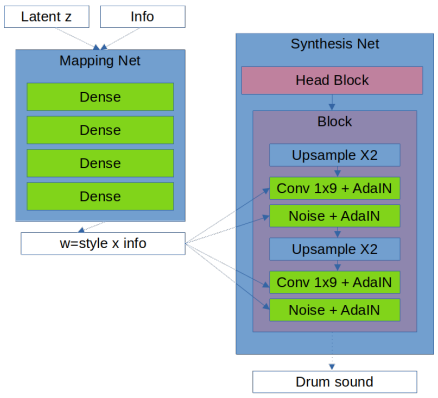

Figure 1: *StyleWaveGAN*

By using a temporal representation, we were following the choice by others like [1, 4, 7]. Working with a temporal representation was backed up by informal perceptual evaluations at the start of the study of StyleWaveGAN. Participants described the temporal representation as producing better audio quality than the spectral counterpart.

2.5. Controlling the network

There are two places where the control information is fed into the network: The first is located before the mapping network where the latent vector z (c.f Figure 1) is concatenated with an embedding of the instrument labels. This augmented latent vector is then fed into the mapping network. The second is located after the mapping network where the output of the mapping network is concatenated with the labels and the target descriptor values. These labels are encoded in form of a one-hot vector and the descriptor values, when used, are encoded as floating point numbers.

To ensure the network uses the information provided by the target descriptor values, we added the control error calculated from the difference between target descriptors values and the descriptor values computed on the the generated sounds to the loss used to optimize the generator. Extending the discussion in [9], section 4.2 will provide arguments for the choice of the L1 norm versus the L2 norm.

2.6. Evaluation of the control

We are reusing the mean absolute error metric introduced in [9] in this study, to compare to the results of our perceptual experiments. This metric uses the Mean Absolute Error (MAE) between the target values and the output values on three regions based on quantiles of the data set values:

- F1: MAE evaluated using only the target descriptor values within the 20th and 50th quantiles
- F2: MAE evaluated using only the target descriptor values within the 50th and 80th quantiles
- F3: MAE evaluated using only the target descriptor values within the 20th and 80th quantiles

We will be using the F1, F2 and F3 values from [9] and compare them to the results of subjective evaluation of the control in the later sections.

Given the very low control errors measured in [9] the question arises whether the remaining error is perceptually relevant. The present paper proposes a perceptual evaluation of this question by means of measuring absolute thresholds of perception of differences for selected reference descriptor values. The measurements and conclusions for this characteristic are presented in the later sections. This psychophysical measurement can also be useful in an attempt to further simplify the synthesis control by providing a discrete set of values instead of a continuum: there is no reason to allow more control if its not perceivable.

3. EXPERIMENTAL SETUP

3.1. Data sets

In the present study we will deal with the following data sets.

ENST-OG: Subset of ENST Drums containing all closed miked drums and hi-hat. For the present evaluation this data set represents real drum sounds. It is used in our perceptual evaluation of sound quality. This data set has all its elements with 44.1kHz sample rate of varying length but greater than 1 second.

ENST-AUG: Augmented version of ENST-OG, as described in [9][1]. It is used as training data set for StyleWaveGAN. This data set has the same sample rate as ENST-OG and length of the sounds are unchanged.

ENST-AUG-EX: Subset of ENST-AUG containing only extreme examples of augmentation. It is used in the evaluation of sound quality to evaluate the perceptual coherence of the augmented training data (ENST-AUG) compared to the original drum samples in (ENST-OG). Same sample rate and duration since it is a subset of ENST-AUG.

SWG-SQ: Samples generated with StyleWaveGAN trained on ENST-AUG without descriptors and with random latent for all labels in ENST-AUG. This set is only used in the sound quality evaluation. These sounds have a duration of 1.5s at 44.1kHz sample rate.

DG-SQ: Samples for kick, snare and cymbals labels provided by Javier Nistal generated with a version of DrumGAN providing drum type conditioning according to [16]. These were

[1] Note that the labels are preserved by the augmentation process.

trained on the private data set used as well in [5]. This set is used in the sound quality evaluation as the state-of-the-art reference. The elements from this data set have a duration of 1.1s and have a sample rate of 16kHz.

SWG-CQ: Snare samples generated with StyleWaveGAN trained on ENST-AUG using descriptor controls. This data set is used for the control quality evaluation. Only snare samples are used to remain consistent with the objective evaluation performed in [9]. The sounds from this data set have the same sample rate and duration as SWG-SQ.

All StyleWaveGAN models used for generating the data sets have been trained exactly as described in [9].

3.2. Reference-free baseline

Comparison to baseline models with the reference-free Frechet Audio Distance [12] has already been done in [9]. The comparison was between StyleWaveGAN, NeuroDrum [1] and WaveGAN [3].

3.3. Perceptual Evaluations

3.3.1. Subjective evaluation of sound quality

In our subjective evaluation framework, we are evaluating the quality of generation among 4 sets of sounds : ENST-OG, ENST-AUG-EX, SWG-SQ and DG-SQ. The comparison of these 4 sets is motivated as follows: ENST-OG is the real world reference, ENST-AUG-EX contains extreme examples used for training and can therefore be seen as a lower bound for prefect training of the model, the evaluation on DG-SQ establishes a state-of-the-art baseline, which has been compared to other approaches in the literature.

First, we will justify our use of DrumGAN (DS-SQ) as the state-of-the-art baseline. In [9] we already have shown that StyleWaveGAN significantly improves the FAD objective measure compared to NeuroDrum and WaveGAN, so for the present investigation we aimed to chose another method. We have considered three other state-of-the-art methods, which are DrumGAN [5], Drysdale et al. [4] and CRASH [7]. [5] and [4] both use GANs and are much closer in technology to StyleWaveGAN than [7], which is based on a completely different approach and requires significantly longer inference times. Accordingly we considered comparison with either [5] and [4] most interesting. A problem here are the varying means of conditioning used in the different methods. While [4] only provides conditioning with drum type, DrumGAN presented in [5] only has perceptual feature conditioning, a later version [16] uses drum type conditioning instead of perceptual feature conditioning. Note that the generation capacity of these models differ considerably. DrumGAN can generate samples of 1.1 s at 16kHZ, while [4] and [7] generate samples of 0.4s and 0.5s at 44.1kHz respectively. We will see in the following discussion that the test participants indicate that the decay time is important to evaluate the realness of drum sounds, which gives an advantage to DrumGAN with its longer samples even if the sample rate is lower.

Note that due to lack of available source code and meta-parameters none of these methods can be faithfully reproduced. Therefore we have to rely on comparing with results produced by the original authors using their respective training sets. We know that Drysdale et al. focus on sample-based electronic music (EM). Samples used in EM are inherently synthetic and are built to sound different from real drums. Since our subjective testing aims to evaluate how close the synthesized sounds are to a real drum, having synthetic samples in the training set will always be evaluated as worse. Given the limitations discussed before we selected DrumGAN [16] as our baseline representing one of the models of the state of the art for our perceptual test as it is produces the longest samples and uses a similar training method to ours.

3.3.2. Subjective evaluation for control quality evaluation

We use one the psychophysical methods described in [17] to measure the absolute threshold. Our experiment uses the constant stimulus method. This means two sounds are played one after the other, one being the reference generated with descriptor value v obtained from the data set and the other one being generated with a descriptor value $v + \Delta$. The test subjects are then asked if the stimuli was perceived as identical or different. Following [17] we determine the absolute threshold as the value of the comparison stimulus judged clearly different than the standard 50% of the times.

We use samples from the SWG-SQ data set for this task. The descriptor values v are computed to match the 20th, 50th and 80th quantiles of the descriptor of interest. The offset Δ is selected from the set

$$D = \{x | x = 0.25z \text{ with } z \in \mathbb{Z} \text{ and } -8 \leq x \leq 8\}. \quad (1)$$

Note that the AudioCommons descriptor values are normalized and range from 0 to 100 so that the variation used in the test always covers $\pm 8\%$ of the total range of the descriptor. We chose this range of variation since the MAE results in [9] and reproduced in table 6 are always lower than 4, which makes the chosen range of variation the double of a global upper bound of the MAE metrics.

The order in which the two samples are played is important and both orderings are used in the test. Since samples are chosen randomly, this ensures that all orderings are equally present. A fade-out is applied to each of these samples to avoid any noisy tails having an impact on the evaluation.

Since subjective equality measures are only valid for a single stimulus intensity, having multiple target values will allow us to extrapolate subjective equality points over a wider range of values. We choose to use the measurement points corresponding to the same points that were used for the MAE metric described in 2.6 and by doing so, we hope to be able to compare the subjective equality points to the MAE and draw conclusions about whether or not the mean error is perceived.

Table 3 summarizes the content of the data set SWG-CQ used in this experiment.

Table 3: *Summary of absolute threshold experiments*

Descriptor	Ordering	Steps	Total
Brightness	2	65	130
Depth	2	65	130
Warmth	2	65	130

3.3.3. Listeners and test conditions

There are two sets of listeners for the two different experiments.

Both of the listening tests took place remotely. For the perceptual quality evaluation, all test participants were presented with 24 samples to rate on a 5-point scale. The five levels of the scale are "Bad", "Poor", "Fair", "Good" and "Excellent" and are represented with values of 1,2,3,4 and 5 respectively. These samples were randomly picked among the ENST-OG, ENST-AUG-EX, SWG-SQ and DG-SQ data sets. The mean opinion score (MOS) is calculated as the average of the score given by the test participants. Part of the samples generated from SGW-SQ are available in the supplementary material. Nine participants took part in this test. Even if the number of participant is low, most of them (5 out 9) are audio professionals. As far as listening equipment goes, the different participants either professional or not used their own listening devices in the form of studio speakers or headphones.

We can now describe the experiment on control quality. In this experiment, all test participants was presented with 32 samples taken at random among the generated pairs from SWG-CQ. We gathered the answers of 31 subjects for this perceptual test. In terms of listening devices for the test, participants were ask to use their own listening devices which were either studio speakers, headphones or earbuds. We asked the participants to find a calm place where they could take the test in one sitting in order to not only avoid noisy data in our experiment due to a bad listening environment but also to get the most consistent listening experience possible.

For both perceptual tests, the details about listening equipment and age groups are listed in table 4

Table 4: *Listening devices and age groups for the perceptual tests*

Experiment / Listening device	Sound Quality	Control Quality
Studio Speakers	2	2
Headphones	7	27
Earbuds	0	2

Experiment / Age Group	Sound Quality	Control Quality
Age 0-17	0	0
Age 18-25	1	13
Age 26-40	4	9
Age 41-65	4	9
Age 66+	0	0

4. EXPERIMENTAL RESULTS

4.1. Results of the subjective evaluation of sound quality

The total Mean Opinion Score (MOS), with their confidence interval at 95%, is shown in Table 5 as well as more detailed per-class results.

The score of the augmented samples is slightly lower than the real samples. This indicates that the extreme cases of our augmentation strategy are a bit too extreme. Less extreme augmentation parameters with more intermediate values should be selected for future work. The main problem that can be found against the augmented samples comes from the pitch changes made by the augmentation process. The pitch change affects negatively the attack of the sound, making them sounding less natural than the real data. While the change is minor, it is sufficiently present to be perceived and graded worse than a real sample. Note however that these extreme parameter combinations remain rather rare in the full set of augmented sounds.

Table 5: *MOS on different data sets depending on the instrument label (1 is lowest, 5 is highest)*

Data set \MOS	All Labels	Cymbals	Kick	Snare
ENST-OG	4.2 ± 0.3	4.1 ± 1.1	4.1 ± 0.6	4.4 ± 0.3
ENST-AUG-EX	3.8 ± 0.5	3.3 ± 1.3	4.0 ± 0.5	3.9 ± 0.5
SWG-SQ	3.5 ± 0.4	3.9 ± 0.7	3.0 ± 0.7	3.6 ± 0.8
DG-SQ	2.3 ± 0.5	2.3 ± 1.3	2.8 ± 0.6	1.6 ± 0.8

Comparing StyleWaveGAN to DrumGAN we can conclude that StyleWaveGAN trained on augmented data produces results that are perceived either similarly close (kick) or significantly closer (snare, cymbals) to real drums than DrumGAN trained on a drum data set obtained from sources that are not further detailed in [5]. We conclude that the StyleWaveGAN model trained on augmented data achieves at least state of the art performance for drum synthesis.

We now discuss the StyleWaveGAN results in details. Given StyleWaveGAN was trained on the full data set of augmented samples, a perfect model should produce results in between the test results of the real data and the extreme examples of the augmented data. We note that StyleWaveGAN achieves this performance only for the cymbals. Snare and kick synthesis remain less natural. A discussion with the participants of the perceptual tests who were audio professionals reveals the following problems: for the kick drum sounds the SWG model does not produce the characteristics long tail of the resonances and is also missing some energy in the frequency band below 100Hz.

For snare drum synthesis the main problem appears to be the fact that SWG creates hybrids of sounds generated with sticks, mallets and brushes. Concatenating for example an attack of a snare sound obtained with a stick with a decay of a snare sound obtained with a brush creates fair sounding but unrealistic samples. These problems with kick and snare sounds indicate that the current implementation of the discriminator is not sufficient and further investigation will be required to improve the discriminator loss such that it avoids these perceptual problems.

4.2. Impact of control loss on control precision

In Table 6, the lines labeled *3D descriptors* show the results when the descriptor of interest is set but the others are taken from a real example from the training data set. Finally, the lines labeled (*3D descriptors, data set*) show the results when all the descriptors values are taken from the training data set. Work regarding the differences between 1D descriptors (i.e one descriptor per model) and 3D descriptors has already been done in [9]. We will only focus on 3D descriptors control here.

The L2 loss performs better than the L1 loss when using brightness and depth with values from the data set. Results with warmth are way worse when using the L2 loss in both cases. The problem with the warmth descriptor and L2 norm can be explained by an offset in the control. This can be seen on Figure 3 when compared to Figure 2 which is showing the results with L1 loss. Red lines show the limit values of the data set and the green vertical lines show the position of the 20th, 50th and 80th quantiles.

Table 6: *Mean absolute error for several configurations with L1 loss, reproduced from [9](lower is better)*

Features	F1	F2	F3
Brightness (3D descriptors)	0.97	1.36	1.17
Depth (3D desc.)	1.33	1.50	1.41
Warmth (3D desc.)	1.29	3.31	2.33
Brightness (data set, 3D desc.)	0.75	0.95	0.85
Depth (data set, 3D desc.)	0.99	1.03	1.0
Warmth (data set, 3D desc.)	1.42	1.37	1.39

Table 7: Mean absolute error for several configurations with L2 loss (lower is better)

Features	F1	F2	F3
Brightness (3D descriptors)	1.16	1.73	1.45
Depth (3D desc.)	1.21	1.29	1.26
Warmth (3D desc.)	4.96	2.49	3.69
Brightness (data set, 3D desc.)	0.66	0.85	0.76
Depth (data set, 3D desc.)	0.77	0.54	0.65
Warmth (data set, 3D desc.)	6.92	6.28	6.60

Please note that Figure 2 is not reproduced from [9] as it shows results when the control values (3 descriptors here) are drawn from the training data set, when the figures in [9] show the effect when the target control values cover the full range of descriptor values and the others are drawn from the training data set.

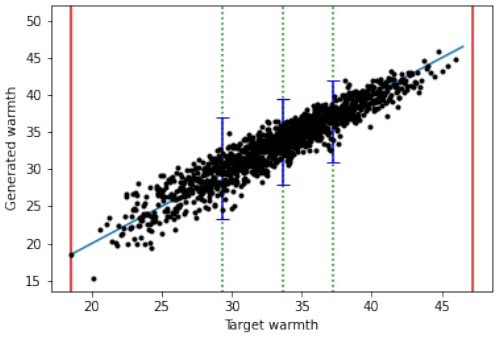

Figure 2: *Generated values for warmth when synthesizing snare drum sounds with control values obtained from the Snare drum sounds in ENST-AUG using StyleWaveGAN trained with L1 loss for the control error and 3D descriptors.*

The offset observed in this experiment when evaluating the model trained with 3D perceptual controls using L2 loss remains unclear. Further studies need to be performed to see whether this is due to a local minimum and may be solved by means retraining with different weight initialization, or whether this shows a systematic problem with the loss weighting that should be solved by means of increasing the weight of the warmth descriptor error in the loss function. Unfortunately training these models takes more than a week and systematic evaluation of many different training runs remains challenging.

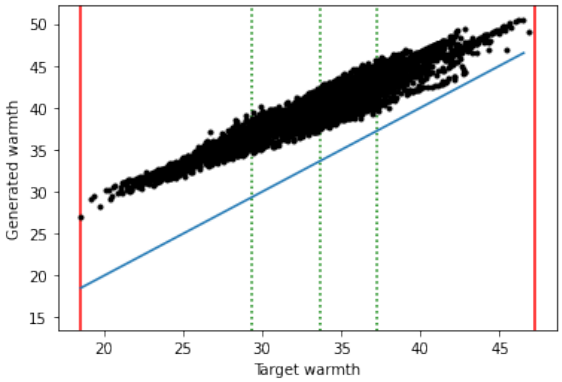

Figure 3: *Generated values for warmth when synthesizing snare drum sounds with control values obtained from the Snare drum sounds in ENST-AUG using StyleWaveGAN trained with L2 loss for the control error and 3D descriptors.*

This confirms our first choice from [9] where we used the L1 loss only.

4.3. Results of subjective evaluation of control error perception

To interpret our results from the subjective evaluation of control error, we need to study different thresholds on the estimated probability distribution of the data. The simplest to interpret is the 50% threshold, also known as absolute threshold : this means the test subjects are randomly choosing if the difference is perceptible or not. We use the absolute threshold on the experiments comprising of only the positive variations (i.e the second sample has a higher descriptor value than the first), the negative variations and the combined (where we take the absolute value of the variation to make our statistics). The results are shown in Table 8 and 9. Figures 4,5 and 6 show the fitted sigmoid alongside the estimated cumulative distribution function (CDF) from three different cases. Figures for all different cases will be available in the supplementary material. The estimated cumulative density function is computed by gathering values in bins spaced by 1 unit and then the cumulated sum on these bins. Dashed black lines on figures 4,5 and 6 represent the 25%, 50% and 75% thresholds estimated with the fitted function.

We calculate an estimated CDF for the probability of hearing a difference in the positive, negative and combined experiment. We then fit a sigmoid function to the data. In other words, we choose the sigmoid function as our psychophysical function. In other words, we try to fit the function ψ such that :

$$\psi(x, b, d) = \frac{1}{1 + \exp(-bx + d)} \tag{2}$$

where b and d are real parameters, estimated with a least-square estimator.

Our main goal with this experiment was to measure if the error between the control and the output in our network is perceptible or not. Our MAE metric measures the error on a segment. We chose the measurement points for the psychophysical experiment to be the endpoints of the segments of our MAE metric to be able to

Table 8: *Summary of absolute threshold experiments for each 3 points of measurement.*

Descriptor	Q20	Q50	Q80
Brightness (combined)	5.00	5.00	6.00
Brightness (positive)	5.00	5.00	6.00
Brightness (negative)	5.00	5.00	6.00
Depth (combined)	5.00	5.00	5.00
Depth (positive)	5.00	4.00	5.00
Depth (negative)	4.00	5.00	4.00
Warmth (combined)	6.00	6.00	6.00
Warmth (positive)	6.00	6.00	7.00
Warmth (negative)	8.00	6.00	4.00

Table 9: *Summary of absolute threshold experiments for each 3 points of measurement with fitted sigmoid*

Descriptor	Q20	Q50	Q80
Brightness (combined)	5.39	5.19	5.75
Brightness (positive)	5.43	5.11	5.83
Brightness (negative)	5.35	5.27	5.75
Depth (combined)	5.11	4.95	4.79
Depth (positive)	5.39	4.23	4.91
Depth (negative)	4.63	5.19	5.19
Warmth (combined)	6.27	5.67	5.35
Warmth (positive)	6.15	5.75	6.27
Warmth (negative)	7.68	5.83	4.75

make some kind of conclusion. For further illustration, we added blue error bars on Figure 2 at the different points of measurements. The error bars show the points of subjective equality, for positive and negative variations. This figure shows the generated samples output values are well within the limits of the points of subjective equality on the segments of interest.

We claim that the error is imperceptible on the segments used for the MAE metric (F1, F2, F3) if and only if the MAE on these segment is lower than the minimum of absolute threshold measurements at the segment endpoints. This condition can be simplified to the MAE being lower than 4 since the points of subjective equality are always higher than 4 in our experiment. This condition is found true on every single segment, wether using the CDF or the fitted sigmoid.

From such a strong result, we can make some conclusions. First, the error should not be noticeable in almost all cases. We also found that the control is not necessarily perceived as symmetric : the same variation but with a different sign could lead to points of subjective equality. Finally, the good results from our perceptual experiment show that a discrete control could work: a step of

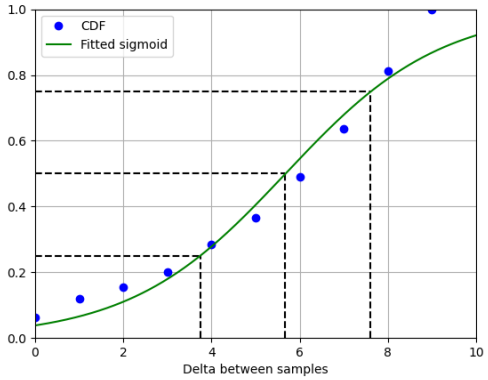

Figure 4: *Psychometric results for brightness for the 80th quantile and combined deltas*

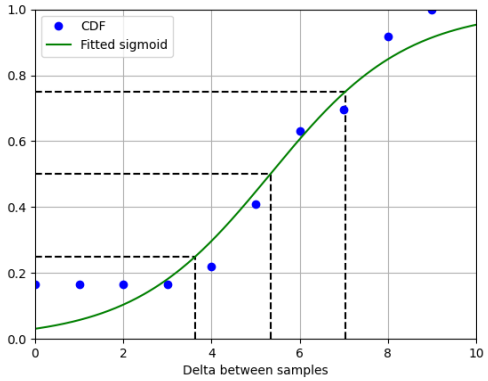

Figure 5: *Psychometric results for depth for the 20th quantile and positive deltas only*

1 unit should not be noticeable in almost all cases and would allow for less fine-grained control, hence making easier to use in a professional audio production context.

5. CONCLUSION

In this paper, we presented a study on the subjective evaluation of the sound quality of the proposed StyleWaveGAN as well as a subjective evaluation of the precision of the control using timbre descriptors from the AudioCommons toolbox.

The perceptual evaluation of the sound quality has confirmed that SWG equates the generation quality of one of the state of the art drum synthesis methods known from the literature and outperforms it significantly for snare drum and cymbals. While the number of participants is low individual discussion with participants has clearly revealed the limitations of the model and the training data set and perspectives for improvements have been developed. A new perceptual study will be performed once the improvements will have been implemented. The perceptual evaluation of the quality of the control with our differentiable features on the snare drum has demonstrated that the mean control error at

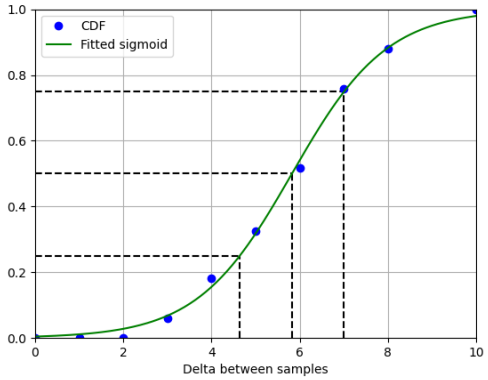

Figure 6: *Psychometric results for warmth for the 50th quantile and negative deltas only*

all the measurement points is consistently lower than the absolute threshold of perception, which leads us to conjecture that control error is perceptually negligible. Note that the perceptual evaluation of the control quality has not been covered in the literature. Our internal investigation with the other drum instruments results in control errors of the same scale which leads us to assume the same conclusion holds for all types of drum instruments. In terms of future work we will continue to work on the sound quality especially for the kick drum for that the perceptual evaluation gave the worst results of all labels that have been tested. We will also work on additional controls, notably regarding velocity.

6. ACKNOWLEDGMENTS

We would like to express our very great appreciation to Patrick Susini for his valuable and constructive suggestions during the planning and development of the psychophysics part of this research work. His willingness to give his time and knowledge so generously has been very much appreciated. We also would like to thank Javier Nistal for providing us with samples generated by DrumGAN.

7. REFERENCES

[1] Antonio Ramires, Pritish Chandna, Xavier Favory, Emilia Gomez, and Xavier Serra, "Neural Percussive Synthesis Parameterised by High-Level Timbral Features," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. nov 2020, vol. 2020-May, pp. 786–790, Institute of Electrical and Electronics Engineers Inc.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," pp. 1–9, 2014.

[3] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," in *ICLR*, 2019.

[4] Jake Drysdale, Maciej Tomczak, and Jason Hockman, "Adversarial synthesis of drum sounds," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx2020)*, 2020.

[5] Javier Nistal Hurle, Stefan Lattner, and Gael Richard, "DrumGAN: Synthesis of drum sounds with timbral feature conditioning using Generative Adversarial Networks," in *21 st International Society for Music Information Retrieval Conference (ISMIR)*, Aug. 2020.

[6] Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres, "Neural Drum Machine : An Interactive System for Real-time Synthesis of Drum Sounds," in *Proc. of the Int. Conference on Computational Creativity*, jul 2019.

[7] Simon Rouard and Gaëtan Hadjeres, "CRASH: raw audio score-based generative modeling for controllable high-resolution drum sound synthesis," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, 2021, pp. 579–585.

[8] Andy Pearce, Tim Brookes, and Russell Mason, "Hierarchical ontology of timbral semantic descriptors," *AudioCommons - Deliverable D5.1*, pp. 1–34, 2016.

[9] Antoine Lavault, Axel Roebel, and Matthieu Voiry, "StyleWaveGAN: Style-Based Synthesis of Drum Sounds with Extensive Controls using Generative Adversarial Network," in *Sound And Music Computing*, 2022, Accepted for publication, accessible under `https://arxiv.org/abs/2204.00907`.

[10] Tero Karras, Samuli Laine, and Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2018.

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and Improving the Image Quality of StyleGAN," dec 2019.

[12] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 2350–2354, 2019.

[13] Geoffroy Peeters, "A Large Set of Audio Features for Sound Description," Tech. Rep. 0, 2004.

[14] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[15] Olivier Gillet and Gaël Richard, "ENST-Drums: An extensive audio-visual database for drum signals processing," *ISMIR 2006 - 7th International Conference on Music Information Retrieval*, pp. 156–159, 2006.

[16] Javier Nistal, *Exploring generative adversarial networks for controllable musical audio synthesis*, Ph.D. thesis, Institut Polytechnique de Paris, 2022.

[17] Bruno L Giordano, Patrick Susini, and Roberto Bresin, "Experimental methods for the perceptual evaluation of sound-producing objects and interfaces," in *Sonic Interaction Design :*. MIT Press, 2012.