OPEN ACCESS Freely available online
M. Cheshire, J. Drysdale, S. Enderby, M. Tomczak and J. Hockman,
“Deep Audio Effects for Snare Drum Recording Transformations”
J. Audio Eng. Soc., vol. 70, no. 9, pp. 742–752, (2022 September).
DOI: https://doi.org/10.17743/jaes.2022.0041

Deep Audio Effects for Snare Drum Recording Transformations

MATTHEW CHESHIRE, JAKE DRYSDALE, SEAN ENDERBY,
(matthew.cheshire@bcu.ac.uk) (jake.drysdale@bcu.ac.uk) (sean.enderby@bcu.ac.uk)
MACIEJ TOMCZAK, AND JASON HOCKMAN
(maciej.tomczak@bcu.ac.uk) (jason.hockman@bcu.ac.uk)

Sound and Music Analysis Group (SoMA), Digital Media Technology Lab (DMT Lab), School of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom

The ability to perceptually modify drum recording parameters in a post-recording process would be of great benefit to engineers limited by time or equipment. In this work, a data-driven approach to post-recording modification of the dampening and microphone positioning parameters commonly associated with snare drum capture is proposed. The system consists of a deep encoder that analyzes audio input and predicts optimal parameters of one or more third-party audio effects, which are then used to process the audio and produce the desired transformed output audio. Furthermore, two novel audio effects are specifically developed to take advantage of the multiple parameter learning abilities of the system. Perceptual quality of transformations is assessed through a subjective listening test, and an object evaluation is used to measure system performance. Results demonstrate a capacity to emulate snare dampening; however, attempts were not successful for emulating microphone position changes.

0 INTRODUCTION

The positioning and recording of a standard acoustic drum kit—comprising of kick, snare, toms, and an assortment of hi-hats and other cymbals—is a technical and time-consuming endeavor. Recording drums may account for as much as 25% of the whole recording project [1]. During a typical session, an engineer must modify a large number of recording parameters to achieve a desired result. Key considerations include the selection of drums, drumheads, tuning, and dampening and the selection, arrangement, and positioning of microphones. These decisions impact the overall timbral quality of a recording, with certain modifications producing greater effects than others [2, 3].

Time permitting, an engineer may test different parameter options to identify an appropriate configuration for a song before committing to the final recording; however, with many variables, this can easily become a lengthy process. As such, the ability to perceptually modify these recording parameters in a post-recording process would be of great benefit to engineers limited by time or equipment, especially during sessions in which compromises may need to be made. In this work, a system is proposed for post-recording modification of the dampening and microphone positioning parameters associated with snare drum capture.

0.1 Background

Several methods for the automatic mixing of drums have been proposed [4–6]. Although these look at emulating processes of the digital mixing stage, the proposed system attempts to emulate techniques that are carried out prior to the recording stage. Two notable techniques an engineer can use to modify snare drum timbre include treating the drum heads directly through dampening or varying the position of the microphones around the drum in order to emphasize or subdue certain timbral characteristics.

Snare batter head dampening is a common timbre manipulation practice in drum recording [7, 8], which involves adding mass to the drumhead to remove unwanted overtones and shorten decay time to produce a perceptually tighter, more controlled sound [9, 10]. Engineers place various materials (e.g., cloth, duct tape, wallet) directly onto the drumhead to achieve subtle to extreme dampening effects. Many commercial products such as Big Fat Snare Drum, Snare Weight, and Moongel allow for the adjustment of dampening amount [11]. The recording engineer may use several of these techniques to create the intended drum sound [12]. Once dampening has been applied, those timbral properties are then committed to the recording, and one loses the ability to apply additional dampening if later required or to remove any if too much was used.

Microphone selection also impacts the timbre of recordings [13, 14]. The authors of [15] modified the spectral characteristics of a snare drum recording to mimic those of another through the use of a 30-band graphic equalizer (EQ); however, a limitation of this work was that access to recordings with target characteristics were required.

Audio effects are an integral part of the music production workflow that can be used to modify sound characteristics, such as dynamics, frequency, and timbre. Utilizing audio effects for a predefined audio transformation can be a laborious task that often requires mastery over a large number of parameters. As a result, there has been an increasing focus on audio effects modeling and intelligent audio effects within the field of music information retrieval.

In recent years, deep learning has demonstrated excellent performance in tasks such as emulating audio effects through end-to-end transformation methods [16–18], estimating audio effect parameters [19], mapping semantic descriptors to the parameter space of audio effects [20], and generating audio through differentiable digital signal processing [21]. More recently, Martinez et al. [22] emulated three common audio production tasks (i.e., mastering, breath/plosive removal, and tube amplification) through the use of a deep encoder, which performs parameterization of third-party audio effects within layers of the network.

0.2 Motivation

The system in [22] facilities training of audio plugin parameters or a chain of plugins for any desired transformation, given the appropriate training data. In this paper, the ability to modify the timbre of an undampened snare recording in order to elicit a perceptual change that corresponds to that of a dampened snare, referred to as Undampened-to-Dampened (U2D), will be explored through the use of multiple audio effects by utilizing the tools presented in [22]. The inverse transformation is also examined, whereby a dampened snare recording is modified to perceptually emulate qualities of an undampened snare recording, referred to as Dampened-to-Undampened (D2U). In addition to these dampening transformations, two positional recording parameter changes are explored: bottom-to-top (B2T) and top-to-bottom (T2B) microphone position.

The remainder of this paper is structured as follows: SEC. 1 outlines the proposed system. SEC. 2 describes the evaluation methodology for subjective objective comparisons. SEC. 3 presents the results from the evaluation, and SEC. 4 provides a discussion. Conclusions and suggestions for future work are presented in SEC. 5.

1 METHODOLOGY

An overview of the system configuration for transforming an undampened snare drum into a dampened snare is provided in Fig. 1. In order to automatically carry out different perceptual transformations, DeepAFx [22] is utilized for its powerful parameter learning and audio processing capabilities. DeepAFx consists of a deep encoder that first analyzes the input audio and then predicts the optimum parameters of one or more effect, which then processes the audio, producing the desired transformed output audio. The system makes use of the LV2 audio plugin open standard and incorporates third-party audio effects as a black box layer within a deep neural network. The authors provide the code used in their experiments.[1]

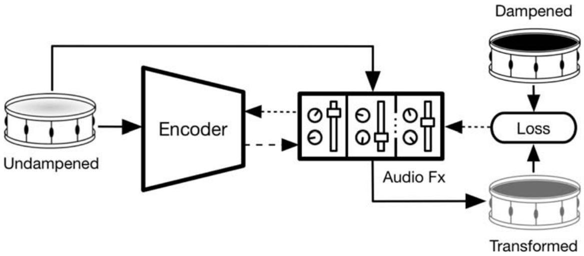

Fig. 1. System overview for snare dampening with DeepAFx with third-party audio effect. Solid lines depict flow of audio, the longer dashed line represents the predicted parameter values, and shorter dashed lines depict gradient flow.

1.1 Network Architecture

Following [22], an inception-based encoder network [23] is implemented to predict the audio effect parameter values required for a desired snare drum transformation. The input to the network is a log-scaled Mel-spectrogram represented as a 4D tensor $t \in \mathbb{R}^{b \times w \times h \times c}$, with batch size b, number of frames w, number of frequency bins h, and channels c. The model consists of 64 convolutional filters with a 5×5 sized kernel followed by 2×2 strided max-pooling. This is followed by six inception blocks with mixed kernel sizes, each comprised of a naive module with a stride of 2 and a dimension reduction module [24]. Rectified linear unit activations are used for all layers apart from the network's last layer, which is a fully connected output layer consisting of r output nodes and a sigmoid activation function, in which r is the number of parameters associated with a particular audio effect. The network outputs estimate audio effect parameter values for each snare drum transformation under observation.

1.2 Audio Effects

For this study, two novel LV2 audio effects are specifically developed to take advantage of DeepAFx's multiple parameter learning abilities; both effects have high parameter counts that would make it tedious and time-consuming for a human engineer to fine tune each control. Typically audio production tools are designed with the audio engineer in mind, graphic user interfaces (GUIs) are implemented, and variables such are parameter amount, layout, size, and color are considered in order to enhance the experience of the user. Allowing DeepAFx to learn the parameters a GUI is not required, nor are any considerations to the impracticality to a human user.

Both effects are investigated for their timbre-transforming abilities: a 10-band dynamic EQ (DEQ10) and

[1] https://github.com/SoMA-group/snarefx.

30-band dynamic EQ (DEQ30). Typically, dynamic EQs will consist of four to seven parametric frequency bands [25, 26], allowing the user to specify center frequency, Q-factor, and shelf or bell filter types [27, 28]. However, unlike traditional dynamic EQ, DEQ10 and DEQ30 are implemented as fixed-band graphic equalizers, with fixed center frequencies based on the specification for octave bands and fractional-octave bands described in [29], allowing for complete dynamic control over the full spectrum.

Dynamic EQ was specifically chosen in order to provide both spectral and temporal manipulation within one audio effect [30], often used in mastering applications [31]. The ability to control specified frequency bands over time lends itself to transformations in which some frequencies may be similar and others are disparate, such as in the case of dampening a snare, and in which both high frequencies are attenuated and their associated envelopes shortened, whereas the lower frequencies remain mostly unaffected. This would be difficult to achieve through the use of a standard full spectrum compressor; thus, dynamic EQ has the potential to perform better than a standard EQ and compressor combined for particular production tasks.

Both DEQ10 and DEQ30 have the same architecture, the signal path consisting of cascaded bi-quad peaking filters. Each frequency band comprises of two such filters; the gain of the first is controlled dynamically and that of the second is controlled through the *make-up gain* parameter for the band. Dynamic control of each band is achieved through a standard feed-forward compressor architecture. Within the side chain for each band, the signal first passes through a bi-quad band-pass filter, with center frequency and bandwidth matching that of the corresponding peaking filter in the signal path. Level detection and ballistics are carried out within the gain computer of the compressor's side chain. The output of this filter undergoes peak amplitude detection and then feeds a gain computer with the following parameters: *threshold*, *attack*, *release*, *ratio*, and *knee*. Each effect has an *output gain* parameter at the end of the signal path. A graphical representation of this architecture is given in Fig. 2. The principle difference between DEQ10 and DEQ30 is that the first uses an octave band layout, whereas the second uses third-octave increments. With six parameters per band and output gain, this gives 61 trainable parameters for DEQ10 and 181 for DEQ30.

In addition to the two novel effects, two open-source plugins were used.[2] Firstly an eight-band parametric EQ (PEQ), was chosen for its frequency sculpting ability and for the ubiquitous nature of parametric EQs in audio engineering. Secondly, because applying dampening to a snare drum alters its envelope characteristic, a transient designer (TD) was chosen as a possible candidate for a tool that might perform well at emulating this feature. A transient designer provides level-independent processing of the signal's envelope by using envelope followers to control output dynamics; this allows transients to be accelerated or slowed down and sustain to be prolonged or shortened [32].

DeepAFx also has the ability to train multiple plugins in a series; chaining multiple effects together is a common practice among mixing engineers [33], so for this reason, this aspect was also investigated. The PEQ and TD were used in conjunction with one another to determine whether they were able to perform better together, providing both spectral and temporal manipulations. The order of PEQ and TD were tested in both configurations, placing TD before and after PEQ. This was found to have very little audible difference on the processed audio; for this reason, only the PEQ+TD configuration was chosen for investigation.

1.3 Loss Function

The objective of the proposed model is to minimize the multi-scale spectrogram loss (MSL) between target snare drums and predicted snare transformations. MSL allows the network to extract information at multiple spectro-temporal resolutions and is calculated as the sum of the L2 difference between magnitude and log magnitude spectrograms computed with different fast Fourier transform resolutions: $r = \{2048, 1024, 512, 256, 128, 64\}$. The spectral loss for each resolution is defined as

$$\begin{aligned}\mathrm{MSL}_{\mathrm{stft}}(S, \hat{S}) = \sum_{r_i} \big[\| S_{r_i} - \hat{S}_{r_i} \|_2 \\ + \| \log S_{r_i} - \log \hat{S}_{r_i} \|_2 \big], \qquad (1)\end{aligned}$$

where magnitude spectrograms S and $\hat{S}$ are computed with a given fast Fourier transform resolution r_i from the target snare drums and predicted snare transformation audio.

1.4 Network Training

The deep encoder takes data x as input and parameters λ. Audio is pre-processed through resampling and conversion to a spectrogram representation. Following [22], snare drum recordings are resampled to 44.1 kHz, and the short-time Fourier transform (STFT) of each snare is calculated using a Hanning window with a size of 1,024 samples and a hop size of 256 samples to facilitate the desired temporal resolution of the network input. The magnitudes of STFT are transformed to log-scaled Mel-spectrograms with 128 Mel-frequency bands.

The model is trained using the Adam optimizer [34] with a learning rate $1e$–4, where each iteration takes a mini-batch of 100 examples. Network weights are initialized using He's constant [35] to promote equalized learning. Once model performance ceases to improve over 25 epochs, early stopping is applied to complete training, and the epoch that achieves the best accuracy on the validation set is used for testing. Training was carried out on a Nvidia TESLA M40.

2 EVALUATION METHODOLOGY

The system presented in Sec. 1 is assessed through two evaluations to determine 1) perceptual quality of the transformations through a subjective listening test and 2) similarity of the transformed audio compared to the target audio through an objective evaluation using various comparative metrics. For each type of transformation under investiga-

[2] http://calf-studio-gear.org/.

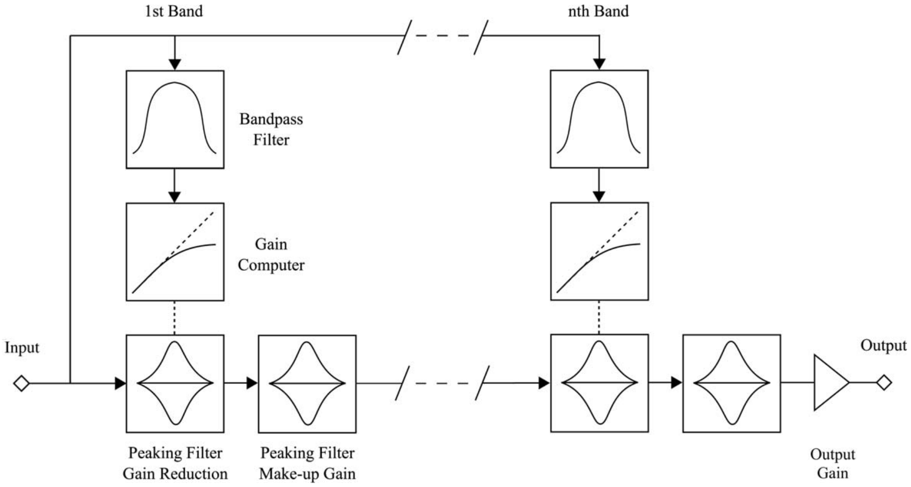

Fig. 2. Architecture of 10-band dynamic equalizer (DEQ10) and a 30-band dynamic equalizer (DEQ30) audio effects.

tion, the unprocessed snare drums from the test dataset of input-target pairs are transformed using the proposed audio effect configurations, where parameter values for each audio effect are inferred from the trained encoder network.

2.1 Dataset

In order to train DeepAFx to learn the most suitable parameters for any given audio processing task it requires input-target paired audio as supervision. The training data for each of the four transformation tasks is comprised of specific subsets from the Snare Drum Data Set (SDDS) [36].[3] From the four subsets, 3,000 input-target pairs were randomly selected to create the test set. SDDS is a comprehensive acoustic snare drum dataset, featuring multi-velocity recordings of ten different snare drums, each captured with 53 studio microphones, using various commercial dampening techniques.

One of the dampening methods used in SDDS was a BigFatSnareDrum (BFSD), a specialized device designed to dampen a snare or tom, placed directly on top of the batter head. This allows for exact repeatability because it covers the entirety of the drumhead and could only be placed in one position unlike other products. Although SDDS included other dampening methods such as MoonGel, BSFD was chosen to be used for the dampening transformations because it produces a distinct timbral change. The BFSD is also used for the D2U transformation. For each U2D and D2U input-target pair, the snare drum, microphone, and mic position were all identical, with the only variable being the dampening, either undampened or dampened with a BFSD.

Individual strikes from each pair were matched based on closest peak amplitude levels and time-aligned using cross correlation. For the positional transformation of T2B and B2T, only eight of the same microphones were used in both top and bottom positions. These pairs were used on all 10 snare drums and for all dampening methods; the paired strikes were identical performances because the top and bottom microphones were recorded simultaneously. For each subset, 80% was used for training, 10% for validation, and the remaining 10% for test data for later evaluation. Once processed by the trained models, the evaluation data was used for the comparative metrics and provided stimuli for the subjective listening tests.

2.2 Subjective Evaluation

A subjective listening test was carried out using a multiple stimulus approach in order to determine whether participants would perceive the transformed audio as comparable to the real-world recording parameter adjustments it was emulating. The test was implemented using the Web Audio Evaluation Tool [37] and was carried out by 25 participants between the ages of 20–42 (mean: 27), and their experience in audio-related fields ranged from 1 to 25 years (mean: 9). Participants were instructed to use the highest-quality playback system available to them. They were required to provide the specification of equipment used, and all systems reported were deemed to be suitably professional.

The four transformations were evaluated on separated pages of the listening test. On each page, participants were presented with seven sliders, each corresponding to a different audio sample. The page and slider order were randomized, and slider starting position was as well. The seven audio stimuli were comprised of the unprocessed input used as a baseline for similarity, with the target acting as a *hidden reference*, and the five samples of the input processed by the five different plugin chains. Participants were instructed to arrange stimuli based on their similarity to the *reference* and use the full range of the scale, placing the most similar at the top and least similar at the bottom. The *hidden reference* was used to ensure participants could accurately identify the identical sample to the *reference*. No low anchor was used, in order to allow participants to rate the perceptually least similar stimuli lowest on the rating scale. Stimuli were loudness normalized to –23 LUFs.

[3] http://dmtlab.bcu.ac.uk/matthewcheshire/audio/sdds.

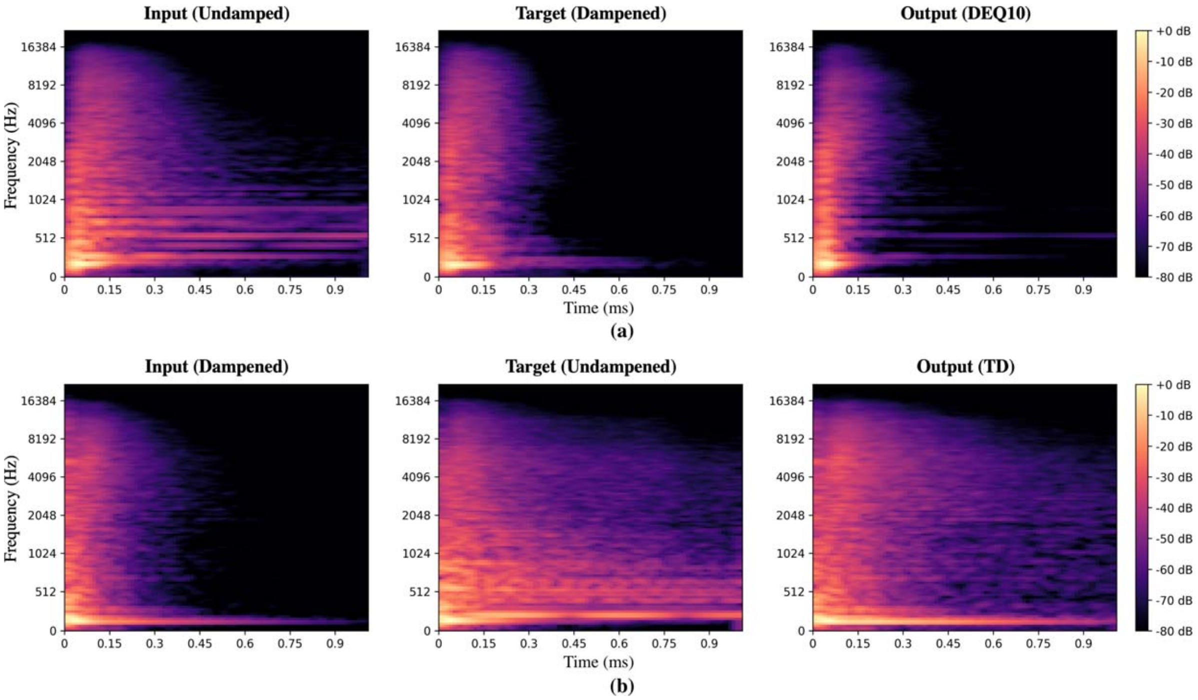

Fig. 3. Mel-scaled log frequency spectrograms for (a) Undampened-to-Dampened with 10-band dynamic equalizer (DEQ10) and (b) Dampened-to-Undampened with a transient designer (TD). Input snare drums (left), target (center), output transformations (right).

Listening test stimuli are available for audition.[4] The input-target pairs for each transformation were randomly selected from the test data subset (SEC. 2.1). Participants could not move on to the next page until all stimuli were played at least once and all sliders were moved. Fig. 3 presents an example of (a) U2D snare transformation using DEQ10 and (b) D2U snare transformation using TD.

2.3 Reconstruction Metrics

In order to evaluate the ability of the model to produce desired transformations of snare recordings, how accurately the transformed examples $\hat{S}$ match the target examples S Recording pairs in the test set introduced in SEC. 2.1 are evaluated using reconstruction metrics in two experiments comparing timbre and pitch characteristics of the transformed snare drums. Each transformation type is grouped into two tasks: 1) dampening (i.e., U2D and D2U) and 2) positional (i.e., T2B and B2T) and is evaluated with a range of spectral representations and metrics focused on timbral (see SEC. 2.3.1) and pitch (see SEC. 2.3.2) reconstruction capabilities of the model.

To extract the selected comparative metrics, a magnitude spectrogram S_{stft} is computed using the STFT for each audio file using an n-length Hann window ($n = 2{,}048$) with a hop size of $\frac{n}{4}$. S_{stft} is additionally mapped onto the Mel-scale or converted to Mel-frequency cepstral coefficients, resulting in S_{Mel} and S_{mfcc}, respectively.

2.3.1 Timbral Reconstruction

Timbral reconstruction metrics in the first experiment include MSL (see SEC. 1.3) and spectral cosine distance (SCD) metrics as used in [22], along with log-spectral distance (LSD) [38] and Pearson correlation (PC) coefficients, which were previously employed in evaluations of deep generative models for music signals as an objective measure of audio quality [39, 40]. Additionally, the cosine similarity (CS) metric based on spectral difference functions (SDFs) used in research on automatic event detection [41] and automatic music remixing [42] are used. The implementation by [22] is followed for the computation of MSL and SCD metrics, in which the former uses STFT magnitudes and latter uses 13 Mel-frequency cepstral coefficients (excluding the first coefficient). The LSD is calculated using Mel-spectrograms as follows:

$$LSD_{Mel}(S, \hat{S}) = \sqrt{\sum [10 \log_{10}(|S|/|\hat{S}|)]^2}. \quad (2)$$

Following [41], spectral difference envelopes E are computed as

$$E_S(t) = \sum_{k=0}^{K-1} \left\{ H(|S_k(t+1)| - |S_k(t)|) \right\}, \quad (3)$$

where S represents a Mel-spectrogram with K bins. The $H(x) = (x + |x|)/2$ is a half-wave rectifier, which returns zero for negative arguments. The calculations of the E_S envelopes is the same for $E_{\hat{S}}$. Following [43], envelope reconstruction of the transformations is evaluated with co-

[4] https://dmtlab.bcu.ac.uk/matthewcheshire/audio/jaes_samples/.

sine similarity calculated between envelopes extracted from target and transformed recordings as follows:

$$CS_{sdf}(S, \hat{S}) = \frac{E_S \cdot E_{\hat{S}}}{\|E_S\| \|E_{\hat{S}}\|}, \quad (4)$$

where · represents a dot product between E. CS_{sdf} will be close to unity for very similar drum envelopes and nearer to zero for dissimilar ones. Spectral difference functions are then calculated as the sum of the first-order difference between each spectrogram (e.g., [44]). The resulting envelopes are then normalized between [0, 1].

All reported timbral reconstruction experiments are presented as means calculated over the test set (see Sec. 2.1) except the MSL_{stft} metric, which is represented as the sum of L2 differences [see Eq. (1)]. Although the computation of PCs is described in the following section, here they are reported as mean PC coefficients averaged over the frequency axis.

2.3.2 Pitch Reconstruction

In the second experiment, a pitch-based reconstruction metric, which was previously used to evaluate the audio quality of pitched instruments generated with an adversarial autoencoder [39], was implemented. This approach is modified to suit snare drum frequency ranges. The use of Mel-spectrograms is opted for, as opposed to constant-Q transform spectrograms used in [39], because a logarithmically-spaced frequency range provides a more even representation over the fundamental frequencies of snare signals than frequency representations spaced over musical octaves (e.g., constant-Q transforms).

3 RESULTS

3.1 Subjective Results

3.1.1 Dampening

Fig. 4 presents normalized violin plots showing the dampening transformation results for the subjective listening test (means are depicted by asterisks, and medians are denoted by black horizontal lines). A one-way analysis of variance (ANOVA) was used to determine whether distributions of the responses have a common mean—that is, whether the plugin chains under evaluation had a different effect on the subjective scores of similarity. U2D (p = 3.12e–14) and D2U (p = 4.81e–14) both had $p < 0.05$. The small p values allow for rejection of the hypothesis that all group means are equal and indicate that the different ratings are not the same as each other.

A post-hoc multiple pairwise comparison was used to establish which of the ratings were significant based on the results from the ANOVA test. As per the recommendations in [45], Tukey's Honestly Significant Difference (HSD) test was used for this comparison. The U2D subjective listening test showed promising results. It can be seen in Fig. 4 that DEQ10 (mean: 0.66) and DEQ30 (mean: 0.58) are rated more similarly to the hidden reference (mean: 1) than the input (mean: 0.3). All participants correctly identified the hidden reference, placing it at the top of the rating scale. The ratings for DEQ10 and DEQ30 were both statistically higher than the input (p = 2.07e–08 and p = 9.84e–06, respectively) using HSD. This suggests that both of these effects moved the processed input perceptually closer in similarity to the reference, which in this instance was a snare drum recording dampened with a BFSD.

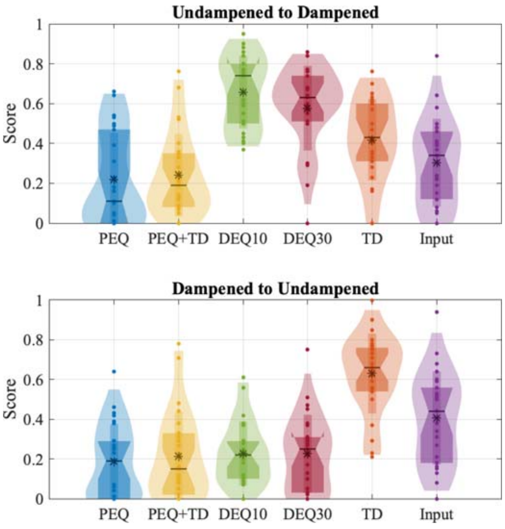

Fig. 4. Dampening results from listening test. DEQ10 = 10-band dynamic equalizer; DEQ30 = 30-band dynamic equalizer; PEQ = parametric equalizer; TD = transient designer.

Although not able to completely emulate the real dampening effect, these results indicate that the transformation is indeed creating a more dampened sound compared with the undampened recording. It should be noted that all participants were able to correctly identify the hidden reference and placed it at the top of the scale for all four test pages. Although TD (mean: 0.42) was rated higher than the input overall, the ratings were not significantly higher (p = 0.054). Likewise, although PEQ and PEQ+TD do have lower overall ratings than the input, they are not statistically different. For D2U, the only effect that had a significantly higher rating (p = 0.0012) than the input (mean: 0.4) was TD (mean: 0.63) based on HSD, which can be seen in Fig. 4.

3.1.2 Positional

The listening test results for the positional transformation are presented in Fig. 5. All participants correctly identified the hidden reference (mean: 1). An ANOVA was used again to determine whether any of the ratings were significantly different; for B2T transformations, it was found that there were no statistical differences between any of the scores (p = 0.42). This can be seen by the relatively close means and overlapping ranges of the different ratings. Although DEQ10 has a higher rating (mean: 0.49) than the input (mean: 0.36), these ratings were not statistically different from each other when the HSD test was conducted.

For T2B, some significant differences were shown based on the results from an ANOVA (p = 1.54e–11). The HSD test revealed that the performance of both PEQ and PEQ+TD (mean: 0.34 and 0.16, respectively) were statisti-

Table 1. Dampening task results using Mel-spectrograms: mean multi-scale loss (MSL), spectral cosine distance (SCD), log-spectral distance (LSD), mean Pearson correlation (PC), and envelope cosine similarity (CS). Lower values indicate greater similarity, except for the PC and CS metrics, for which higher values do.

	MSL_{stft}		SCD_{mfcc}		LSD_{Mel}		PC_{Mel}		CS_{sdf}	
Name	U2D	D2U	U2D	D2U	U2D	D2U	U2D	D2U	U2D	D2U
PEQ	8.31	65.57	0.75	0.90	2.53	3.09	0.68	0.52	0.86	0.69
TD	6.92	12.90	0.73	**0.85**	2.78	**2.72**	0.64	0.60	0.70	**0.91**
PEQ+TD	8.91	39.96	0.64	0.87	2.45	3.49	0.62	0.45	0.61	0.52
DEQ10	**4.77**	11.83	**0.55**	0.80	**2.13**	4.32	**0.70**	**0.68**	**0.89**	0.90
DEQ30	5.46	**8.01**	0.63	0.87	2.25	4.71	0.69	**0.68**	0.86	0.90

Bold values indicate best score (highest or lowest based on metric). DEQ10 = 10-band dynamic equalizer; DEQ30 = 30-band dynamic equalizer; D2U = Dampened-to-Undampened; mfcc = Mel-frequency cepstral coefficient; PEQ = parametric equalizer; sdf = spectral difference function; stft = short-time Fourier transform; TD = transient designer; U2D = Undampened-to-Dampened.

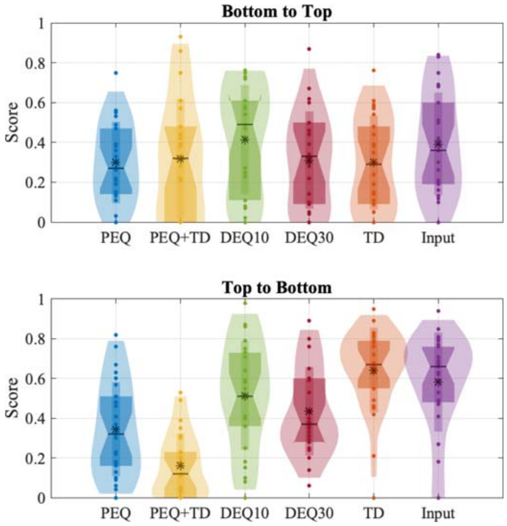

Fig. 5. Positional results from listening test. DEQ10 = 10-band dynamic equalizer; DEQ30 = 30-band dynamic equalizer; PEQ = parametric equalizer; TD = transient designer.

cally lower than the input (mean 0.58), with PEQ+TD being rated least similar to the target. TD had slightly higher ratings (mean 0.64) than the input, but again, these ratings were not statistically different from each other. This showed that for T2B positional changes, no method was successful at moving the input perceptually closer to the target, with both PEQ and PEQ+TD statistically worsening similarity. For B2T, no significant effects were seen, either positively or negatively, by any of the transformations.

3.2 Objective Results

Several of the objective metrics for U2D shown in Table 1 display similar trends to the subjective evaluations. For U2D all metrics showed DEQ10 to be most similar to the target. For D2U, TD rated most similar in the subjective evaluation and measured most similar when using SCD, LSD, and CS; however, unlike the subjective ratings when using MSL and PC, DEQ30 performed the best.

The objective metrics for the positional tasks can be seen in Table 2, DEQ10 had the highest similarity for T2B and B2T when measured with MSL and PC, respectively. PEQ also showed favorable results for T2B when using LSD and B2T when using both MSL and CS. TD was another effect that performed well across different metrics because it displayed the highest similarity with both SCD and PC for the T2B transformation. PEQ+TD was the only effect that presented strong similarity for one metric alone, with it scoring most similarly when using SCD for B2T.

4 DISCUSSION

The results from the listening test indicate that D2U may be a harder transformation to emulate than U2D, with both DEQ10 and DEQ30 being rated statistically more similar

Table 2. Positional task results, metrics are the same as those used in Table 1. Lower values indicate greater similarity, except for the PC and CS metrics, for which higher values do.

	MSL_{stft}		SCD_{mfcc}		LSD_{Mel}		PC_{Mel}		CS_{sdf}	
Name	B2T	T2B	B2T	T2B	B2T	T2B	B2T	T2B	B2T	T2B
PEQ	**7.86**	10.63	0.39	0.43	2.09	**2.11**	0.64	0.53	**0.91**	0.87
TD	10.16	7.35	0.40	**0.39**	2.34	2.07	0.61	**0.64**	0.89	**0.92**
PEQ+TD	17.86	23.09	**0.35**	0.42	**1.81**	2.45	0.52	0.38	0.48	0.57
DEQ10	8.17	**5.83**	0.54	0.50	2.39	2.54	**0.66**	0.54	0.83	0.87
DEQ30	8.27	6.33	0.68	0.62	2.61	3.01	0.65	0.54	0.81	0.88

Bold values indicate best score (highest or lowest based on metric). B2T = bottom-to-top microphone position; CS = cosine similarity; DEQ10 = 10-band dynamic equalizer; DEQ30 = 30-band dynamic equalizer; LSD = log-spectral distance; mfcc= Mel-frequency cepstral coefficient; MSL = multi-scale loss; PC = Pearson correlation; PEQ = parametric equalizer; SCD = spectral cosine distance; sdf = spectral difference function; stft = short-time Fourier transform; T2B = top-to-bottom microphone position; TD = transient designer.

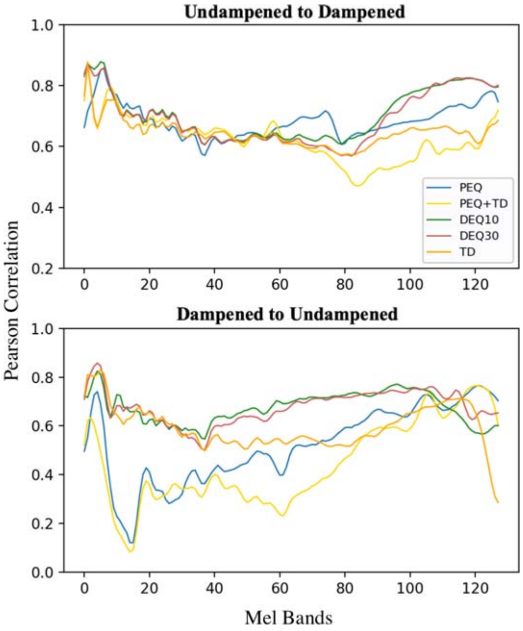

Fig. 6. Mean smoothed Pearson correlation results computed with Mel-spectrograms for the dampening task. DEQ10 = 10-band dynamic equalizer; DEQ30 = 30-band dynamic equalizer; PEQ = parametric equalizer; TD = transient designer.

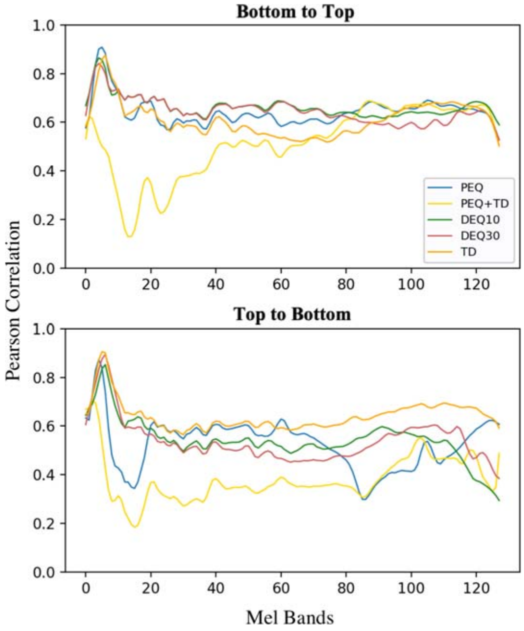

Fig. 7. Mean smoothed Pearson correlation results computed with Mel-spectrograms for the positional task. DEQ10 = 10-band dynamic equalizer; DEQ30 = 30-band dynamic equalizer; PEQ = parametric equalizer; TD = transient designer.

to the target for U2D but had ratings that were not significantly different to the input when used for D2U. Dampening a snare drum removes high-frequency energy, whereas removing dampening increases higher frequencies. When dealing with a heavily dampened snare recording, the high-frequency content has already been removed, and it shows that DeepAFx was not able to learn optimal parameters for the effects to enhance the missing information.

TD was most successful for the D2U transformation, likely because of TD's *release boost* parameter, shaping the envelope of the drum recording to better emulate an undampened strike. One possible alternation to DEQ10 and DEQ30 that may have facilitated better results for D2U would be to change the *ratio* parameter to allow values below 1. This would create an expansion effect instead of a compression effect for each frequency band, which could possibly be used to create a similar effect to that of the TD. Fig. 6 displays the mean smoothed PC results for the dampening tasks. High degrees of similarity to the target can be observed by both DEQ10 and DEQ30 only for the higher-frequency ranges for U2D. Little difference is seen between any of the plugins for the lower-frequency bands. Because high frequencies are most affected by dampening, the high measure of similarity in these important bands is likely responsible for the significantly higher ratings in the subjective evaluation.

For D2U, DEQ10 and DEQ30 have the highest measures of similarity in the mid-range and upper-mid–range frequency bands; however, this similarity is not reflected in the subjective tests. Although TD was subjectively the most similar to the target, the PC in Fig. 6 shows that it does not outperform DEQ10 or DEQ30, suggesting that envelope similarity is more important for D2U than spectral similarity. The subjective evaluation for B2T did not show any effect chain to statistically produce different ratings. In the case of T2B, PEQ and PEQ+TD produced ratings that were statistically lower than the input. A possible cause for this may be that the input is rated similar to the target. With little timbral disparity between input and target, it may be more difficult for DeepAFx to use the provided plugins to make the necessary improvement. PEQ and PEQ+TD also showed very low similarity for the mean smoothed PC results for T2B seen in Fig. 7, with the most notable dissimilarity being in the lower-frequency ranges and upper-mid range. PEQ+TD also showed very poor similarity in the lower frequencies for B2T; however, this was not reflected in the subjective evaluations.

The stimuli selected for the listening tests may not best exemplify the ideal transformation because the input-target pairs were chosen randomly from the available evaluation data. Thus, more representative samples that were not able to be assessed during the subjective evaluations may exist. Other variables, such as timbral differences associated with velocity disparity, could also play a part in the subjective perception of similarity. The effects of dampening or microphone placement may be less pronounced when the snare is played very lightly. Certain microphones may be more adept at capturing timbral subtleties making it easier for a listener to distinguish changes, or particular snare drums may emphasis the effects of parameter changes more so than others. The relationship between subjective ratings and objective metrics cannot be strongly linked because the objective measures made use of all samples from the evaluation data. In most cases DEQ10 outperformed DEQ30, which indicates that octave-band control (i.e., DEQ10) had sufficient timbral shaping abilities and third-octave band (i.e., DEQ30) had no additional benefits.

5 CONCLUSION

In this study, a deep learning system for automatic modification of snare drum recording parameters has been investigated. Two novel audio effects, an octave-band and third-octave–band dynamic EQ with fixed center frequency bands and trainable parameters, were created specifically for use within this system. Results from a subjective evaluation demonstrated that with particular effects, the system was able to move perceptually closer to the real-world targets for dampening tasks but was unsuccessful in positional transformations. Objective metrics also revealed a tendency toward improvements in similarity for certain transformations. Most notably, DEQ10 performed best at Undampened-to-Dampened in all measures.

A possible direction for future research in this area would be to assess the benefits of additional computational power, larger datasets, and alternative architectures to improve the quality of the transformations. The authors would also like to explore more aspects of the recording process, for example, transformations between different drum shell materials and investigation of other audio effects, such as distortions or reverbs for their timbral shaping capabilities. Additionally, subsequent studies could investigate methods for navigating the network's latent space. Navigation controls could be provided as a GUI to creatively interpolate between transformations or refine the estimated parameters.

6 REFERENCES

[1] R. Toulson, "The Perception and Importance of Drum Tuning in Live Performance and Music Production," in *Proceedings of the 4th Art of Record Production Conference* (Lowell, MA) (2008 Nov.).

[2] B. Owsinski and D. Moody, *The Drum Recording Handbook* (Hal Leonard, Milwaukee, WI, 2009).

[3] B. Bartlett and J. Bartlett, *Practical Recording Techniques: The Step-by-Step Approach to Professional Audio Recording* (Focal Press, Waltham, MA, 2008), 5th ed.

[4] K. Yoshii, M. Goto, and H. Okuno, "INTER:D: A Drum Sound Equalizer for Controlling Volume and Timbre of Drums," in *Proceedings of the 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, pp. 205–212 (London, UK) (2005 Nov.). https://doi.org/ds85t4.

[5] D. Moffat and M. B. Sandler, "Machine Learning Multitrack Gain Mixing of Drums," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), e-Brief 527.

[6] M. A. Martínez Ramírez, D. Stoller, and D. Moffat, "A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net," *J. Audio Eng. Soc.*, vol. 69, no. 3, pp. 142–151 (2021 Mar.). https://doi.org/h4zh.

[7] M. Seymour, "Engineer's Guide To Tuning and Damping Drums," https://www.soundonsound.com/techniques/engineers-guide-tuning-and-damping-drums (2010 Aug.).

[8] N. D'Virgilio, "How to Control Drum Sustain With Dampening," https://www.sweetwater.com/insync/how-to-control-drum-sustain-with-dampening/#:~:text=Tear%20a%20small%20piece%20of,to%20really%20dampen%20the%20head.&text=Moongel%20is%20a%20great%20product%20to%20keep%20in%20your%20stickbag (2014 Sep.).

[9] M. Major, *Recording Drums: The Complete Guide* (Course Technology PTR, Boston, MA, 2014).

[10] M. H. Parsons, *The Drummer's Studio Survival Guide: How to Get the Best Possible Drum Tracks on Any Recording Project* (Modern Drummer Publications, Cedar Grove, NJ, 1996).

[11] B. Gibson, *Sound Advice on Recording & Mixing Drums* (Alfred Music, Los Angeles, CA, 2004).

[12] B. Owsinski, *The Recording Engineer's Handbook* (Bobby Owsinski Media Group, Burbank, CA, 2017), 4th ed.

[13] M. Cheshire, J. Hockman, and R. Stables, "Microphone Comparison for Snare Drum Recording," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 10040.

[14] K. Pedersen and M. Grimshaw-Aagaard, *The Recording, Mixing, and Mastering Reference Handbook* (Oxford University Press, New York, NY, 2019).

[15] M. Cheshire, R. Stables, and J. Hockman, "Microphone Comparison: Spectral Feature Mapping for Snare Drum Recording," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10263.

[16] M. A. Martínez Ramírez, E. Benetos, and J. D. Reiss, "Deep Learning for Black-Box Modeling of Audio Effects," *Appl. Sci.*, vol. 10, no. 2, paper 638 (2020 Jan.). https://doi.org/gppmr6.

[17] A. Wright, E.-P. Damskägg, L. Juvela, and V. Välimäki, "Real-Time Guitar Amplifier Emulation With Deep Learning," *Appl. Sci.*, vol. 10, no. 3, paper 766 (2020 Jan.). https://doi.org/ggvv8t.

[18] S. Hawley, B. Colburn, and S. I. Mimilakis, "Profiling Audio Compressors With Deep Neural Networks," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10222.

[19] J. Rämö and V. Välimäki, "Neural Third-Octave Graphic Equalizer," in *Proceedings of the 22nd International Conference on Digital Audio Effects*, paper 21 (Birmingham, UK) (2019 Sep.).

[20] S. Stasis, R. Stables, and J. Hockman, "Semantically Controlled Adaptive Equalisation in Reduced Dimensionality Parameter Space," *Appl. Sci.*, vol. 6, no. 4, paper 116 (2016 Apr.). https://doi.org/h4zj.

[21] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *Proceedings of the International Conference on Learning Representations* (Addis Ababa, Ethiopia) (2020 Apr.).

[22] M. A. Martínez Ramírez, O. Wang, P. Smaragdis, and N. J. Bryan, "Differentiable Signal Processing With Black-Box Audio Effects," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Sig-*

nal Processing, pp. 66–70 (Toronto, Canada) (2021 May). https://doi.org/h4zk.

[23] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Metric Learning vs Classification for Disentangled Music Representation Learning," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, pp. 439–445 (Montreal, Canada) (2020 Oct.).

[24] C. Szegedy, W. Liu, Y. Jia, et al., "Going Deeper With Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (Boston, MA) (2015Jun.).

[25] A. Fox, "Top 10 Best Dynamic EQ Plugins for Your DAW in 2022," https://mynewmicrophone.com/top-best-dynamic-eq-plugins-for-your-daw/ (2022).

[26] M. Hahn, "How to Use Dynamic EQ for A Better Mix," https://blog.landr.com/dynamic-eq/ (2020 Jul.).

[27] I. Stewart, "How to Use Dynamic EQ in Mastering," https://www.izotope.com/en/learn/how-to-use-dynamic-eq-in-mastering.html (2021 Sep.).

[28] G. Brown, "Multiband Compressors vs. Dynamic EQs: Differences and Uses," https://www.izotope.com/en/learn/multiband-compressors-vs-dynamic-eqs.html (2020 Apr.).

[29] ANSI, "Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters," *Standard ANSI S1.11-2004* (2009 Jun.).

[30] D. K. Wise, "Concept, Design, and Implementation of a General Dynamic Parametric Equalizer," *J. Audio Eng. Soc.*, vol. 57, no. 1/2, pp. 16–28 (2009 Jan.).

[31] R. Izhaki, *Mixing Audio: Concepts, Practices, and Tools* (Taylor & Francis, New York, NY, 2017), 3rd ed. https://doi.org/h4zn.

[32] H. Gier and P. White, "Transient Designer - Dual-Channel, Model 9946," https://spl.audio/wp-content/uploads/transient_designer_2_9946_manual.pdf (1999 Jul.).

[33] B. Owsinski, *The Mixing Engineer's Handbook* (Bobby Owsinski Media Group, Burbank, CA, 2016), 4th ed.

[34] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the International Conference on Learning Representations* (San Diego, CA) (2015 May).

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034 (Santiago, Chile) (2015 Dec.). https://doi.org/gfw55h.

[36] M. Cheshire, "Snare Drum Data Set (SDDS): More Snare Drums Than You Can Shake a Stick At," presented at the *149th Convention of the Audio Engineering Society* (2020 Oct.), e-Brief 626.

[37] N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, "Web Audio Evaluation Tool: A Browser-Based Listening Test Environment," in *Proceedings of the 12th Sound and Music Computing Conference*, pp. 147-152 (Kildare, Ireland) (2015 Jul.).

[38] A. Bitton, P. Esling, A. Caillon, and M. Fouilleul, "Assisted Sound Sample Generation With Musical Conditioning in Adversarial Auto-Encoders," in *Proceedings of the International Conference on Digital Audio Effects*, paper 24 (Birmingham, UK) (2019 Sep.).

[39] J. W. Kim, R. Bittner, A. Kumar, and J. P. Bello, "Neural Music Synthesis for Flexible Timbre Control," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 176–180 (Brighton, UK) (2019 May). https://doi.org/h4zq.

[40] M. Tomczak, J. Drysdale, and J. Hockman, "Drum Translation for Timbral and Rhythmic Transformation," in *Proceedings of the International Conference on Digital Audio Effects*, paper 25 (Birmingham, UK) (2019 Sep.).

[41] J. P. Bello, L. Daudet, S. Abdallah, et al., "A Tutorial on Onset Detection in Music Signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047 (2005 Sep.). https://doi.org/dw7kcc.

[42] M. E. P. Davies, P. Hamel, K. Yoshii, and M. Goto, "AutoMashUpper: Automatic Creation of Multi-Song Music Mashups," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1726–1737 (2014 Dec.). https://doi.org/h4zr.

[43] M. Tomczak, M. Goto, and J. Hockman, "Drum Synthesis and Rhythmic Transformation With Adversarial Autoencoders," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2427–2435 (Seattle, WA) (2020 Oct.). https://doi.org/h4zs.

[44] S. Dixon, F. Gouyon, and G. Widmer, "Towards Characterisation of Music via Rhythmic Patterns," in *Proceedings of the International Conference on Music Information Retrieval*, pp. 509 (Barcelona, Spain) (2004 Jan.).

[45] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).

THE AUTHORS

Matthew Cheshire

Jake Drysdale

Sean Enderby

Maciek Tomczak

Jason Hockman

Matthew Cheshire is a Ph.D. student in the SoMA Group at Birmingham City University. His main interests are timbral modifications, recording techniques, and perceptual evaluations. He has been a member of the AES since 2018, and he was the recipient of the Saul Walker Award (2019) and the John Eargle Award (2020 and 2021).

•

Jake Drysdale is a Ph.D. student in the SoMA Group, where he specializes in deep learning methods to assist in sample-based music production. His main interests are neural audio synthesis and music sample retrieval.

•

Sean Enderby completed his Ph.D. in Intelligent Music Production at Birmingham City University in 2017. He is currently a Lecturer and Researcher in the SoMA Group. His main areas of interest are music digital signal processing, virtual analog, and "intelligent" parameterization of effects.

•

Maciek Tomczak is a Ph.D. student in the SoMA Group. His main areas of interest are machine learning, computational rhythm analysis, and audio style transfer.

•

Jason Hockman received a doctorate in Music Research from McGill University in 2014. He is currently Associate Professor of Audio Engineering at Birmingham City University, where he leads the SoMA Group. His main areas of interest are in music informatics and computational rhythm and meter analysis.