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. X, no. X, pp. 1–10, (2021 X).
DOI: https://doi.org/10.17743/jaes.2020.0031

A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net*

MARCO A. MARTÍNEZ RAMÍREZ,[1]* DANIEL STOLLER,[1]* AND DAVID MOFFAT,[2] *AES Member*
(m.a.martinezramirez@qmul.ac.uk) (business@dstoller.net) (david.moffat@plymouth.ac.uk)

[1] *Centre for Digital Music, Queen Mary University of London, London, United Kingdom*
[2] *Interdisciplinary Center for Computer Music Research, University of Plymouth, Plymouth, United Kingdom*

The development of intelligent music production tools has been of growing interest in recent years. Deep learning approaches have been shown as being a highly effective method for approximating individual audio effects. In this work, we propose an end-to-end deep neural network based on the Wave-U-Net to perform automatic mixing of drums. We follow an end-to-end approach where raw audio from the individual drum recordings is the input of the system and the waveform of the stereo mix is the output. We compare the system to existing machine learning approaches to intelligent drum mixing. Through a subjective listening test we explore the performance of these systems when processing various types of drum mixes. We report that the mixes generated by our model are virtually indistinguishable from professional human mixes while also outperforming previous intelligent mixing approaches.

0 INTRODUCTION

Audio mixing is the process of blending multitrack recordings by manipulating the dynamics, spatialization, or timbre of the respective musical sources. This manipulation is achieved through a set of linear and nonlinear audio effects, such as gain, panning, equalization (EQ), dynamic range compression (DRC), and artificial reverberation [1].

A fully automatic mixing system is one where a set of audio tracks can be fully mixed and produced by a computational system without human intervention [2]. There have been a number of different approaches to attempt to solve this task [3]; however there is no single unified approach or solution to the mixing problem [4].

Intelligent music production has been a growing field, with many focuses on the latest growth in machine learning to be applied to audio. The aim of this work is to identify if it is possible for a deep learning approach to learn the entire audio transformation. In this work we will use the task of mixing a set of drums together to simplify the problem from all musical mixes to a specific subgroup mix [5]. Subgrouping is a common practice in music production where individual parts of a mix are broken down into smaller mixes or stems, and drum subgroups are one of the most commonly used stems [6].

This paper presents an end-to-end deep neural network (DNN) approach to mixing of audio content where raw audio is both the input and output of the system. A data-driven approach based on a variant of the *Wave-U-Net* [7] is used to represent the mixing process to be undertaken, and the audio will be manipulated directly by the neural network to produce an output mix. Taking into account that the *Wave-U-Net* is originally designed to perform audio source separation, we investigate automatic mixing by *swapping* input and output of the *Wave-U-Net*, i.e., the input corresponds to the musical sources and the output to the mixture.

The performance of this model is compared to a range of human-made mixes. In addition, alternative approaches to automatic mixing based on signal processing and random forest regressors are used as baseline models. It is expected that a neural network approach to automatic mixing considerably outperforms mixing approaches derived from traditional signal processing. We explore the performance of these systems via a listening test and find that the *Wave-U-Net* mixes are virtually indistinguishable from a

*These authors contributed equally to this work.

*DM is supported by EPSRC Grant EP/L019981/1, EPSRC grant EP/L01632X/1, EPSRC Grant EP/S026991/1 RadioMe: Real-time Radio Remixing.

human-made mix while also achieving higher ratings than mixes from baseline models.

The rest of the paper is structured as follows. Sec. 1 presents the state-of-the-art in intelligent mixing systems, discussing existing machine learning approaches to music production and identifying relevant preceding work. Sec. 2.1 explains the deep learning approach that is implemented and compared to a previous machine learning approach, presented in Sec. 2.2. Sec. 2.3 presents the dataset that we will use for training and evaluation of this paper. The resulting generated mixes are then evaluated using a subjective listening experiment, described in Sec. 3. The effectiveness of the deep learning mixing methods are evaluated in Sec. 4, and conclusions presented in Sec. 5.

1 RELATED WORK

There are a number of different approaches that have attempted to automate the process of music mixing. Pachet and Delerue [8] first framed the music mixing process as a constraint optimization problem, where rules are defined and simultaneously evaluated to find the most optimal solution in a dynamically changing environment. The concept of mixing rules has since been taken up in a large portion of literature, including by Bocko et al. [9], who proposed building an expert knowledge fully automatic mixing system, where audio features are related to mixing rules. De Man and Reiss [10] developed this approach further, defining some potential rules for how they can be applied. Everardo [11] proposed formalizing the logical rule structure, defining a rule syntax and structure, which was then developed further and formalized within the semantic web context by Moffat et al. [12].

Conversely there are a number of approaches that rely on some grounded theory and predefined signal processing paths to perform music mixing. Matz et al. [13] combined a range of different automatic mixing processes and constructed a flow diagram as to how different tracks should be mixed, including source separation. They demonstrate the importance of gain and equalization in automatic mixing. De Man et al. [14] discussed many different approaches for automating each individual audio effect and how this can be applied to automatic mixing process; however there is limited acknowledgment as to the impact that audio effects will have on each other.

Moffat and Sandler [2] present an overview of automatic mixing approaches. It is identified that instead of using formal logic structures for rule mixing, a data-driven approach to understanding the music mixing process would be very valuable. This is demonstrated by Martínez Ramírez and Reiss [15], who propose a proof-of-concept of a data-driven, neural network approach to stem audio mixing. Furthermore Mimilakis et al. [16] investigate DNNs to predict gain coefficients in order to perform automatic DRC for mastering applications. Moffat and Sandler [17] developed a machine learning, data-driven approach for prediction of gain mixing parameters, where the gain parameters are approximates using a reverse engineering approach [18].

The approach we take in this paper leverages the recent progress made in other audio processing tasks—mainly audio source separation, which requires the prediction of audio sources from a given audio mixture. While it intuitively represents the "opposite" to the mixing problem we consider, there is considerable overlap in the types of challenges involved. Firstly, in both cases there is uncertainty about the correct output for a given input—for separation, since different sources can sometimes produce the same mixture, and for automatic mixing, because of the wide variety of suitable effects that could be applied to the stems. Both tasks also require models that process audio signals as input and output audio signals, which is challenging, as they are very high-dimensional because of their high sampling rate. To deal with the audio input and output, almost all previous separation approaches convert the audio input to a spectrogram-based representation. Most systems use this representation to predict spectrograms for each source before converting them to time-domain audio outputs [19–23].

However, spectrogram-based approaches suffer from two problems. Firstly, they cannot take input phase into account, which can be detrimental to performance. Secondly, the spectrogram inversion step to obtain time-domain audio signals is only approximate, which introduces artifacts. When applying these approaches to our automatic mixing task, where time-based effects cause phase manipulations and oftentimes only subtle changes need to be applied, the above issues can be expected to be especially noticeable. Furthermore, preserving the phase of the stems when performing mixing is a desirable property of an automatic mixing approach to avoid unexpected phase interactions between the stem signals.

To avoid the above issues, novel time-domain approaches have been proposed in recent literature. Martínez Ramírez et al. [24–27] investigated DNNs for audio processing tasks, such as modeling of various types of audio effects. Similarly, Wright et al. [28] explored variants of the *WaveNet* architecture [29] and recurrent neural networks to model distortion audio effects. Hawley et al. [30] proposed a DNN based on *U-Net* [31] and *Time-Frequency* [32] networks to model DRC. Instead of requiring a spectrogram inversion step, most of these systems are simply trained to minimize a mean-absolute-error (L1) or mean-squared-error (L2) between the predicted and the ground truth time-domain signal.

The aforementioned methods model only static or parametric configurations of individual audio processors; therefore a much more complex task such as mixing may not be feasible. In the domain of speech enhancement, various time-domain approaches have been proposed [33, 34]. The *Wave-U-Net* [7] has been shown to achieve high performances for the more general problem of audio source separation. Because of its generic, convolutional architecture, we hypothesize that it can be used as a generic audio-to-audio transformation model with only little adaptation. Thus we explore automatic mixing by *swapping* the inputs and outputs of the *Wave-U-Net*.

2 METHODOLOGY

2.1 Wave-U-Net

The main model used in this paper is a variant of the *Wave-U-Net* proposed in [7] for the task of audio source separation. It uses raw audio input and output combined with a series of downsampling (*DS*) and upsampling (*US*) blocks consisting of 1D convolutions followed by resampling operators to compute features at multiple-timescales that can be used for prediction.

Since the model was successfully employed for mono as well as stereo separation simply by changing the number of input and output channels in the first and last convolution, respectively, we suspect it can be generally used for audio-to-audio transformation tasks and adapted to drum mixing.

Considering the model has raw audio input and output, we did not use any preprocessing of the audio signals, which are sampled at 44.1 kHz. Our goal is to blend K monophonic waveforms or stems $S^1, \ldots, S^K$ into a stereo mixture waveform M.

Besides changing the number of input channels to the number of mono sources to be mixed ($K = 8$) and the number of output channels to 2 because of the stereo mixture output required, we modify the original *Wave-U-Net* as follows. The number of layers L is reduced from 12 to 10 to accelerate training and avoid overfitting.

This reduces the receptive field to 32,750 samples or 0.74 seconds. However we use the *Wave-U-Net* variant with extra input context, which allows the model to consistently learn temporal correlations and also reduces the output artifacts. The current receptive field together with the extra input context should be sufficiently large for drum mixing as the mixed output at a certain time point likely does not depend on stem activity occurring multiple seconds before or after.

Each input consist of 121,843 samples or 2.76 seconds and the output corresponds to the center part of the inputs and consists of 89,093 samples or 2.02 seconds. In order to obtain the prediction for a full drum mixture, we concatenate the output frames as non-overlapping segments.

We remove the tanh nonlinearity from the last convolution and instead allow all real-valued outputs during training. This is to avoid the need to saturate the nonlinearity with extreme values when outputting amplitudes close to +1 or −1. To ensure the outputs are between +1 and −1 at test time we simply clip the outputs accordingly. The rest of the convolutional layers are followed by the LeakyReLU activation function.

A block diagram can be seen in Fig. 1 and its structure is described in detail in Table 1. The *DS* blocks perform 1D convolutions of $F_c \cdot i$ filters of size $f_d = 15$, with layer $i \in [1, L]$, where $F_c = 24$ and corresponds to the number of initial filters. The resulting feature map is followed by a decimate operation that discards values for every other time step to reduce the time resolution by 50%.

The *US* blocks use linear interpolation to perform an upsampling of a factor of two and concatenate the resulting feature map with the cropped output of the respective *DS* block before decimation. This is followed by a 1D convolution of $F_c \cdot i$ filters of size $f_u = 5$, with $i \in [L, 1]$. All convolutions are along the time dimension without implicit padding and all strides are of unit value.

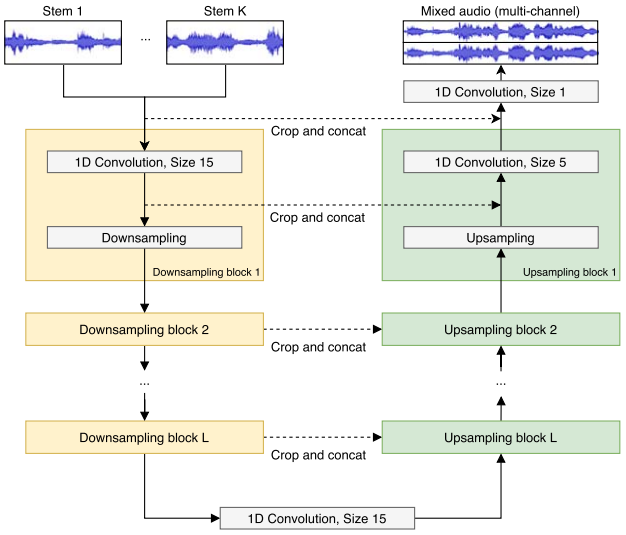

Fig. 1. Block diagram of the adapted *Wave-U-Net* for automatically mixing K stems using L layers.

Table 1. Detailed architecture of the modified *Wave-U-Net* with input and output frame sizes of 121843 and 89093 samples, respectively. DS_i corresponds to the output of the ith *DS* block before decimation.

Block	Operation	Shape
	Input	(121,843, 8)
DS, repeated for $i = 1, \ldots L$	Conv1D($F_c \cdot i, f_d$)	
	Decimate	
	Conv1D($F_c \cdot (L+1), f_d$)	(119, 264)
US, repeated for $i = L, \ldots 1$	Upsample	
	Crop and Concat(DS_i)	
	Conv1D($F_c \cdot i, f_u$)	
	Crop and Concat(Input)	
	Conv1D(K,1)	(89,093, 2)

Training is performed as in the original paper [7] but using the L1 instead of the L2 distance as training loss, based on previous observations with neural models that output raw audio obtaining perceptually more convincing results [24, 26]. *Adam* is used as the optimizer and we use an early stopping patience of 20 epochs followed by a fine-tuning step. The initial learning rate is 10^{-4} and the batch size is 16. We do not double the batch size in the fine-tuning stage because of memory requirements but still lower the learning rate to 10^{-5}. We select the model with the lowest loss for the validation subset. A full implementation and trained models can be found online.[1]

2.2 Random Forest-Based Approach

For the purposes of evaluation, we felt it appropriate to perform comparison with an existing *data-driven* automatic mixing approach. This is used to understand how our automatic mixing approach works, in comparison to other

[1] https://github.com/f90/Mix-Wave-U-Net

state-of-the-art approaches. As there is currently no existing end-to-end automatic mixing system published, we have combined a series of different approaches, where gain, compression, and reverberation are applied to the audio tracks in the mixing process.

For the gain balancing, we implement an alternative machine learning driven automatic mixing approach. The random forests mixing method, as presented by [35], was used for comparison. This approach uses a multiple estimator random forest regressor to predict the gain parameters used to mix each of the different tracks together. These gain predictions are based on reverse engineering of a mix [18] and provide the ground truth for training. For this case, the training and validation datasets were combined, as the validation set is not required, due to the bootstrap aggregation approach inherent within the random forest method. The objective results of this model on the training dataset are $R^2 = 0.984$, and on the combined test and validation dataset we achieve $R^2 = 0.915$. Once gains for each track were calculated the tracks were summed together to produced the *dry mix*.

To create the *wet mix*, dynamic range compression and reverberation are applied once the individual tracks are mixed. Both of these approaches use a *knowledge-engineering* approach, where effect parameters were automated based on signal analysis. This was done as there are no data-driven approaches to apply the dynamic range compression or reverberation to audio. The compression uses the work described in [36] and automatic reverberation is applied using the approach given in [37]. The compressor is specifically designed for application to drums with the aim of emphasizing transients.

The timing parameters are often considered the most important settings on a compressor [38]. A beat tracker is used in combination with a tail decay envelope follower to identify the appropriate attack and release times on the compressor. The threshold of the compressor was set to the Root Mean Square of the signal and the ratio was set as a function of the crest factor.

Reverberation is applied to give a spatial balance. An algorithmic reverberator is taken [39] and the control parameters automated based on the a list of known mixing rules [40]. The tempo of the track is extracted and used to control the diffusion and tail decay parameters of the reverb, based on a mapping between the tempo and RT60 of a musical piece [41, 42]. Pre-delay is calculated as a function of the tempo and the Haas fusion point [43] while reverb gain and damping are both calculated as functions of the transient nature of the audio signal.

In both the compressor and the reverb cases, these are signal processing and audio feature derived approaches to automate the effect based on known mixing rules, and as such, it is expected that these approaches will produce greatly different results to the proposed machine learning methods. It is hypothesized that the signal processing derived mixing methods will under-perform when compared to the machine learning approaches, as these approaches rely on manual interpretation of mixing approaches rather than an analysis of data to reproduce a mix—which, it is expected, will be simpler in the case of the drum stem, as presented in this work.

Table 2. Distribution of the number of drum recording types into training, validation, and test set.

	Training	Validation	Test	Total
Hits	96	7	5	103
Phrase	122	7	6	129
Solo	8	2	1	10
Minus-One	58	4	2	62
Total	284	20	14	304
Percentage	89.3%	6.3%	4.4%	100%

2.3 Dataset

The ENST drum dataset [44] is a dataset that contains multitrack drum recordings and two human expert-made stereo mixes for each drum track; the *dry* mix, which consists of panning and loudness gain; and the *wet* mix, which consists of the aforementioned effects, plus EQ, compression, artificial reverberation, and mastering DRC.

This dataset includes recordings of three different drummers and three different drum kits, playing a variety of different musical styles and playing techniques. In each case, the multitrack recordings are obtained with 7 or 8 mono microphones: bass drum, snare drum, hi-hat, mid tom, low tom, mid-low tom (if available), left overhead, and right overhead. We use a silence signal of zeros as the input track when the mid-low tom is not available.

The ENST dataset is arranged in four different types of recording: hits; phrase; solo; and accompaniment. Each track is between 7 seconds and 84 seconds long, with a median duration of 19.8 seconds. The total duration of audio material is around 225 minutes.

hits The drummer hits a single drum a number of times with a single type of stick

phrase The drummer performs a short excerpt in a requested style and then at a range of different tempo and complexity levels, determining whether fills would be included.

solo A specific phrase designed to last around 30 seconds, in which the drummer is free to perform throughout the entire drum kit as they see fit.

accompaniment or minus-one The drummer performs along with either a CD or generated MIDI file, playing to a strict rhythm.

All tracks in the dataset were used and split between training, validation, and test, as described in Table 2. No data augmentation techniques were used to increase the data size.

3 EXPERIMENTS

The training procedures were performed for each method and mixing task. Then the models were tested with sam-

ples from the test subset, from which resulting mixes were evaluated using a subjective listening experiment.

3.1 Participants

Twenty participants were recruited as volunteer staff and students from the University of Plymouth and University of London. All participants reported that they trusted their ears and had some experience with critical listening. There were no financial incentives provided. Six participants identified as female and fourteen participants identified as male. No other genders were identified by the participants. The mean age of participants was 26.5, with a standard deviation of 8.48, and all participants were over 18 years old. No test took longer than 38 minutes, so fatigue was not an issue [45].

3.2 Setup and Procedure

Participants were asked to complete an online listening experiment, which was constructed using the Web Audio Evaluation Tool [46]. Participants were asked to use a web browser to access the online experiment, where they were asked to use a good pair of quality headphones to complete the experiment. The headphone type and listening conditions were self reported and any poor quality budget headphones or excessively noisy environments were removed from the listening experiment.

All the different mixes of a single drum loop were presented on the same screen, and participants had to finish evaluating all samples before moving to the next screen. The interface is shown in Fig. 2, where each vertical bar can be selected to play back an audio sample. The participants were asked to rate each sample based on their preference and within a continuous scale. The experiment conducted was MUSHRA inspired [47]; however we used a slightly different protocol to encourage direct comparison between audio samples rather than comparison to a reference and to ensure that participants select the preferred mix. This approach allows participants to present results in a specific order, which is appropriate for the intended analysis, which is the Mann-Whitney Rank Sum test. This follows a methodology similar to that of other subjective evaluation experiments [48, 49].

Participants were initially asked to set the volume to a comfortable level and refrain from adjusting the device volume level for the duration of the experiment, instead using the in-experiment volume control. No participant changed the volume by more than 6 dB using the in-experiment volume control.

3.3 Materials

Participants were asked to rate a series of different mixes of drum loops on preference. There were six different drum loops from the test dataset, which were presented one at a time. In each case, there were seven different mixes of each drum loop, all presented together, and participants were asked to inter-compare the samples presented and rank on preference.

The seven different mixes were as follows—two mixed by a human professional engineer, as the *reference* samples, both provided from the ENST Drum dataset [44]. Two were produced by our deep learning approach presented in Sec. 2.1, which will be named the *Wave-U-Net* approach. Two were produced by the *Random Forest* approach presented in Sec. 2.2, and the final sample was the hidden *anchor*. The latter makes use of two overhead microphones, with hard panning, and a 3.5 kHz low-pass filter, as per the MUSHRA standard [50]. Two samples are each presented for the reference, *Wave-U-Net* and *Random Forest*, since one represents a *dry* and the other one a *wet* mix.

The drum loops to be evaluated were taken from the *phrases*, *solo*, and *minus-one* sample groupings, as the ability to mix the *hits* grouping was not considered to be as important. The tracks used were also selected to ensure that there were two examples from each drummer, and each drum setup, to evaluate generalizability. All samples were loudness normalized to -28 dBFS, in accordance with ITU-R BS.177-2 [51]. All samples were played back at 44.1 kHz, the native sample rate of the tracks. The order of the different drum loops and the individual track placement and naming was randomized to remove bias. All audio samples used for evaluation can be found online.[2]

4 RESULTS

In this section, we will present the results of our listening test and analyze how well our proposed approach compares to the baseline approach as well as the reference mixes.

4.1 Quantitative

Fig. 3 shows the violin plot of the results. Firstly, the references and anchors were chosen appropriately, since they received extremely high and low ratings, respectively, supporting the validity of our experimental results. Overall our *Wave-U-Net* approach performs very well, producing ratings similar to those given to the reference mixes, with small differences between the dry and wet tasks. The performance of the *Random Forest* approach is substantially lower and strongly depends on whether the wet or dry mixing task is considered. The low ratings for the wet task are due to the *Random Forest* method being a combination of previous approaches for automatic gain, DRC, and reverb. These methods have been shown to work well only in isolation, and based on our results, a further exploration of the parameters of each method is required.

To determine the statistical significance of these differences between approaches, we perform a series of statistical tests. Firstly, the Shapiro-Wilk test is used to check for normality and homogeneity, with the null hypothesis being that the data came from a normally distributed population. Table 3 presents the results of the Shapiro-Wilk test and as can be seen, for all mix types, the results are statistically significant ($\alpha = 0.001$), the null hypothesis is rejected in all cases, and the data are considered to be non-normal, confirmed

[2] https://mchijmma.github.io/drum-mixing-wave-u-net/

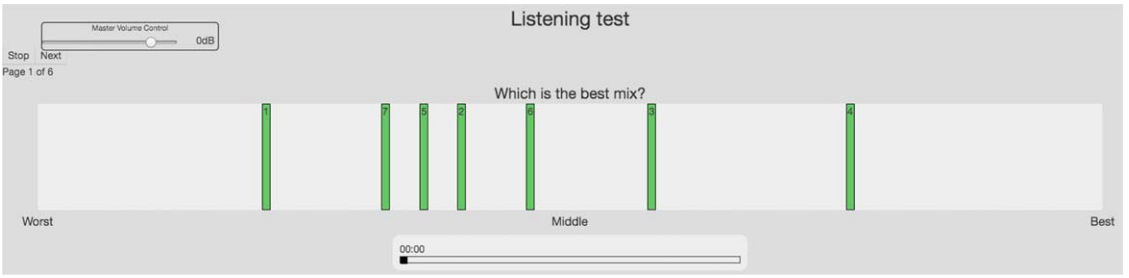

Fig. 2. Screenshot of the listening experiment interface.

Table 3. Shapiro-Wilk test for normality.

Mix	p	W
Anchor	2.22×10^{-14}	0.698
Random Forest Dry	1.48×10^{-5}	0.933
Random Forest Wet	7.66×10^{-19}	0.475
Reference Dry	6.22×10^{-5}	0.942
Reference Wet	1.63×10^{-4}	0.948
Wave-U-Net Dry	2.16×10^{-4}	0.950
Wave-U-Net Wet	3.82×10^{-4}	0.953

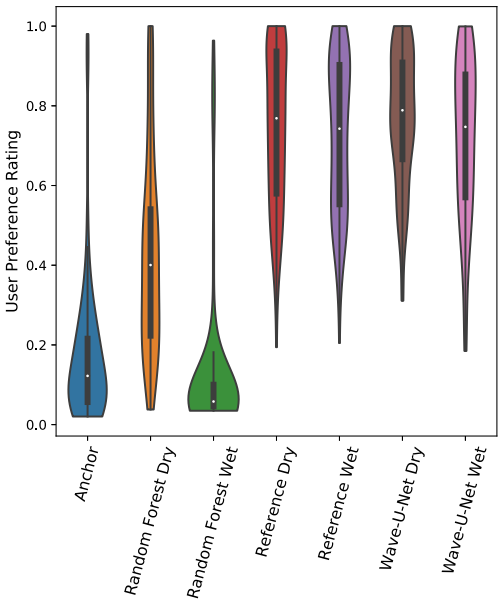

Fig. 3. Violin plot of preference ratings.

by visual inspection of Fig. 3. As such, non-parameteric analysis of the data is required.

A Kruskal-Wallis test is then conducted based on the non-parametric nature of the data to identify whether a statistically significant difference exists in the user preferences between each of the different mixing approaches ($H =$ 484.9921423799316, $p = 1.4363346737459683e - 101$). The Kruskal-Wallis test showed that the null hypothesis, of all mixing approaches deriving from the same distribution, was rejected, as $p < \alpha$. There is a significant difference between the preference rating between each of the different mixing approaches. To identify exactly which groups are significantly different, a post hoc pairwise Mann-Whitney test is performed, with Bonferroni correction. The results of this are presented in Table 4.

Table 4. Post hoc Mann-Whitney test results of pairwise comparison of mix creation method on preference rating, with Bonferroni Correction. o ≥ 0.05, * < 0.01, *** < 0.0001, · (no comparison).

Mix	Anchor	Random Forest		Reference		Wave-U-Net	
		Dry	Wet	Dry	Wet	Dry	Wet
Anchor	.	***	*	***	***	***	***
Random Forest Dry	***	.	***	***	***	***	***
Random Forest Wet	*	***	.	***	***	***	***
Reference Dry	***	***	***	.	o	o	o
Reference Wet	***	***	***	o	·	o	o
Wave-U-Net Dry	***	***	***	o	o	·	o
Wave-U-Net Wet	***	***	***	o	o	o	·

Through inspection of Fig. 3 and Table 4 it can be seen that the anchor and each of the *Random Forest* approaches are significantly different from all other approaches, including each other. It should also be noted that there is no significant difference in the user ratings for the *Wave-U-Net* mixed approaches and the reference human-made mixes. Furthermore the *Wave-U-Net* mixes have very similar median values to the reference audio samples.

4.2 Qualitative

For both dry and wet mixing tasks and from the test subset, Fig. 4 shows the waveform and spectrogram of selected mixes. From the spectrograms, it can be seen that the *Random Forest* wet mix highly diverges from reference, hence the reported poor perceptual ratings. This performance is due to the model adding large amounts of artificial reverberation. From the waveforms, the *Random Forest* dry model produces a better fit than the *Random Forest* wet model; thus it applies gain and panning more effectively.

However, dry and wet *Wave-U-Net* mixes, both in the time and frequency domains, are indistinguishable from the reference samples. The latter is displayed in greater detail in Fig. 5, where a segment of the wet mixes in Fig. 4(c) is shown. The segment corresponds to the onset transient of a bass drum and, as expected, the reference and *Wave-U-Net* mixes are almost identical sample by sample. In particular, the most noticeable effects properly applied by the *Wave-U-Net* models are gain, panning, EQ, and DRC. Empirically, the wet mixes produced by the model seldom

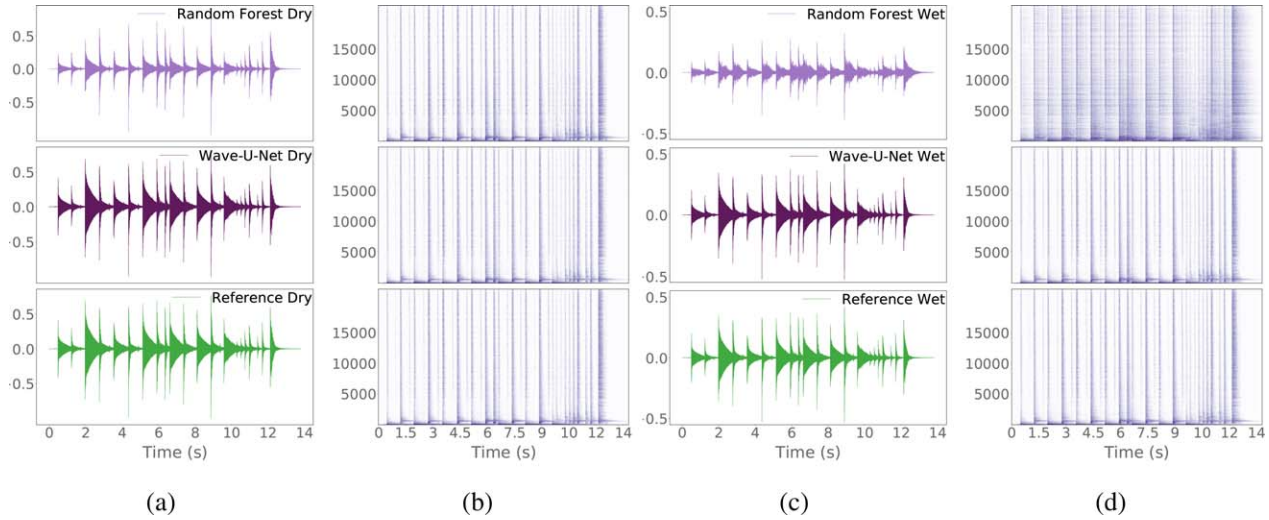

Fig. 4. Results with a selected *phrase* mix from the test dataset. Figs. 4(a) and 4(c) show the mono waveforms of the dry and wet mixes, respectively. Vertical axes represent amplitude. Figs. 4(b) and 4(d) show their respective spectrograms; color intensity represents higher magnitude and vertical axes represent frequency (Hz).

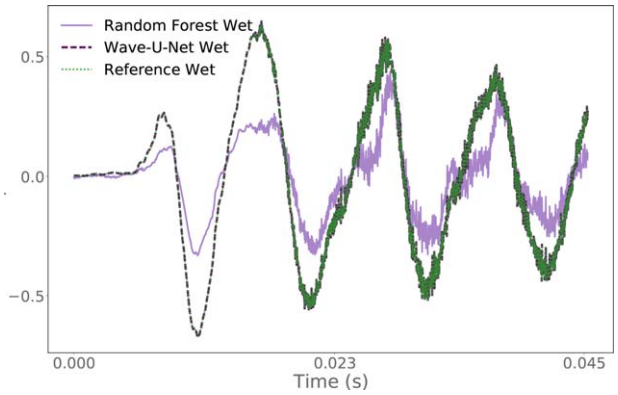

Fig. 5. Wet mixes from the drum mixes of Fig. 4(c), a segment that corresponds to the onset of a bass drum. Vertical axis represents amplitude.

lack some reverb that is present in the reference mixes, although additional testing is required.

5 CONCLUSION

The results in Sec. 4 clearly show that the *Wave-U-Net* approach performed significantly better than the alternative automatic mixing approach and that the user preference ratings are not significantly different from that of a human-made mix. As such it can be deemed that the *Wave-U-Net* mixing approach is indistinguishable from a human mixer in terms of preference rating. A user would not be able to state that the human or computer made mix is better using this system.

We believe that this is the first time where an automatic mixing system is able to produce a mix that is not significantly different from a human-made mix in terms of user preference. Not only that, but this system has the capacity to learn not just a single audio effect or predefined audio effect chain but to directly learn the audio transformation and all signal processing involved in producing a musical mix. Thus the ability learn and apply all audio effects simultaneously is highly beneficial to the mixing process, as it acknowledges the fact that all aspects and parameters of a mix are inherently intertwined and interrelated [52].

It is clear from this demonstration that neural network approaches to mixing can produce impressive results. These results were due to the mix being performed specifically on drum content. The latter consists of a well-structured dataset, with a relatively consistent set of input channels, channel ordering, and source content. Thus, even with the variation over three different recording setups, three different drummers, and a range of different drumming styles, there is clear need for further development and analysis of this approach to investigate how generalizable our results are to other mixing contexts.

We have also demonstrated that a musical mixing approach, where there is a focus on mixing based on subgroups and specific audio content, could lead to much more effective intelligent mixing systems than developing intelligent mixing tools making use of specific audio effects.

5.1 Further Work

There are a range of developments that could be utilized to further improve this intelligent mixing system. Our experiments are limited to drum mixing in particular, as this problem is well defined and mixing is usually relatively consistent and occurs in a large number of styles and genres of music. However it is not known how well our approach performs for similar mixing tasks. Additionally the input to the network is static—a fixed number of stems needs to be provided in a certain order. Future developments could aim to make the network invariant to the order of the different inputs and still produce suitable mixes even if some inputs are missing.

This more generalizable approach would then allow for an extension to musical content with varying instrumentation beyond just drums. In addition, further exploration of model performance for each type of drum kit could also be investigated.

The musical mixes were all produced by a single individual, who mixed the tracks to fit in with a given musical content. The musical context of the mix will be very important to the system, so perhaps improvements could be provided with an associated musical context to fit the mix into. Furthermore obtaining multiple mixes per set of stems would allow probabilistic modeling of the potential space of effects that could be applied in each case, rather than predicting a single mix.

We did not use any preprocessing for the input recordings or target stereo mixes, as throughout the dataset the mixing and recording engineer maintained similar amplitude levels for each type of input and mix. Therefore a robustness analysis of the model is required for inputs with different volume levels. A further approach to improve the results could be to apply data augmentation techniques to make the end result invariant to preprocessing on any or all audio tracks. Tackling the above challenges and the availability of a more general, large dataset of annotated multitrack audio would certainly facilitate a more generalizable intelligent music production approach. The standardization of these datasets, with associated musical mixes, both human and computer-produced mixes, would greatly aid in the development of larger and more impressive intelligent mixing systems.

To that end, there are a number of technical improvements that could be made to this work. The use of a more appropriate loss function that contains more perceptually relevant parameters could be advantageous. Equipping the model with some additional control parameters so the user can shape certain aspects of the generated musical mix would also be beneficial. As future work, different mixing tasks such as automatic drum time-alignment and removal of drum bleed could be explored.

6 REFERENCES

[1] P. Pestana and J. Reiss, "Intelligent Audio Production Strategies Informed by Best Practices," in *Proceedings of the AES 53rd International Conference: Semantic Audio* (2014 Jan.), paper S2-2.

[2] D. Moffat and M. B. Sandler, "Approaches in Intelligent Music Production," *Arts*, vol. 8, no. 4, p. 14 (2019 Sep.). https://doi.org/10.3390/arts8040125.

[3] B. De Man, J. D. Reiss, and R. Stables, "Ten Years of Automatic Mixing," in *Proceedings of the 3rd Workshop on Intelligent Music Production* (2017 Sep.).

[4] T. Wilmering, D. Moffat, A. Milo, and M. B. Sandler, "A History of Audio Effects," *Appl. Sci.*, vol. 10, no. 3, p. 791 (2020 Jan.). https://doi.org/10.3390/app10030791.

[5] D. Ronan, H. Gunes, D. Moffat, and J. D. Reiss, "Automatic Subgrouping of Multitrack Audio," in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, pp. 1–8 (Trondheim, Norway) (2015 Nov.).

[6] D. Ronan, B. De Man, H. Gunes, and J. D. Reiss, "The Impact of Subgrouping Practices on the Perception of Multitrack Music Mixes," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9442.

[7] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018)* (2018 Jun.).

[8] F. Pachet and O. Delerue, "On-the-Fly Multi-Track Mixing," presented at the *109th Convention of the Audio Engineering Society* (2000 Sep.), paper 5255.

[9] G. Bocko, M. F. Bocko, D. Headlam, J. Lundberg, and G. Ren, "Automatic Music Production System Employing Probabilistic Expert Systems," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), paper 8255.

[10] B. De Man and J. D. Reiss, "A Knowledge-Engineered Autonomous Mixing System," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), paper 8961.

[11] F. Everardo, "Towards an Automated Multitrack Mixing Tool Using Answer Set Programming," in *Proceedings of the 14th Sound and Music Computing Conference* (2017 Jul.).

[12] D. Moffat, F. Thalmann, and M. B. Sandler, "Towards a Semantic Web Representation and Application of Audio Mixing Rules," in *Proceedings of the 4th Workshop on Intelligent Music Production (WIMP)* (2018 Sep.).

[13] D. Matz, E. Cano, and J. Abeßer, "New Sonorities for Early Jazz Recordings Using Sound Source Separation and Automatic Mixing Tools," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 749–755 (Málaga, Spain) (2015 Oct.).

[14] R. Stables B. De Man, and J. D. Reiss, *Intelligent Music Production* (Focal Press, Waltham, MA, 2019).

[15] M. A. Martínez Ramírez and J. D. Reiss, "Deep Learning and Intelligent Audio Mixing," in *Proceedings of the 3rd Workshop on Intelligent Music Production* (2017 Sep.).

[16] S. I. Mimilakis, E. Cano, J. Abeßer, and G. Schuller, "New Sonorities for Jazz Recordings: Separation and Mixing Using Deep Neural Networks," in *Proceedings of the 2nd AES Workshop on Intelligent Music Production (WIMP)* (2016 Sep.).

[17] D. Moffat and M. Sandler, "Automatic Mixing Level Balancing Enhanced Through Source Interference Identification," presented at the *146th Convention of the Audio Engineering Society* (2019 Mar.), paper 497.

[18] D. Barchiesi and J. Reiss, "Reverse Engineering of a Mix," *J. Audio Eng. Soc.*, vol. 58, no. 7/8, pp. 563–576 (2010 Jul.).

[19] F. -R. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *Proceedings of the International Conference on Latent Variable Analysis*

and Signal Separation (2018 Jul.). https://doi.org/10.1007/978-3-319-93764-9_28.

[20] P. -S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012 Mar.). https://doi.org/10.1109/ICASSP.2012.6287816.

[21] P. -S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-Voice Separation From Monaural Recordings Using Deep Recurrent Neural Networks," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)* (2014 Oct.).

[22] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A Joint Separation-Classification Model for Sound Event Detection of Weakly Labelled Data," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8462448.

[23] F. -R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A Reference Implementation for Music Source Separation," *J. Open Source Software*, vol. 4, no. 41, p. 1667 (2019 Sep.). https://doi.org/10.21105/joss.01667.

[24] 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, p. 638 (2020 Jan.). https://doi.org/10.3390/app10020638.

[25] M. A. Martínez Ramírez, E. Benetos, and J. D. Reiss, "Modeling Plate and Spring Reverberation Using a DSP-Informed Deep Neural Network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020 May). https://doi.org/10.1109/ICASSP40776.2020.9053093.

[26] M. A. Martínez Ramírez and J. D. Reiss, "Modeling Nonlinear Audio Effects With End-to-End Deep Neural Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019 May). https://doi.org/10.1109/ICASSP.2019.8683529.

[27] M. A. Martínez Ramírez and J. Reiss, "End-to-End Equalization With Convolutional Neural Networks," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18)*, pp. 296–303 (Aveiro, Portugal) (2018 Sep.).

[28] A. Wright, E. -P. Damskägg, L. Juvela, and V. Välimäki, "Real-Time Guitar Amplifier Emulation With Deep Learning," *Appl. Sci.*, vol. 10, no. 3, p. 766 (2020 Jan.). https://doi.org/10.3390/app10030766.

[29] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499* (2016).

[30] S. H. 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.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241 (Munich, Germany) (2015 Oct.). https://doi.org/10.1007/978-3-319-24574-4_28.

[32] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, "Time-Frequency Networks for Audio Super-Resolution," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8462049.

[33] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proceedings of INTERSPEECH 2017*, pp. 3642–3646 (Stockholm, Sweden) (2017 Aug.). https://doi.org/10.21437/Interspeech.2017-1428.

[34] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8462417.

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

[36] D. Moffat and M. Sandler, "Adaptive Ballistics Control of Dynamic Range Compression for Percussive Tracks," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 484.

[37] D. Moffat and M. Sandler, "An Automated Approach to the Application of Reverberation," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10264.

[38] G. Bromham, D. Moffat, M. Barthet, and G. Fazekas, "The Impact of Compressor Ballistics on the Perceived Style of Music," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 10080.

[39] J. Dattorro, "Effect Design, Part 1: Reverberator and Other Filters," *J. Audio Eng. Soc.*, vol. 45, no. 9, pp. 660–684 (1997 Sep.).

[40] P. D. L. G. Pestana, *Automatic Mixing Systems Using Adaptive Digital Audio Effects*, Ph.D. thesis, Universidade Católica Portuguesa, Porto, Portugal (2013 Feb.).

[41] J. Weaver, M. Barthet, and E. Chew, "Analysis of Piano Duo Tempo Changes in Varying Convolution Reverberation Conditions," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 10108.

[42] J. Weaver, M. Barthet, and E. Chew, "Filling the Space: The Impact of Convolution Reverberation Time on Note Duration and Velocity in Duet Performance," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10266.

[43] H. Haas, "The Influence of a Single Echo on the Audibility of Speech," *J. Audio Eng. Soc.*, vol. 20, no. 2, pp. 146–159 (1972 Mar.).

[44] O. Gillet and G. Richard, "ENST-Drums: An Extensive Audio-Visual Database for Drum Signals Processing," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)* (2006 Oct.).

[45] R. Schatz, S. Egger, and K. Masuch, "The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings," *J. Audio Eng. Soc.*, vol. 60, no. 1/2, pp. 63–73 (2012 Jan.).

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

[47] ITU-R, "BS 1534-1, Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," Technical Report (2001).

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

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

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

[51] ITU-R, "Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," *Recommendation BS.1770-2* (2011 Mar.).

[52] M. N. Lefford, G. Bromham, and D. Moffat, "Mixing With Intelligent Mixing Systems: Evolving Practices and Lessons From Computer Assisted Design," presented at the *148th Convention of the Audio Engineering Society* (2020 May), paper 10376.

THE AUTHORS

Marco A. Martínez Ramírez

Daniel Stoller

David Moffat

Marco A. Martínez Ramírez is a music technology researcher at Sony in the Tokyo R&D Center, where he is a member of the AI & Music Technology Group. Previously Marco was an audio research intern at Adobe and received his PhD from Queen Mary University of London. His research interests lie at the intersection of machine learning, digital signal processing, and intelligent music production, where his primary focus of research is on deep learning architectures for music and audio processing. Marco is also a music producer and mixing engineer.

•

Daniel Stoller received his PhD from Queen Mary University of London. As part of the Centre for Digital Music, he developed methods to improve the generalization capability of deep learning methods when faced with little training data. Applications include a variety of audio and music processing tasks, such as audio source separation and singing voice detection.

•

Dr. David Moffat is a Lecturer in Sound and Music Computing at the University of Plymouth. He received his PhD and MSc from Queen Mary University. His research focuses on intelligent and assistive mixing and audio production tools through the implementation of semantic tools and machine learning. David has been a member of the Audio Engineering Society since 2014 and is a member of the AES UK committee and Vice Chair of the Technical Committee on Semantic Audio Analysis.