# Reverse engineering of a recording mix with differentiable digital signal processing

Joseph T. Colonel and Joshua Reiss

Citation: The Journal of the Acoustical Society of America **150**, 608 (2021); doi: 10.1121/10.0005622
View online: https://doi.org/10.1121/10.0005622
View Table of Contents: https://asa.scitation.org/toc/jas/150/1
Published by the Acoustical Society of America

JASA ARTICLE

# Reverse engineering of a recording mix with differentiable digital signal processing[a)]

Joseph T. Colonel[b)] and Joshua Reiss
*Centre for Digital Music, Queen Mary University of London, London, United Kingdom*

**ABSTRACT:**
A method to retrieve the parameters used to create a multitrack mix using only raw tracks and the stereo mixdown is presented. This method is able to model linear time-invariant effects such as gain, pan, equalisation, delay, and reverb. Nonlinear effects, such as distortion and compression, are not considered in this work. The optimization procedure used is the stochastic gradient descent with the aid of differentiable digital signal processing modules. This method allows for a fully interpretable representation of the mixing signal chain by explicitly modelling the audio effects rather than using differentiable blackbox modules. Two reverb module architectures are proposed, a "stereo reverb" model and an "individual reverb" model, and each is discussed. Objective feature measures are taken of the outputs of the two architectures when tasked with estimating a target mix and compared against a stereo gain mix baseline. A listening study is performed to measure how closely the two architectures can perceptually match a reference mix when compared to a stereo gain mix. Results show that the stereo reverb model performs best on objective measures and there is no statistically significant difference between the participants' perception of the stereo reverb model and reference mixes.  https://doi.org/10.1121/10.0005622
(Received 1 February 2021; revised 24 May 2021; accepted 24 June 2021; published online 27 July 2021)
[Editor: Peter Gerstoft] Pages: 608–619

## I. INTRODUCTION

Multitrack mixing refers to the process of a mixing engineer combining the elements of a song, audio piece, or recording into a "mixdown." In the case of a band recorded in a studio, for example, a mixing engineer will collate the band's various recordings, prepare and align those tracks, and apply sound effects and other processing to those tracks to create a stereo mix. It is up to the mixing engineer to ensure that the sonic elements of the composition are fixed, fitted, and featured properly in accordance with the band's artistic vision.

Mixing engineers often will use both analog and digital processing to create their mixdown. Frequently these include gain, pan, equalisation (EQ), delay, reverb, distortion, and dynamic range compression. In the absence of automation, these first five effects are often implemented as linear time-invariant processing. The latter two effects are nonlinear.

As is now often the case, a mixing engineer will use a session in a digital audio workstation (DAW) to create a mixdown. This session contains most if not all of the information regarding what effects and processing were applied to each track in the multitrack recording. In the absence of this session, it is very difficult to recreate the processing used to create a specific mixdown, a process that is referred to as reverse engineering the mix. Even in the case in which the original session is available, sharing sessions can prove difficult as sessions cannot be shared across different DAWs or sometimes across versions of the same DAW. Furthermore, the session may use specific (and often costly) effect plugins unavailable to another mixing engineer, which could make the entire session unusable should the first mixing engineer employ complicated signal chains. This reverse engineering problem may be phrased as, "Given a set of raw multitrack recordings and a mixdown, how can one derive all the effects and their parameters that were used to produce the mixdown?"

## II. BACKGROUND

The problem of reverse engineering a mix remains relatively niche in the field of machine learning for audio. However, much attention in the field has been paid to recovering an audio signal in a blind manner after some processing has been applied. Such problems include denoising (Yu *et al.*, 2008; Grais and Plumbley, 2017), dereverberation (Lebart *et al.*, 2001; Feng *et al.*, 2014), and source separation (Belouchrani *et al.*, 1997; Stoller *et al.*, 2018).

In Jourjine *et al.* (2000), the authors reframe a source separation problem as a reverse engineering of a mix. Given that $N$ sources are disjoint-orthogonal and anechoic, the method can estimate the attenuation and delay applied to each source from a two channel mix. Unfortunately, the disjoint-orthogonal and anechoic assumptions rarely hold in the case of a multitrack recording and mixdown given that multiple sources may bleed into several tracks and delay and reverb are echoic by definition.

[a)]This paper is part of a special issue on Machine Learning in Acoustics.
[b)]Electronic mail: aj.t.colonel@qmul.ac.uk

https://doi.org/10.1121/10.0005622

In Gorlow and Marchand (2013), a two stage cascaded encoder and decoder are used to estimate the compression, gain, and panning effects applied to create a mixdown. This is performed using only the spectrogram of the mixdown, i.e., without the raw multitrack. The method requires fixing the number of tracks used to create the mixdown and does not attempt to model the EQ or reverb. Without access to the raw multitracks, this problem statement and the problem statement by Jourjine *et al.* (2000) land closer to source separation or upmixing than to reverse engineering a mix.

A considerabe amount of work has been published on modelling individual sound effects and processing. Presented in Gorlow and Reiss (2013) is a method for estimating the parameters of a dynamic range compressor applied to a signal. More recently, blackbox modelling of audio effects using neural networks has received attention. Examples can be found for modelling the EQ (Ramírez and Reiss, 2018), distortion (Ramírez and Reiss, 2019), compression (Hawley *et al.*, 2019), and other nonlinear effects (Ramírez *et al.*, 2020). The blackbox approach outlined in Steinmetz *et al.* (2020) simultaneously models the EQ, compression, and reverb and can be controlled by passing parameters relating to each effect to the blackbox.

The methods in Barchiesi and Reiss (2010) explicitly try to answer this reverse engineering of a mix question and inform the primary motivations of this paper. In Barchiesi and Reiss (2010), a method for reverse engineering a mix is presented, which combines separate modelling of the linear processing and nonlinear processing used to create a mixdown. This method uses both the mixdown and raw multitrack. The nonlinear processing is estimated using a frame-based approach in which the dynamic range compression is modelled as a time-varying gain envelope. A time domain least squares approach is used to model the linear processing, including gain, pan, delay, and EQ. This method theoretically holds when estimating a convolutional reverb impulse response, but the length of these impulse responses makes a least squares estimate impractical.

The improvements to Barchiesi and Reiss (2010) that are presented in this work combine that work's explicit modelling of the mixing chain with the capabilities of modern neural networks to model complex audio effects.

A neural network whitebox approach to audio synthesis and sound effect modelling, called differentiable digital signal processing (DDSP), has been developed in Engel *et al.* (2019). The DDSP toolbox contains modules often found in sound synthesis, such as harmonic oscillators and subtractive synthesizers, that can be integrated into neural networks. The toolbox also contains a convolutional reverb module and finite impulse response (FIR) EQ module.

The method presented in this paper expands on the linear processing modelling of Barchiesi and Reiss (2010) with the aid of modules developed in Engel *et al.* (2019) but does not incorporate nonlinear processing such as distortion or compression. Whereas these effects are often used by mixing engineers, there is currently no published literature on whitebox neural network implementations of these effects. By using explicit whitebox modelling of linear audio effects, the algorithm presented here produces a fully characterized, interpretable signal chain that could be further modified upon inspection.

The paper is organized as follows. Section III will outline the architecture and optimization procedure used to reverse engineer a mix using EQ, gain, pan, delay, and reverb. Section IV will show the outputs of the reverse engineering algorithm as well as the result of a listening study conducted to evaluate the quality of the reverse engineered mixdowns. Section V will discuss the results and Sec. VI will conclude with remarks and directions for future work.

## III. METHOD AND THEORY

### A. Formal problem statement

Let $y(n)$ represent a target mixdown, and let $\hat{y}(n)$ represent the mixdown produced by some mixing chain characterized by a set of parameters $\Theta$. The goal is to find values that correspond to the parameter settings in a mixing chain that will minimize $||y(n) - \hat{y}(n)||$, where $||\cdot||$ denotes some norm that will be used as a cost function.

The signal processing chain applied to each input raw track is as follows: dry input $\rightarrow$ FIR EQ $\rightarrow$ gain $\rightarrow$ pan $\rightarrow$ reverb and wet/dry mix $\rightarrow$ sum with other stems. Note that because these effects are all linear time-invariant, the order of application of the effects is arbitrary. In mix engineering, a "stem" refers to a raw track that has had processing applied to it. To drive each module, a set of parameters $\Theta_{\text{module}}$ are estimated. For example, $\Theta_{\text{EQ}}$ refers to the set of parameters estimated to drive the EQ module. A stem which has been processed by applying both the EQ and gain to a raw track $x(n)$ can be written as

$$\text{stem}_{\text{EQ,Gain}}(n) = \text{Gain}(\text{EQ}(x(n)|\Theta_{\text{EQ}})|\Theta_{\text{Gain}}), \tag{1}$$

and a stereo mixdown of $N$ raw tracks $x_i(n)$ with EQ, gain, pan, and reverb applied can be written as

$$\begin{aligned}\hat{y}_L(n) &= \sum_{i=1}^{N} \text{Reverb}(\text{Pan}_L(\text{Gain}(\text{EQ}(x(n)_i|\Theta_{\text{EQ}})|\Theta_{\text{Gain}})|\Theta_{\text{Pan}})|\Theta_{\text{Reverb}}),\\ \hat{y}_R(n) &= \sum_{i=1}^{N} \text{Reverb}(\text{Pan}_R(\text{Gain}(\text{EQ}(x(n)_i|\Theta_{\text{EQ}})|\Theta_{\text{Gain}})|\Theta_{\text{Pan}})|\Theta_{\text{Reverb}}).\end{aligned} \tag{2}$$

### B. EQ

The frequency transfer curve module is used for EQ by multiplying an input signal's short-term Fourier transform (STFT) magnitude response with a user specified curve in the frequency domain (Engel *et al.*, 2019). In this work, a 1025 point frequency transfer curve $\Theta_{\text{EQ}}$ is used. This corresponds to a FIR EQ with 2048 taps in its impulse response. Given a raw track $x(n)$, the EQ module can be written as

$$\text{EQ}(x(n)|\Theta_{\text{EQ}}) = \text{ISTFT}\big(\text{STFT}(x(n)) \times \Theta_{\text{EQ}}\big), \tag{3}$$

where ISTFT is the inverse short-time Fourier transform and "×" refers to pointwise multiplication.

In this work, the EQ is modelled after a ten band FIR graphical EQ (Välimäki and Reiss, 2016), which can be characterized using a ten-dimensional $\Theta_{\text{EQ gains}}$. The ten values specify the gain of each of the octave band filters, which are centered at 30, 60, 125, 250, 500, 1000, 2000, 4000, 8000, and 16 000 Hz respectively. Shelving filters are used for frequencies below 30 Hz and above 16 000 Hz that match the attenuation specified at the lowest and highest octave band, respectively.

The following procedure is used to calculate the 1025-dimensional $\Theta_{\text{EQ}}$ that will approximate a ten band FIR graphical EQ. First a ten-dimensional $\Theta_{\text{EQ gains}}$ is generated. Then, these values are transformed via

$$\Theta_{\text{EQ gains}} \leftarrow 1 - \sigma(\Theta_{\text{EQ gains}}), \tag{4}$$

where $\sigma$ denotes the sigmoid function

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{5}$$

The values in the transformed $\Theta_{\text{EQ gains}}$ range from (0,1) due to the bounds of the sigmoid function.

Finally, a piecewise linear frequency transfer curve $\Theta_{\text{EQ}}$ is constructed using linear interpolation between the octave band attenuations specified by $\Theta_{\text{EQ gains}}$. Thus, the EQ module's frequency transfer curve is bounded from (0,1) at all points The estimated values are initialized with the random uniform noise from [–1,1], which initializes the octave band gains from −6 to −1 dB.

### C. Gain and pan

The gain module is formulated as

$$\text{G}(x(n)|\Theta_{\text{Gain}}) = \text{LReLU}(\Theta_{\text{Gain}}) \times x(n), \tag{6}$$

where LReLU( ) refers to the leaky rectified linear unit (Maas *et al.*, 2013)

$$\text{LReLU}(x) = \begin{cases} \beta x, & x < 0 \\ x, & x \geq 0, \end{cases} \tag{7}$$

with tunable parameter $\beta$. For this work, $\beta = 0.5$ has been chosen. Note that these gains can go negative, which corresponds with applying a phase shift to the equalised stem. The gain parameters $\Theta_{\text{Gain}}$ are initialized with random uniform noise from [0.9,1.1], which corresponds to gains from −0.915 to 0.828 dB.

The pan module uses a linear panning law and is formulated as

$$\begin{aligned} \text{Pan}_L(x(n)|\Theta_{\text{Pan}}) &= (0.5 + (0.5 \times \tanh(\Theta_{\text{Pan}}))) \times x(n), \\ \text{Pan}_R(x(n)|\Theta_{\text{Pan}}) &= (1 - \text{Pan}_L) \times x(n), \end{aligned} \tag{8}$$

where "×" denotes pointwise multiplication and tanh( ) denotes the hyperbolic tangent function

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{9}$$

The panning module applies a gain of $\text{Pan}_L$ to the signal before sending it to the left channel and a gain of $\text{Pan}_R$ to the signal before sending it to the right channel. The pan parameters $\Theta_{\text{Pan}}$ are initialized with mean 0, variance $10^{-6}$ Gaussian noise.

### D. Reverb

Similar to the EQ module, the reverb module also performs convolution with a given impulse response via multiplication in the frequency domain. Instead of estimating a frequency transfer curve, however, the reverb module directly estimates an impulse response $\Theta_{\text{IR}}$.

Two reverb architectures were tested in this work, one using a stereo reverb bus (thus, requiring two estimated impulse responses for a mixdown) and one using two impulse responses per channel [thus, requiring (2 × number of tracks) impulse responses for a mixdown]. Figures 1 and 2 show the block diagrams for each of the two architectures.

For the stereo reverb bus architecture, a wet/dry mix is produced by performing a weighted sum between the input stem and stem with reverb applied. Hence, the module estimates $\Theta_{\text{IR}}$ for the reverb's impulse response and $\Theta_{W/D}$ for the module's wet/dry mix. For the individual bus architecture, $\Theta_{W/D}$ is omitted.

In the stereo reverb bus case, the module's output is formulated as

$$\begin{aligned} \text{Reverb}(x(n)|\Theta_{W/D}, \Theta_{\text{IR}}) = x(n) + \sigma(\Theta_{W/D}) \\ \times (x(n) * \Theta_{\text{IR}}), \end{aligned} \tag{10}$$

where $\sigma( )$ denotes the sigmoid function, "×" denotes multiplication, and "*" denotes convolution. $\Theta_{W/D}$ is initialized with uniform random noise from $[-0.3, 0.3]$, which corresponds to a range of −7 to −5 dB.

In the individual reverb bus case, the output becomes

$$\text{Reverb}(x(n)|\Theta_{\text{IR}}) = x(n) + x(n) * \Theta_{\text{IR}}, \tag{11}$$

where "*" denotes convolution. In both cases, $\Theta_{\text{IR}}$ is initialized with mean 0, variance $10^{-6}$ random Gaussian noise.

https://doi.org/10.1121/10.0005622

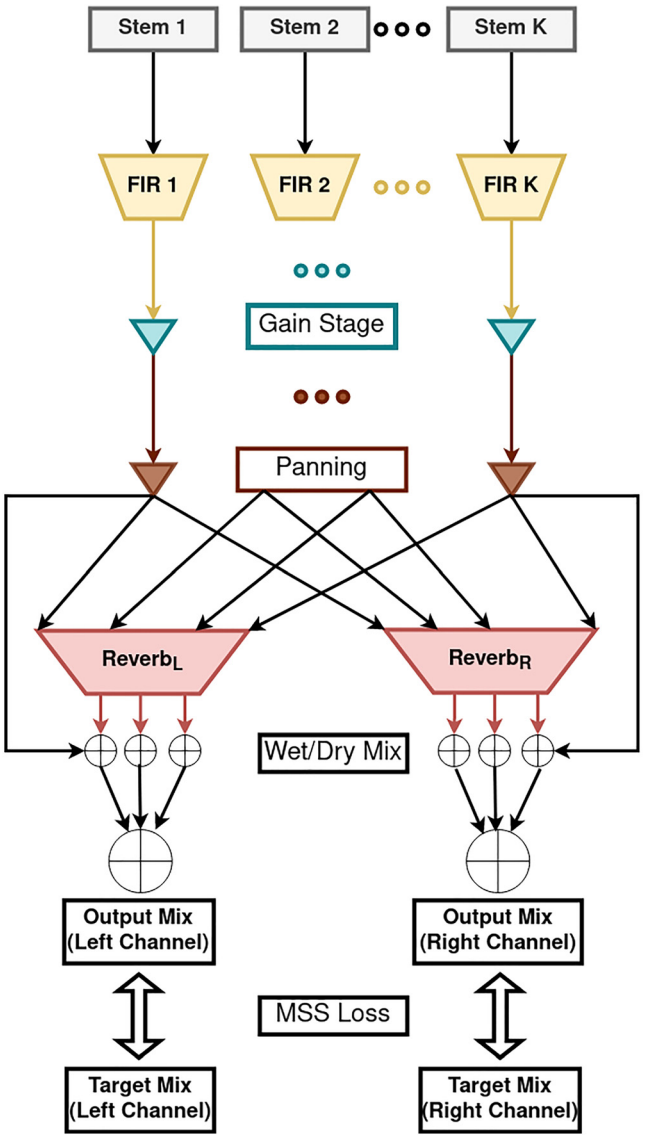


FIG. 1. (Color online) The mixing chain diagram for the "stereo bus" architecture.

The choice to investigate two reverb architectures stems from a desire to balance the method's modelling capacity with the network size and complexity. Mathematically speaking, both a left and right convolutional reverb impulse response must be estimated for each raw track in a multitrack to fully characterize the mixing chain. This is necessary because mixing engineers typically use stereo reverb impulse responses that are decorrelated in the left and right channels to increase spatialisation (Kendall, 1995).

However, the number of learned reverb parameters $\Theta_{\text{IR}}$ is orders of magnitude larger than $\Theta_{W/D}$, $\Theta_{\text{EQ}}$, $\Theta_{\text{Pan}}$, and $\Theta_{\text{Gain}}$ combined. As formulated in this paper, the number of parameters needed to describe the mixing chain before the reverb module is 12: 1 for gain, 1 for pan, and 10 for the graphic EQ. To model a 1 s convolutional reverb impulse response sampled at compact disc (CD) quality, 44 100 values are needed. Given that multitracks often contain more than 10 raw tracks, at least 20 impulse responses would have to be estimated for a full characterization, which balloons $\Theta_{\text{IR}}$ to 882 000 estimated values. The stereo reverb model would cap $\Theta_{\text{IR}}$ at 88 200 parameters, in this case, regardless of how many raw tracks make up the multitrack. In this formulation, $\Theta_{W/D}$ is necessary to control the "amount of reverb" applied to each raw track sent to the left and right channels.

### E. Loss and optimization procedure

Given randomly initialized $\Theta_{\text{gain}}$, $\Theta_{\text{pan}}$, $\Theta_{\text{EQ}}$, $\Theta_{\text{IR}}$, and $\Theta_{W/D}$, target mixdown $y(n)$, raw tracks $x_i(n)$, and estimated mixdown $\hat{y}(n)$ as described in Eq. (2), the stochastic gradient descent can be used to minimize $||y(n) - \hat{y}(n)||$ by updating the module parameters $\Theta$, where $||\cdot||$ denotes some norm used as a cost function.

In this work, a multi-scale spectrogram (MSS) loss is used as the cost function $||\cdot||$ (Engel *et al.*, 2019), which was inspired by the multi-resolution spectral amplitude distance demonstrated in Wang *et al.* (2020). As the name implies, the MSS computes a norm by measuring the distance between the spectrograms of two audio signals with varying STFT window sizes and performing a weighted sum of these differences. Although the mean absolute error (MAE) in the time domain is often used in audio applications and is cheaper to compute than the MSS loss, the latter was chosen because it ignores the phase differences between the target and estimated signals, which mimics human perception (Chi *et al.*, 2005). The resolutions for the spectrograms used are 2048, 512, and 128 samples. At a 44.1 kHz sampling rate, these correspond to windows of sizes 50, 12, and 3 ms. An L1 loss is computed on these spectrograms, which is the absolute value of the difference between the spectrograms reduced across both the frame and frequency dimensions.

The stochastic gradient descent was performed with learning rate scheduling and early stopping (Darken *et al.*, 1992). The descent begins using the ADAM optimizer with learning rate $10^{-3}$ (Kingma and Ba, 2015). Once the loss reaches an early stopping criterion, the learning rate is dropped to $10^{-4}$. After the same procedure happens again, the learning rate is further dropped to $10^{-5}$. The gradient descent concludes thereafter.

## IV. RESULTS AND ANALYSIS

All audio used in this work is sampled at 44.1 kHz, corresponding to CD quality audio. A professional mixing engineer was tasked with producing mixdowns for five separate multitrack recordings. The multitracks were chosen because they ranged from roughly 10 to 20 raw tracks each, had representative excerpts between 20 and 30 s in length, and were diverse in instrumentation and genre. All stereo tracks were converted to two mono tracks. All multitracks were downloaded from the Cambridge Multitracks dataset (Senior, 2011). Table I shows the artist and song title for each song used in this work. For each multitrack, three approximated mixdowns were calculated: a stereo bus approximation (two

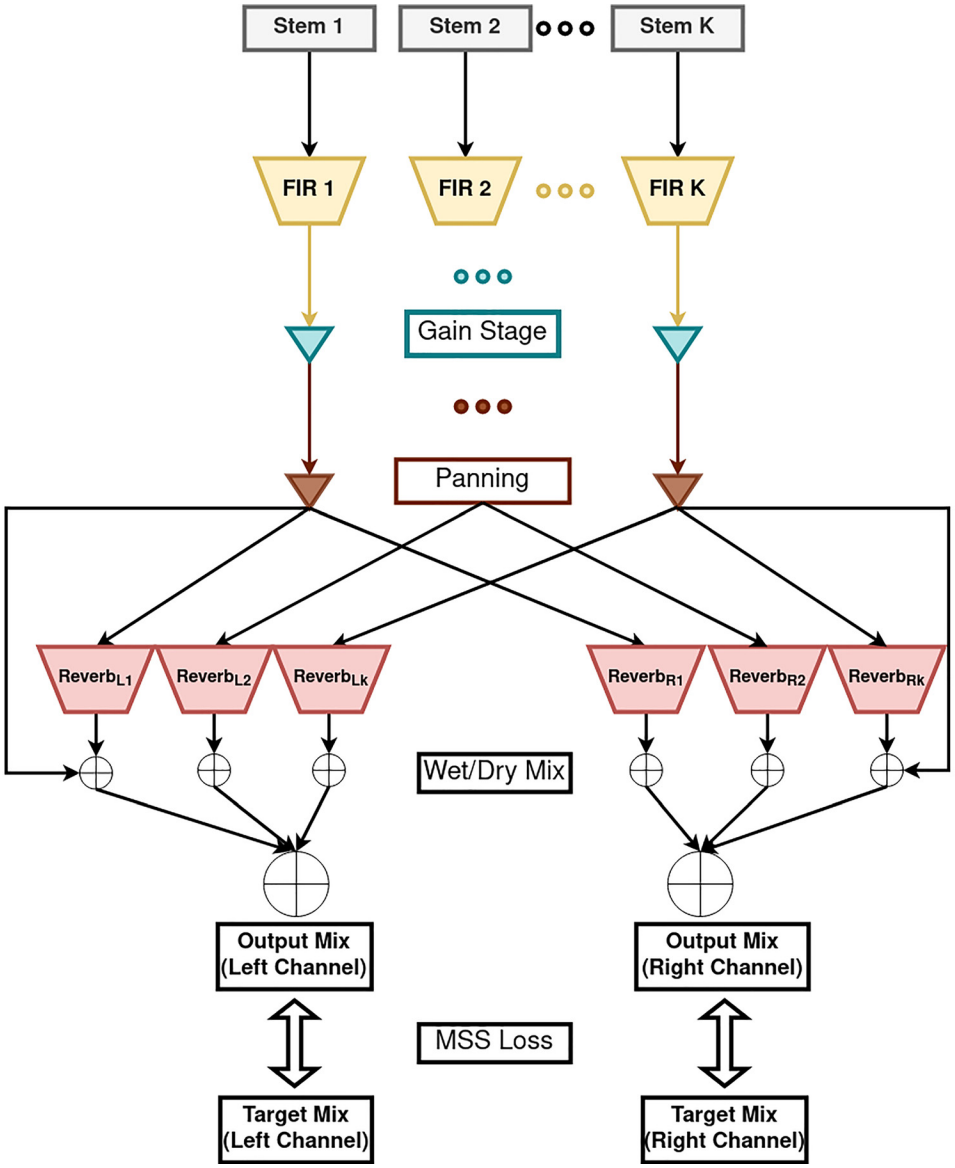


FIG. 2. (Color online) The mixing chain diagram for the "individual bus" architecture.

bus), an individual bus approximation (individual bus), and a stereo gain mix approximation using the least squares estimatation method in Barchiesi and Reiss (2010) (gain mix).

Similar to the procedure followed in Barchiesi and Reiss (2010), the mixing engineer only used linear processing to create the mixdown, including gain, pan, EQ, delay, and reverb. No distortion or dynamic range compression was used. Moreover, no automation was used on the linear effects. All audio discussed in this paper can be found online.[1]

TABLE I. The songs chosen for mixing and shortened names used in this paper.

| Artist | Song name | Genre | Reference name |
|---|---|---|---|
| Araujo | The Saga of the Harrison Crabfeathers | Jazz | Saga |
| Blue Lit Moon | Dad's Glad | Alternative rock | Glad |
| Carol Dant | I am the Desert | Electronica | Desert |
| The Complaniacs | Etc | Punk | Etc |
| Timboz | Pony | Metal | Pony |

### A. Objective evaluation

To objectively measure how close the estimated mixdowns matched the target mixdowns, a set of low-level audio features was calculated and compared according to the methodology in Wilson and Fazenda (2016). These features can be subgrouped into spectral measures (Bogdanov *et al.*, 2013), loudness measures (Recommendation, ■), stereo features (Tzanetakis *et al.*, 2007), and envelope probability mass function (PMF) features (Wilson and Fazenda, 2014). The PMF features are calculated by making a histogram of the values that a digital audio signal takes, normalizing this historgram so it becomes a PMF, and then calculating the statistical measures of this PMF.

https://doi.org/10.1121/10.0005622

A full account of each mixdown's feature values can be found in the supplementary materials included with this paper.[2] To aggregate each mixdown's performance, the average relative error of each subgroup of features can be calculated. The relative errors must be taken as different features have different units of measurements and, furthermore, different feature measures can be orders of magnitude apart from one another. Observing an average relative error across subgroups of features allows for an overall picture of how each mixdown matched the reference across perceptual correlates.

Table II provides a summary of the average errors across each feature subgroup for each approximated mixdown relative to the reference mixdown. Within each song, the two bus architecture outperforms both the individual bus architecture and gain mix across all subgroups of the features. For the songs "Saga," "Pony," "Desert," and "Dad's Glad," the gain mix outperforms the individual bus mix in average relative error across each subgroup of features. For the song "Etc," the individual bus mix outperforms the gain mix in average relative error for spectral, PMF, and stereo measures with the gain mix performing best for the loudness features.

There are certain instances in which the gain mix or the individual bus mix matches the reference more closely than the two bus mix in a given feature. When measuring 95% spectral roll-off, the individual bus mix performs best in the song "Saga." When measuring spectral spread, the gain mix outperforms the two bus mix in the songs "Dad's Glad" and "Saga." These are the only instances in spectral measure in which the two bus architecture does not perform best.

When measuring the PMF centroid, the gain mix performs best for the song "Dad's Glad," and the individual bus performs best for the song "Etc." For the PMF skew, the gain mix performs best in the songs "Desert" and "Pony." For the PMF kurtosis, the gain mix performs best in the song "Desert." In all other PMF measures, the two bus architecture performs best.

When measuring both the stereo panning spectrum across all bands and across high frequency bands, the individual bus mix performs best for the songs "Dad's Glad" and "Etc." For the mid band stereo panning spectrum, the individual bus performs best for the song "Etc." For the stereo left-right ratio, the gain mix performs best for the song "Desert." In all other stereo measures, the two bus architecture performs best.

For the loudness range measure, the gain mix performs best for the song "Pony." For the average crest factor measured with a resolution of 100 ms, the individual bus mix performs best for the song "Desert," and the gain mix performs best for the song "Pony." When the resolution of the crest factor is increased to 1000 ms, the gain mix performs best for the song "Desert." The two bus mix performs best for all other measures of loudness.

TABLE II. The average relative errors by feature subgroup for approximated mixdowns compared to the reference mixdown.

| Mixdown | Spectral | PMF | Stereo | Loudness | Total |
|---|---|---|---|---|---|
| Desert two bus | 0.95% | 14.86% | 5.99% | 1.68% | 4.88% |
| Desert gain mix | 5.87% | 15.96% | 13.00% | 4.80% | 9.21% |
| Desert individual bus | 35.55% | 104.67% | 133.84% | 25.24% | 70.52% |
| Etc two bus | 2.52% | 6.57% | 14.07% | 3.67% | 6.54% |
| Etc gain Mix | 19.21% | 95.99% | 33.83% | 5.94% | 32.91% |
| Etc individual bus | 9.63% | 62.78% | 20.37% | 22.64% | 25.07% |
| Glad two bus | 7.10% | 6.04% | 32.49% | 1.80% | 12.16% |
| Glad gain mix | 23.16% | 38.72% | 51.85% | 8.76% | 29.60% |
| Glad individual bus | 34.83% | 91.09% | 92.46% | 17.52% | 55.13% |
| Pony two bus | 1.70% | 21.34% | 9.95% | 7.67% | 8.82% |
| Pony gain mix | 28.24% | 49.42% | 35.33% | 15.68% | 30.49% |
| Pony individual bus | 61.16% | 55.86% | 159.22% | 16.65% | 74.21% |
| Saga two bus | 2.25% | 3.26% | 3.53% | 0.41% | 2.28% |
| Saga gain mix | 13.57% | 25.52% | 17.44% | 12.03% | 16.26% |
| Saga individual bus | 27.37% | 356.85% | 137.33% | 61.17% | 122.17% |

### B. Human listener evaluation

Whereas much research has been done to numerically characterize the timbre (see Peeters *et al.*, 2011) and relate closeness of timbres within a perceptual space (e.g., Caclin *et al.*, 2005; Elliott *et al.*, 2013), there are no perfect numerical measures for determining how close two timbres are. Furthermore, the timbral complexity of multitrack mixes renders numerically characterizing mixes challenging (Wilson and Fazenda, 2015; Colonel and Reiss, 2019). Although the results presented in the objective evaluation show the two bus mix outperforming the other two mixes in numerical measures, this is no guarantee that the two bus mix sounds most similar to the reference mix. Thus, a listening study was performed to assess how well each method perceptually matched the reference mix.

The listening study included 23 participants to evaluate the approximated mixdowns of the five multitracks. Fourteen participants reported having no audio production or mixing experience, whereas nine participants reported having some audio production or mixing experience. Participants were tasked with rating a set of mixes based on how closely the mixes matched a reference mix on a continuous scale of 0–1, where "0" represented a mix "very far" from the reference, and "1" represented a mix "matching exactly" the reference mix.

During the test, participants were presented each of the mulitracks in a random order one-by-one. For a given multitrack, participants were presented with four stimuli to rate against the reference mix. One stimulus was the identical reference mix. The other presented stimuli were three approximated mixdowns: a stereo bus approximation, an individual bus approximation, and a stereo gain mix approximation. Participants evaluated mixes for each of the 5 multitracks, therefore, providing a total of 20 ratings. Participants were encouraged to use the full 0–1 rating range when appropriate. Furthermore, participants were given no

TABLE III. The $f$ stats and $p$-values for each multitrack and all multitracks, including all participants.

| Group | $f$ stat | $p$-value |
|---|---|---|
| All songs | 316.178 | $3.737 \times 10^{-107}$ |
| Desert | 143.595 | $4.127 \times 10^{-32}$ |
| Etc | 100.860 | $4.195 \times 10^{-27}$ |
| Glad | 91.140 | $9.891 \times 10^{-26}$ |
| Pony | 70.035 | $2.575 \times 10^{-22}$ |
| Saga | 42.554 | $1.529 \times 10^{-16}$ |

time limit for the test, and no limit was placed on how many times a participant could listen to a given stimulus.

The analysis that follows is adapted from the perceptual study presented in Moffat and Reiss (2018). The null hypothesis is that the perceptual evaluation scores are from the same distribution. A one-way analysis of variance (ANOVA) with Bonferroni correction shows that for all mixdowns, the effect the method used to reverse engineer the mix had on the user perception was statistically significant. This result holds when analysing the ratings separated by each multitrack as well. Table III lists the $f$ stats and $p$-values. Figures 3 and 4 show the box plots of the overall results of the listening study.

With the null hypothesis rejected, a *post hoc* Tukey pairwise comparison with Bonferroni correction to reduce the chance of type I errors was used. Table V shows the results of these pairwise comparisons for all of the architectures used. The pairwise comparisons demonstrate that the mean of the participants' ratings for the reference mix and stereo bus mix do not differ significantly. All other pairwise comparisons do differ significantly.

When breaking down the data by song, the above results hold for three of the five multitracks: "Glad," "Desert," and "Saga." For the multitrack "Etc," the pairwise comparisons demonstrate that the mean of the participants' ratings differs significantly for all pairs, including the reference and stereo bus models as shown in Table VI. For the multitrack "Pony," the reference/stereo bus model pair and gain mix/individual bus model do not have participant rating distributions that differ significantly as shown in Table VII.

Most of these results hold when excluding the listeners with no audio production or mixing experience. Box plots can be found in Figs. 5 and 6. A one-way ANOVA with Bonferroni correction shows for all mixdowns that the effect the method used to reverse engineer the mix had on user perception was statistically significant. This result holds when analysing the ratings separated by each multitrack as well. Table IV lists the $f$ stats and $p$-values for this subset of participants with some audio production or mixing experience.

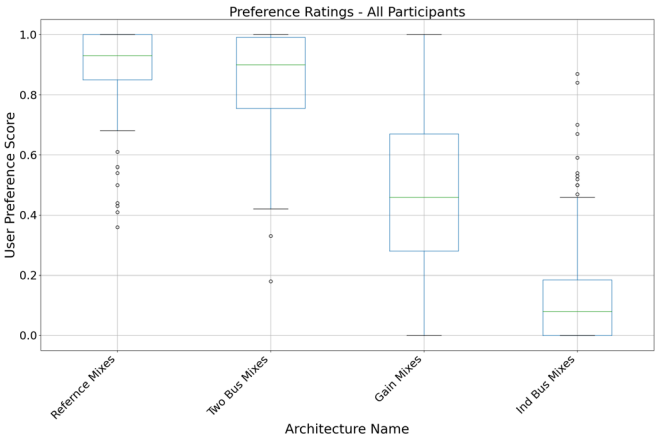


FIG. 4. (Color online) The box plots of all of the listener ratings, broken down by each mixdown.

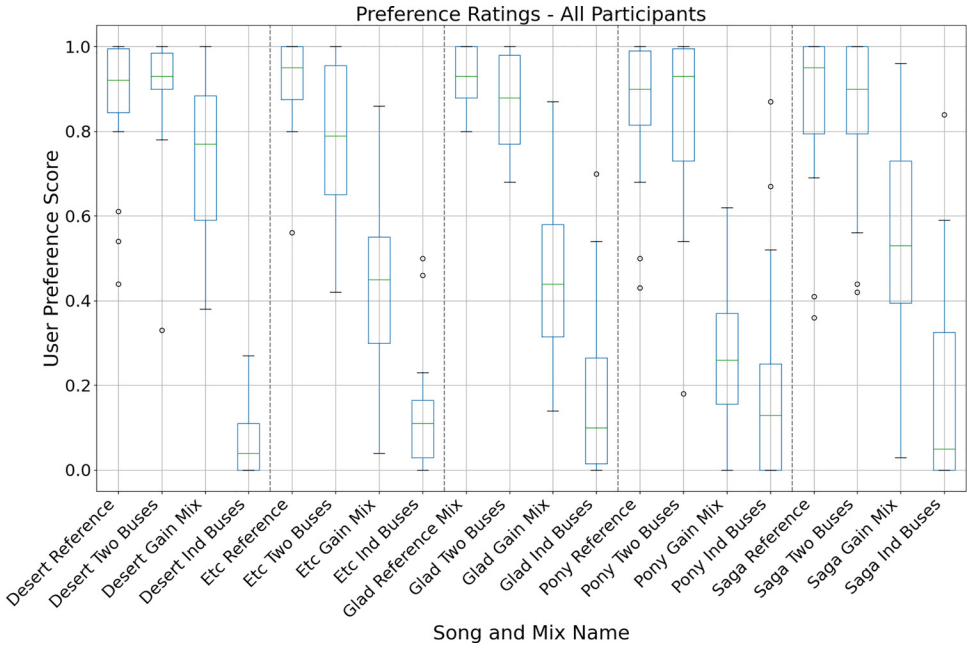


FIG. 3. (Color online) The box plots of all of the listener ratings, broken down by estimation architecture.

https://doi.org/10.1121/10.0005622

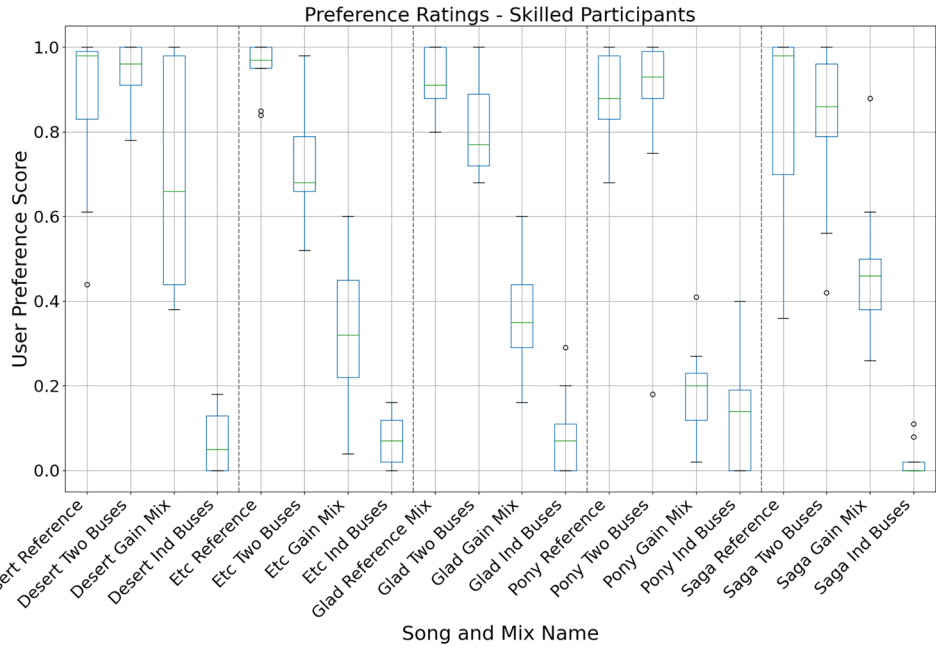


FIG. 5. (Color online) The box plots of the ratings made by listeners with audio experience, broken down by architecture.

*Post hoc* Tukey pairwise comparisons of the ratings provided by this subset of participants are similar to those for the whole group. When comparing across all songs and mixdowns, the values of the Tukey analysis match those presented in Table V. This also remains true for the songs "Glad" and "Saga." For the multitrack "Etc," the pairwise comparisons again demonstrate that the mean of the participants' ratings differs significantly for all pairs, including the reference and stereo bus models as shown in Table VI. And, again, in "Pony," the reference/stereo bus model pair and gain mix/individual bus models do not have participant rating distributions that differ significantly as shown in Table VII. This subset's ratings do differ for the song "Desert," where there is no statistical difference between the stereo bus/reference pair and gain/reference pair. The results are shown in Table VIII. As Fig. 3 demonstrates, participants rated the stereo bus model as nearly identical to the reference mix, then rated the gain mix next closest, and finally rated the individual bus mix the furthest from the reference.

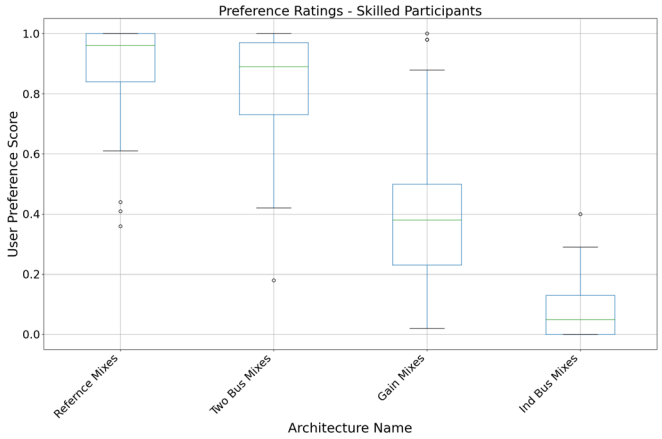


FIG. 6. (Color online) The box plots of the ratings made by listeners with audio experience, broken down by mixdown.

## V. DISCUSSION

The results of both the objective evaluation and listening test suggest that the stereo reverb bus model outperforms both the gain mix and individual bus models in a reverse engineering of a mix task. Furthermore, the gain mix outperforms the individual bus model in almost all cases.

It is interesting to note that the individual bus architecture performs poorer than a gain mix in both objective and subjective measures given that the gain mix does not apply EQ or reverb. Even with the explicit ability to modify a raw track's spectral content, the individual bus architecture does a poorer job than the gain mix of matching the reference mix's spectral features in four of the five songs. "Etc" is the only song where the individual bus model outperforms the gain mix in spectral measures and, incidentally, is the only song where the individual bus mix outperforms the gain mix in all other measures as well. Yet, the results of the listening test place the individual bus estimate of "Etc" lower than

TABLE IV. The $f$ stats and $p$-values for each multitrack and all multitracks, including only participants with audio production or mixing experience.

| Group | $f$ stat | $p$-value |
|---|---|---|
| All songs | 207.896 | $1.322 \times 10^{-57}$ |
| Desert | 47.444 | $6.977 \times 10^{-12}$ |
| Etc | 91.529 | $8.701 \times 10^{-16}$ |
| Glad | 117.997 | $2.208 \times 10^{-17}$ |
| Pony | 53.478 | $1.450 \times 10^{-12}$ |
| Saga | 35.129 | $3.062 \times 10^{-10}$ |

TABLE V. The results of the pairwise comparison of the mixdown architecture on the perceptual similarity rating across multitracks with the Bonferroni correction, $o > 0.05$, * $< 0.001$ · = no comparison.

| | Reference mix | Gain mix | Individual bus mix | Two bus mix |
|---|---|---|---|---|
| Reference mix | · | * | * | *o* |
| Gain mix | * | · | * | * |
| Individual bus mix | * | * | · | * |
| Two bus mix | *o* | * | * | · |

TABLE VII. The results of the pairwise comparison of the mixdown architecture on the perceptual similarity rating within the "Pony" multitrack with the Bonferroni correction, $o > 0.05$, * $< 0.001$ · = no comparison.

| | Reference mix | Gain mix | Individual bus mix | Two bus mix |
|---|---|---|---|---|
| Reference mix | · | * | * | *o* |
| Gain mix | * | · | *o* | * |
| Individual bus mix | * | *o* | · | * |
| Two bus mix | *o* | * | * | · |

the gain mix for both experienced and inexperienced listeners. This highlights the difficulty in using objective features to characterize the multitrack mixes. In its estimate, the individual bus architecture placed a prominent reverb on the vocal stem that does not match the reference mix, which is most likely why it performed so poorly in the listener evaluation. In general, the mixes estimated by the individual bus architecture frequently apply much more reverb than the reference.

Across all of the tests, the "Desert" mix from the individual bus model performed the worst in the listening test, and the "Desert" mix from the stereo bus model performed the best. In "Desert," several synthesizers (Synth1–Synth6) are layered within the composition and provide backing to a layered vocal (Vox1–Vox3). The reference mix applies both delay and reverb to most of the song's elements in keeping with a "washed-out electronica" style mix, as well as the EQ, gain, and pan. Figures 7 and 8 show $\Theta_{IR}$ and the frequency transfer curves for select stems from the song "Desert," produced by the individual bus model and stereo bus model, respectively. Observing Fig. 7, one can see the stark differences between the learned reverb impulse responses across the stems. Synth3 appears to have both echo and reverb applied, Synth6 has a dense reverb with significant energy in the tail applied, and Vox2 has a light echo applied. Note as well that the gradient descent produces distinct reverbs for the left and right channels for each of these stems. When listening to this mix, however, the vocals are barely audible, and the synthesizers dominate the mix. Figure 8 shows that the stereo bus model has combined both a reverb with less energy in the tail than the individual bus model with a prominent echo into the left and right channels learned impulse responses. The result is a mixdown that is nearly indistinguishable from the reference mix.

TABLE VI. The results of the pairwise comparison of the mixdown architecture on the perceptual similarity rating within the "Etc" multitrack with the Bonferroni correction, $o > 0.05$, * $< 0.001$ · = no comparison.

| | Reference mix | Gain mix | Individual bus mix | Two bus mix |
|---|---|---|---|---|
| Reference mix | · | * | * | * |
| Gain mix | * | · | * | * |
| Individual bus mix | * | * | · | * |
| Two bus mix | * | * | * | · |

This failure by the individual bus model may be due to the relatively large parameter space that the optimization has to navigate. It may be the case that given random parameter initialization for the 20 raw tracks, the optimization begins too far away from the parameters of the reference and instead converges to a random local minimum. Note that with 20 raw tracks, $\Theta_{IR}$ consists of $1.764 \times 10^6$ parameters. It is interesting to note that the individual bus model's optimization also does not match the reference when panning certain raw tracks. While the stereo reverb bus model pans Vox2 nearly center with $\text{Pan}_L = 0.502$, the individual reverb bus model pans Vox2 to the right with $\text{Pan}_L = 0.314$. This again suggests that the individual reverb bus model is exploring some area of the parameter space distant from the reference mix.

Future improvements may be made to the individual bus model by bypassing the reverb module for the beginning of the stochastic gradient descent. For example, the bypass could be activated until the first early stopping, which would allow for the network to best fit the gain, pan, and EQ parameters before attempting to apply reverb. A full study of how DDSP performs in reverb estimation may also shed light on the issues of the individual bus model, including how a gradient descent performed on a FIR reverb IR can match infinite impulse response (IIR) reverb implementations and reverb impulse responses with non-integer delays.

There are several directions that future work can take. The most pressing is implementing nonlinear processing in the signal chain, including dynamic range compression and distortion. These can either come in the form of some white-box approximation of the effects or blackbox modules inserted into the signal chain, which model specific effects like those presented in Steinmetz and Reiss (2021) or Choi *et al.* (2021). Another direction the work can take is

TABLE VIII. The results of the pairwise comparison of the mixdown architecture on the perceptual similarity rating within the "Desert" multitrack, including only participants with audio production experience, with Bonferroni correction, $o > 0.05$, * $< 0.001$ · = no comparison.

| | Reference mix | Gain mix | Individual bus mix | Two bus mix |
|---|---|---|---|---|
| Reference mix | · | *o* | * | *o* |
| Gain mix | *o* | · | *o* | * |
| Individual bus mix | * | *o* | · | * |
| Two bus mix | *o* | * | * | · |

https://doi.org/10.1121/10.0005622

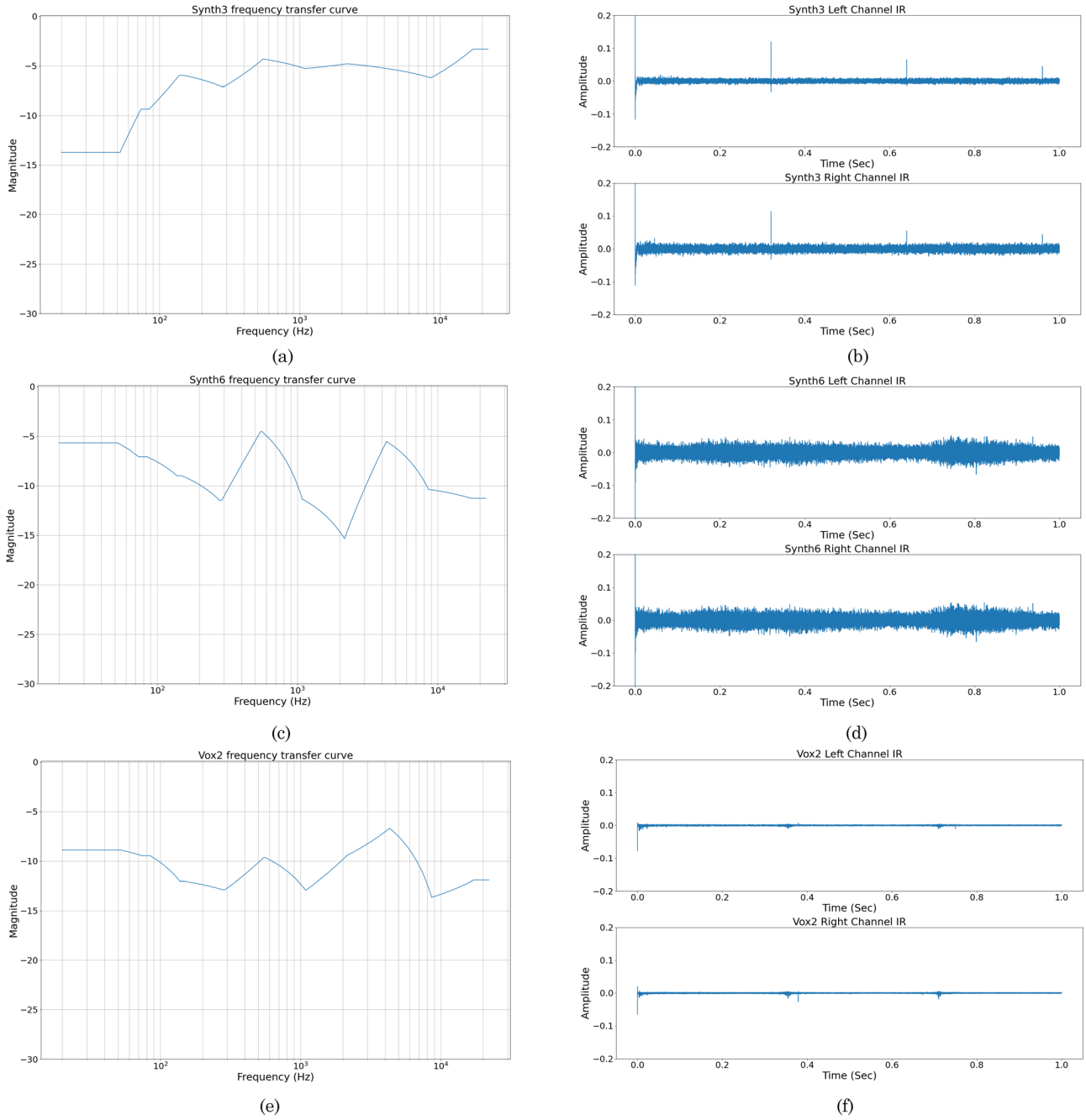


FIG. 7. (Color online) The estimated frequency transfer curves and reverb impulse responses for select stems in the song "Desert," using the individual bus model.

modelling automation, which is frequently used by mixing engineers. This could be realised using frame-by-frame approximations of mixing parameters or some other control scheme. Another direction would be the implementation of differentiable IIR filters for the EQ as suggested in Nercessian (2020) or Kuznetsov *et al.* (2020) rather than the FIR filters presented here. Of particular note would be a comparison in the phase behaviour of DDSP FIR EQs and differentiable IIR EQs.

This reverse engineering work may also aid in numerically characterizing mixing engineers' behaviour by analysing and extracting mix parameters from a corpus of professional mixes. This corpus could then be used to improve objective measures of multitrack mixes for perceptual correlation to avoid issues such as those encountered when objectively measuring the "Etc" mixdowns.

## VI. CONCLUSION

This paper presents two architectures that can reverse engineer the linear processing, including the gain, pan, EQ, delay, and reverb, of a mix. This architecture takes as its inputs the raw tracks of a multitrack and targets a reference mixdown. This is achieved using whitebox neural network

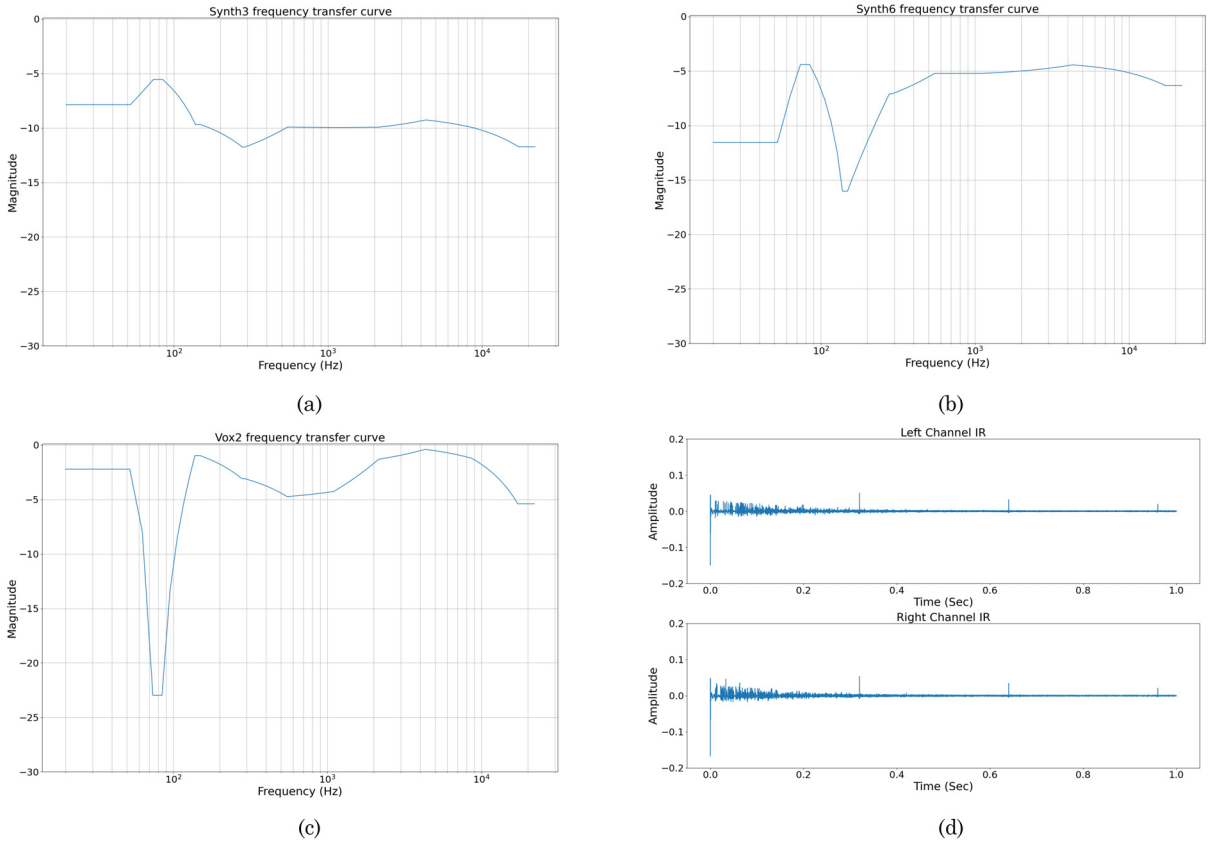


FIG. 8. (Color online) The estimated frequency transfer curves and reverb impulse responses for select stems in the song "Desert," using the stereo bus model.

audio effect modelling using DDSP and performing gradient descent. The results of objective measures and a listening test demonstrate that the stereo bus reverb model outperforms the individual reverb bus model, and the participant's perceptual ratings between the stereo reverb model and reference mix do not differ significantly.

## ACKNOWLEDGMENTS

This work was partly supported by the research development division of Yamaha Corporation, Japan. We would like to thank Angeliki Mourgela for providing us with quality references mixes. We would also like to thank Dave Moffat for his conversations when designing these system architectures.

[1]See https://jtcolonel.github.io/RevEng/ (Last viewed July 19, 2021).
[2]See Supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0005622 for a table with all objective features measured from each mixdown.

Barchiesi, D., and Reiss, J. (**2010**). "Reverse engineering of a mix," J. Audio Eng. Soc. **58**(7/8), 563–576.
Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E. (**1997**). "A blind source separation technique using second-order statistics," IEEE Trans. Signal Process. **45**(2), 434–444.
Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (**2013**). "Essentia: An audio analysis library for music information retrieval," in *14th Conference of the International Society for Music Information Retrieval (ISMIR)*, edited by A. Britto, F. Gouyon, and S. Dixon, November 4–8, Curitiba, Brazil, pp. 493–498.
Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (**2005**). "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," J. Acoust. Soc. Am. **118**(1), 471–482.
Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**(2), 887–906.
Choi, W., Kim, M., Ramírez, M. A. M., Chung, J., and Jung, S. (**2021**). "Amss-net: Audio manipulation on user-specified sources with textual queries," arXiv:2104.13553.
Colonel, J., and Reiss, J. D. (**2019**). "Exploring preference for multitrack mixes using statistical analysis of mir and textual features," in *Audio Engineering Society Convention 147* (Audio Engineering Society, New York, New York).
Darken, C., Chang, J., and Moody, J. (**1992**). "Learning rate schedules for faster stochastic gradient search," in *Neural Networks for Signal Processing* (Citeseer, Helsinoger, Denmark), Vol. 2.
Elliott, T. M., Hamilton, L. S., and Theunissen, F. E. (**2013**). "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones," J. Acoust. Soc. Am. **133**(1), 389–404.
Engel, J., Hantrakul, L., Gu, C., and Roberts, A. (**2019**). "DDSP: Differentiable digital signal processing," in *International Conference on Learning Representations*.
Feng, X., Zhang, Y., and Glass, J. (**2014**). "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 1759–1763.
Gorlow, S., and Marchand, S. (**2013**). "Reverse engineering stereo music recordings pursuing an informed two-stage approach," in *2013 International Conference on Digital Audio Effects (DAFx-13)*, pp. 1–8.
Gorlow, S., and Reiss, J. D. (**2013**). "Model-based inversion of dynamic range compression," IEEE Trans. Audio, Speech, Lang. Process. **21**(7), 1434–1444.

Grais, E. M., and Plumbley, M. D. (**2017**). "Single channel audio source separation using convolutional denoising autoencoders," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (IEEE, New York), pp. 1265–1269.
Hawley, S., Colburn, B., and Mimilakis, S. I. (**2019**). "Profiling audio compressors with deep neural networks," in *Audio Engineering Society Convention 147* (Audio Engineering Society, New York, New York).
Jourjine, A., Rickard, S., and Yilmaz, O. (**2000**). "Blind separation of disjoint orthogonal signals: Demixing *n* sources from 2 mixtures," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 2985–2988.
Kendall, G. S. (**1995**). "The decorrelation of audio signals and its impact on spatial imagery," Comput. Music J. **19**(4), 71–87.
Kingma, D. P., and Ba, J. (**2015**). "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (Poster)*.
Kuznetsov, B., Parker, J. D., and Esqueda, F. (**2020**). "Differentiable IIR filters for machine learning applications," in *Proc. Int. Conf. Digital Audio Effects (eDAFx-20)*, pp. 297–303.
Lebart, K., Boucher, J.-M., and Denbigh, P. N. (**2001**). "A new method based on spectral subtraction for speech dereverberation," Acta Acust. Acust. **87**(3), 359–366.
Maas, A. L., Hannun, A. Y., and Ng, A. Y. (**2013**). "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning*, Vol. 30, p. 3.
Moffat, D., and Reiss, J. D. (**2018**). "Perceptual evaluation of synthesized sound effects," ACM Trans. Appl. Percept. **15**(2), 1–19.
Nercessian, S. (**2020**). "Neural parametric equalizer matching using differentiable biquads," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-20)*, Vienna, Austria.
Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (**2011**). "The timbre toolbox: Extracting audio descriptors from musical signals," J. Acoust. Soc. Am. **130**(5), 2902–2916.
Ramírez, M. A. M., and Reiss, J. D. (**2018**). "End-to-end equalization with convolutional neural networks," in *21st International Conference on Digital Audio Effects (DAFx-18)*.
Ramírez, M. A. M., and Reiss, J. D. (**2019**). "Modeling nonlinear audio effects with end-to-end deep neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, New York), pp. 171–175.
Ramírez, M. M., Benetos, E., and Reiss, J. D. (**2020**). "Deep learning for black-box modeling of audio effects," Appl. Sci. **10**(2), 638.
Recommendation, E. (2014). "R 128—Loudness normalisation and permitted maximum level of audio signals" (PDF) tech.ebu.ch, European Broadcasting Union, Geneva, June 2014 (Retrieved 19 July 2021).
Senior, M. (**2011**). *Mixing Secrets for the Small Studio* (Taylor and Francis, London).
Steinmetz, C. J., Pons, J., Pascual, S., and Serrà, J. (**2020**). "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," arXiv:2010.10291.
Steinmetz, C. J., and Reiss, J. D. (**2021**). "Efficient neural networks for real-time analog audio effect modeling," arXiv:2102.06200.
Stoller, D., Ewert, S., and Dixon, S. (**2018**). "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*, Paris, France, September 23–27, pp. 334–340.
Tzanetakis, G., Jones, R., and McNally, K. (**2007**). "Stereo panning features for classifying recording production style," in *International Society for Music Information Retrieval Conference (ISMIR)* (Citeseer, Vienna, Austria), pp. 441–444.
Välimäki, V., and Reiss, J. D. (**2016**). "All about audio equalization: Solutions and frontiers," Appl. Sci. **6**(5), 129.
Wang, X., Takaki, S., and Yamagishi, J. (**2020**). "Neural source-filter waveform models for statistical parametric speech synthesis," IEEE/ACM Trans. Audio, Speech, Lang. Process. **28**, 402–415.
Wilson, A., and Fazenda, B. (**2014**). "Categorisation of distortion profiles in relation to audio quality," in *International Conference on Digital Audio Effects (DAFx)*.
Wilson, A., and Fazenda, B. (**2015**). "101 mixes: A statistical analysis of mix-variation in a dataset of multi-track music mixes," in *Audio Engineering Society Convention 139* (Audio Engineering Society, New York, New York).
Wilson, A., and Fazenda, B. (**2016**). "Variation in multitrack mixes: Analysis of low-level audio signal features," J. Audio Eng. Soc. **64**(7/8), 466–473.
Yu, G., Mallat, S., and Bacry, E. (**2008**). "Audio denoising by time-frequency block thresholding," IEEE Trans. Signal Process. **56**(5), 1830–1839.