Deep learning approaches to multitrack mixing

Joseph Colonel[1,4] and Christian Steinmetz[2,3]

[1]Centre for Digital Music, Queen Mary University of London
[2]Music Technology Group, Universitat Pompeu Fabra, Barcelona
[3]Dolby Laboratories
[4]Collaboration with the Yamaha Corporation

Who are we?

Joseph Colonel
@josephtcolonel
josephtcolonel.com
j.t.colonel@qmul.ac.uk

Christian Steinmetz
@csteinmetz1
christiansteinmetz.com
c.j.steinmetz@qmul.ac.uk

Previous work

(pre-deep learning)

Two major approaches

Expert systems

Machine Learning

Expert systems
(Knowledge engineering)

1. Develop a knowledge base → Consult textbooks and audio engineers
2. Define a set of rules and logic → Formalize rules based on instrument class
3. Use rules to perform task → Perform processing based on instruments

Brecht De Man and Joshua D. Reiss, “A knowledge-engineered autonomous mixing system,” 135th Convention of the Audio Engineering Society, October 2013.

Pro: Produces explainable decisions

Con: Lacks sufficient complexity

Machine Learning

(Classical ML algorithms)

1. Construct relevant dataset → ENST-drums dataset gain mixes
2. Apply learning algorithms → Random forests
3. Perform inference with model → Predict gain coefficients

D. Moffat, and M. Sandler, "Machine Learning Multitrack Gain Mixing of Drums," Audio Engineering Society, Engineering Brief 527, (2019 October.)

Pro: Provides greater model flexibility

Con: Absence of large scale parametric data

LEVEL | EQUALIZATION | COMPRESSION | PANNING | REVERB | MULTIPLE | MACHINE LEARNING | KNOWLEDGE-BASED | OVERVIEW | CLEAR

Show 10 entries

Search:

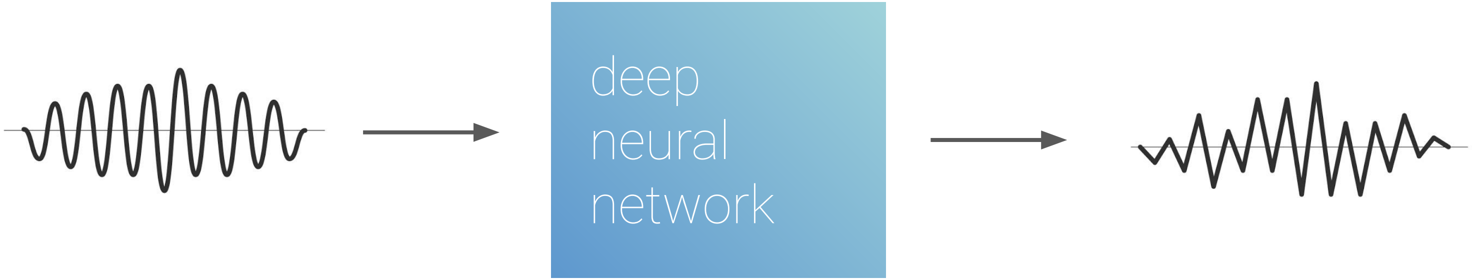
Year	Title	Author(s)	Category	Approach	Code
2020	One-shot parametric audio production style transfer with application to frequency equalization	S. I. Mimilakis, N. J. Bryan, and P. Smaragdis	Equalization	ML	
2020	Mixing with intelligent mixing systems: evolving practices and lessons from computer assisted design	M. N. Lefford, G. Bromham, and D. Moffat	Review	Multiple	
2019	An automatic mixing system for multitrack spatialization for stereo based on unmasking and best panning practices	A. Tom, J.D. Reiss, and P. Depalle	Panning	KBS	CODE
2019	Automatic mixing level balancing enhanced through source interference identification	D. Moffat and M. B. Sandler	Level	KBS	
2019	Background ducking to produce esthetically pleasing audio for TV with clear speech	M. Torcoli et al.	Level	KBS	

For a more complete review of the field see this webpage, which features a searchable table of relevant papers.

https://csteinmetz1.github.io/AutomaticMixingPapers

These systems often fail to generalize to real-world music production use cases.

*...but recent successes in **deep learning** for audio motivates the application of new methods*



End-to-end **deep learning** for multitrack mixing

1. Learning directly from waveforms, no knowledge of parameters
2. Surpass performance of previous ML and expert systems
3. Greater processing flexibility to create “detailed” mixes

Key challenges

in applying deep learning

1. **Limited training data** *We need the original tracks and good mixes.*
2. **Evaluation of mixes** *What makes a good mix? According to who?*
3. **Highly variable inputs** *No consistent size and structure to inputs.*
4. **High-fidelity required** *High sampling rates and no artifacts.*
5. **User interaction** *Audio engineers need to tweak the output.*

Outline

Three existing deep learning approaches

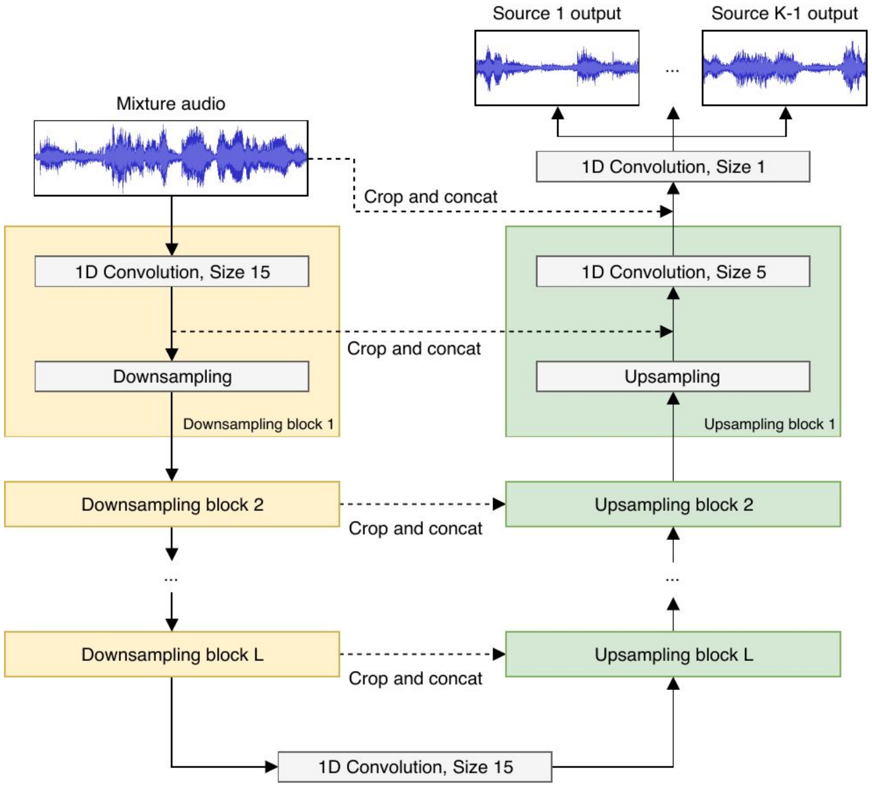
1. **Wave-U-Net for multitrack mixing**
   Work from Martínez Ramírez, Stoller, and Moffat
2. **DDSP for multitrack mixing**
   Work from Colonel and Reiss
3. **Differentiable mixing console**
   Work from Steinmetz and Serrà

Wave-U-Net

for multitrack mixing

Wave-U-Net

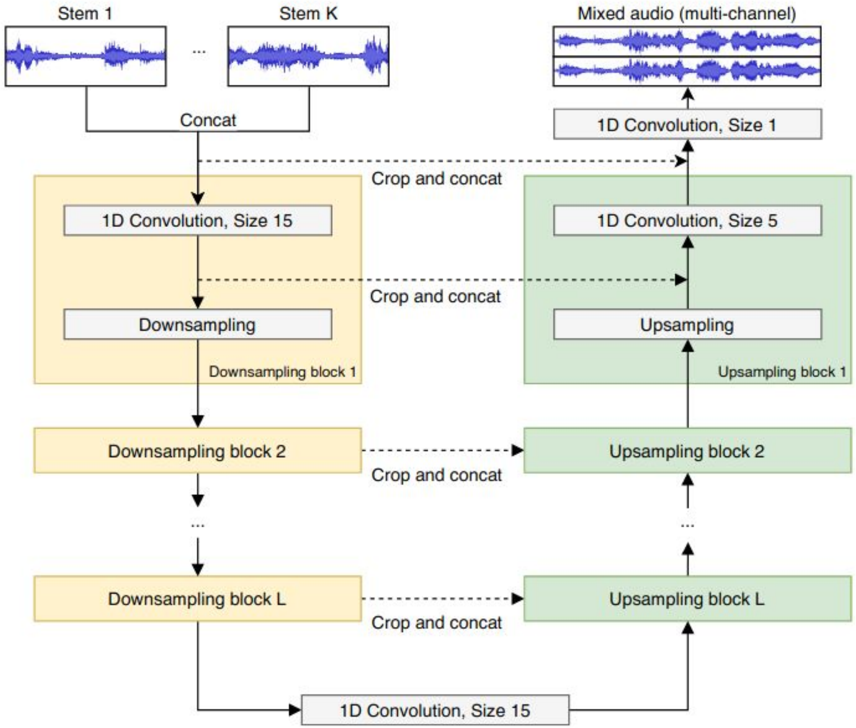
- Architecture originally proposed for source separation task
- Convolutional, U shaped network
- Input waveform retained at final layer to inform separation



Stoller, D., S. Ewert, and S. Dixon. "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation." ISMIR. 2018.

Wave-U-Net for Drum Mixing

- “Reverse” source separation
- ENST-Drums dataset
- Convolutional, U shaped network
- Input stems retained at final layer to inform mixing
- Learns EQ, reverb, compression in “black box” manner



Gillet, Olivier, and Gaël Richard. "ENST-Drums: an extensive audio-visual database for drum signals processing." ISMIR. 2006.

M. Martinez, D. Stoller, and D. Moffat "A Deep Learning Approach to Intelligent Drum Mixing with the Wave-U-Net" Journal of the Audio Engineering Society, Accepted Manuscript
https://mchijmma.github.io/drum-mixing-wave-u-net/

DDSP

for multitrack mixing

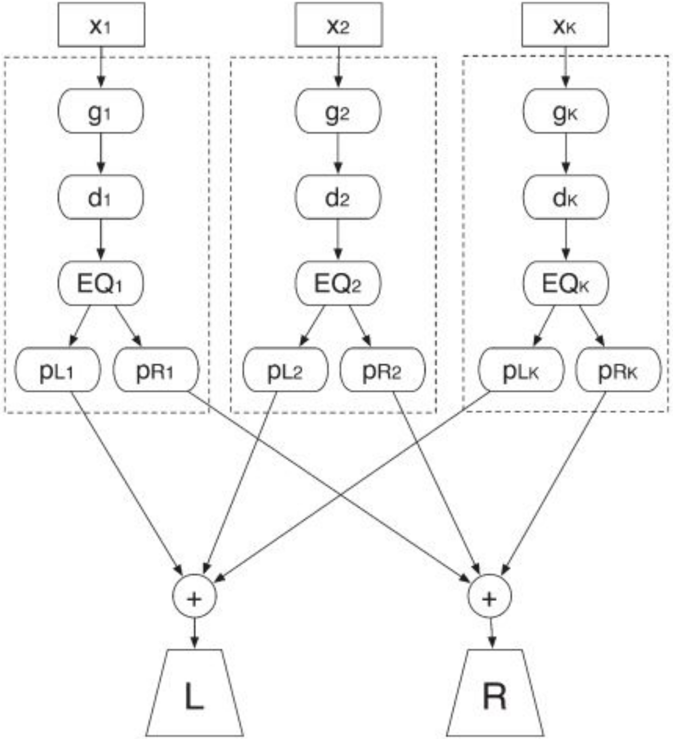
Differentiable Digital Signal Processing (DDSP)

- Python library developed by Magenta
- Casts common DSP modules for use in neural networks
  - Convolutional reverb, FIR filters, etc.
- Demonstrated uses in sound synthesis and timbre transfer
  - Harmonic oscillators, filtered noise, etc.

Engel, Jesse, Chenjie Gu, and Adam Roberts. "DDSP: Differentiable Digital Signal Processing." International Conference on Learning Representations. 2019.

Reverse Engineering a Mix

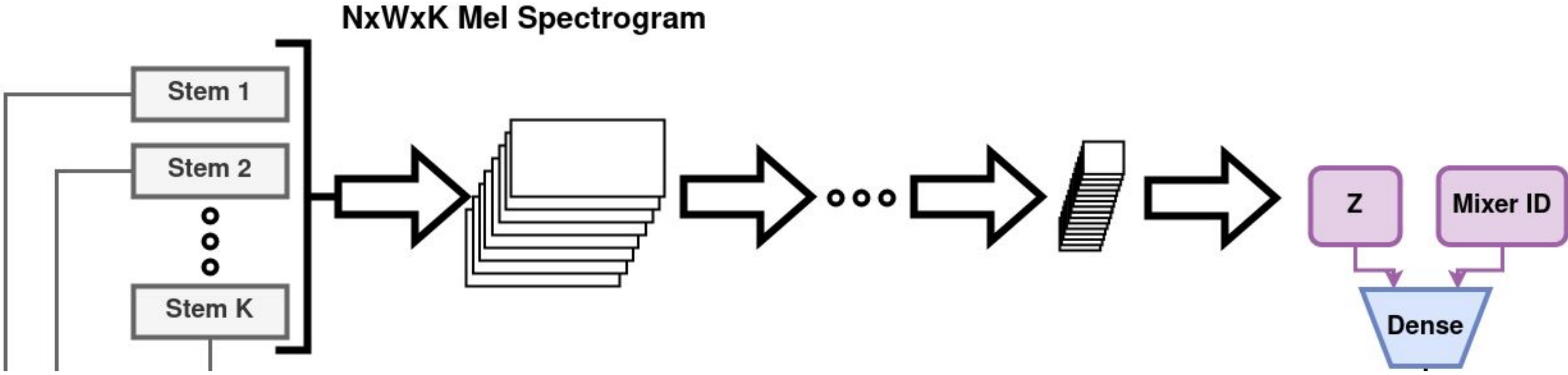
- Estimate mix parameters using stems and mixdown
    - Model both linear time-invariant (LTI) and dynamic processing
- DDSP approach can model reverb as well



Barchiesi, Daniele, and Joshua Reiss. "Reverse engineering of a mix." Journal of the Audio Engineering Society 58.7/8 (2010): 563-576.

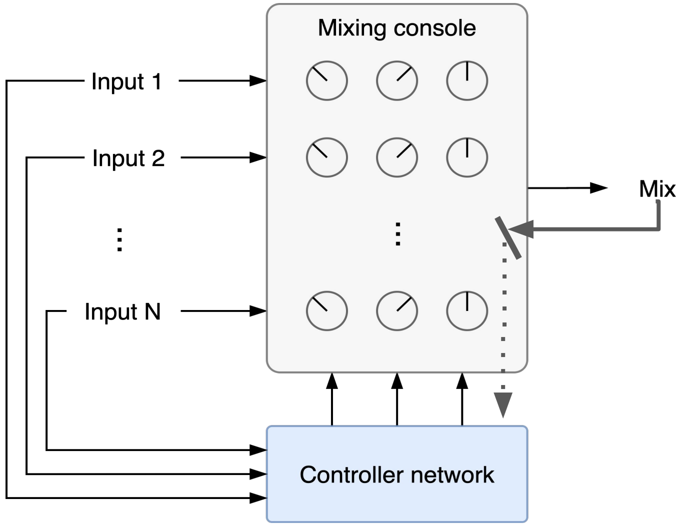
Mixing System - in Development

- Working with ENST Drums dataset
- Explicit modelling of mixing chain with human readable outputs
- Decisions made in stem-aware fashion

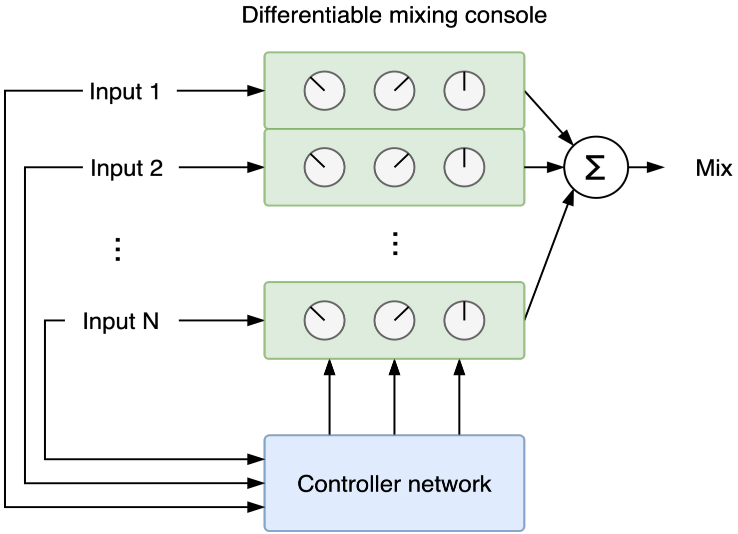


Differentiable mixing console

We could use traditional DSP effects as a strong inductive bias for the mixing task

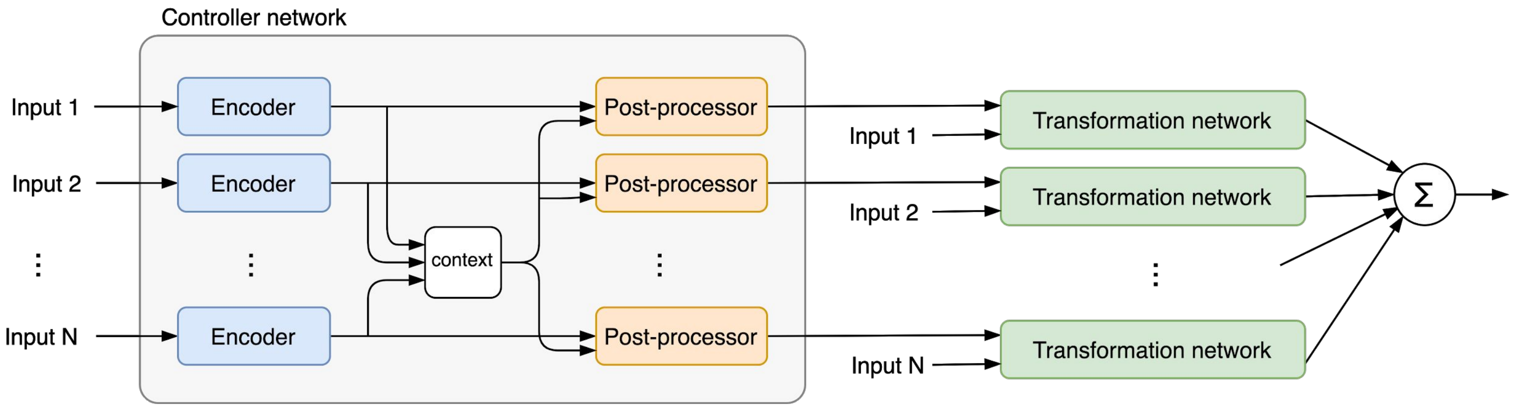


unfortunately, the mixing console is not differentiable



...but we can train a differentiable model to emulate a channel

Differentiable mixing console



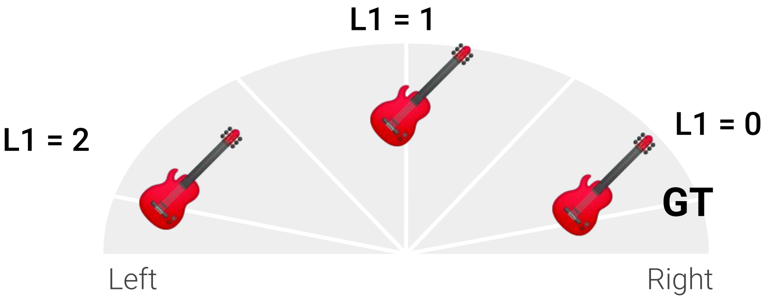
Limited data *Strong inductive bias with **pre-trained** subsystems*

Variable inputs ***Weight sharing** at each subnetwork across input channels*

High fidelity *Audio processing network operates at **44.1 kHz***

User interaction *Produces common mixing parameters users can **tweak***

Stereo loss function



Panning here is more perceptually similar but gives a higher L1 loss

L1 and L2 loss on stereo signals encourage panning all elements to the center.

$$y_{\text{sum}} = y_{\text{left}} + y_{\text{right}}$$

$$y_{\text{diff}} = y_{\text{left}} - y_{\text{right}}$$

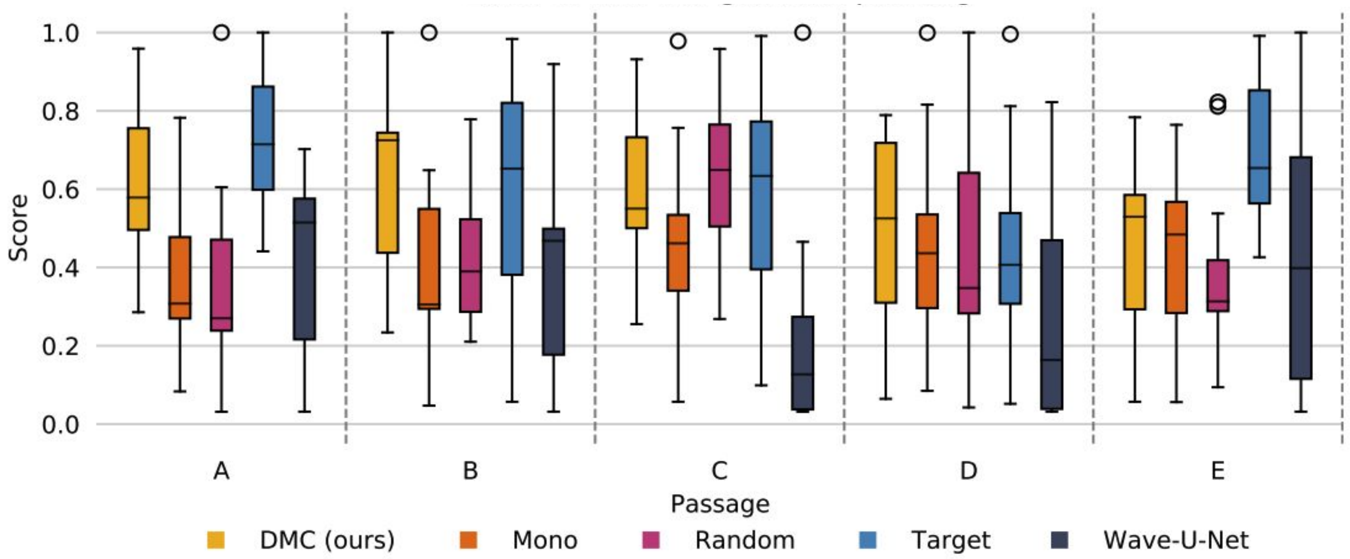
$$\ell_{\text{Stereo}}(\hat{y}, y) = \ell_{\text{MR-STFT}}(\hat{y}_{\text{sum}}, y_{\text{sum}}) + \ell_{\text{MR-STFT}}(\hat{y}_{\text{diff}}, y_{\text{diff}})$$

Achieves invariance to stereo (left-right) orientation

Evaluation of mixes *Loss function that encourages realistic mixes*

Perceptual evaluation

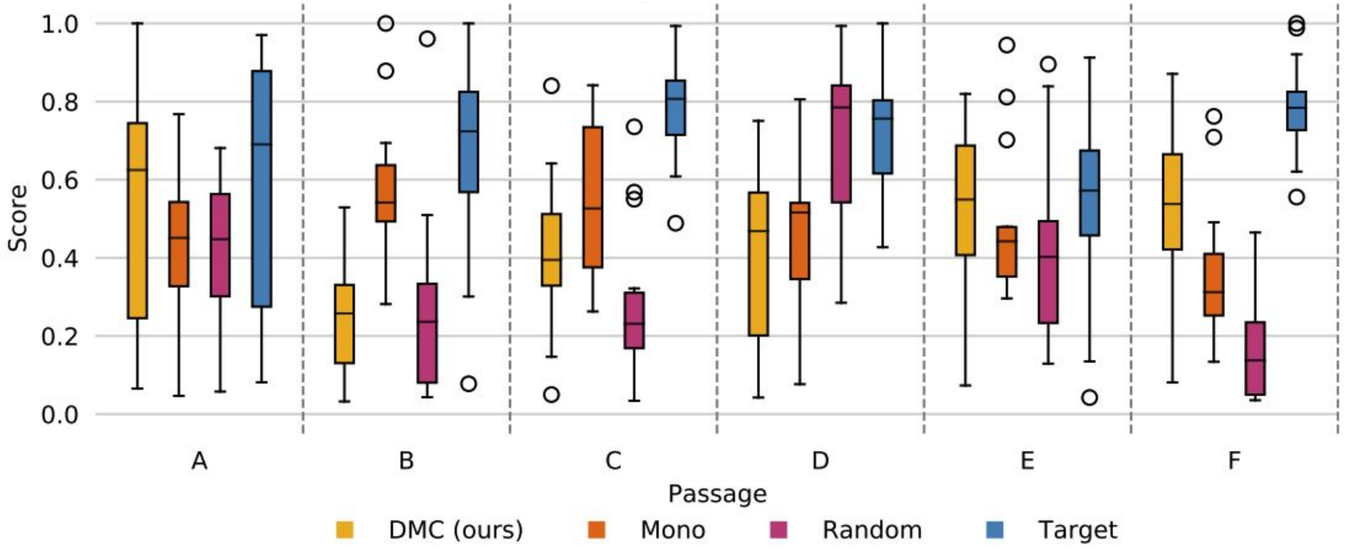
ENST-drums (8 channels)
Gain and panning



On average we outperform the baseline (Mono and Random) mixes.
Wave-U-Net underperform due to artifacts from transposed convolutions.
In some passages, our method (DMC) outperforms the target mixes.

Perceptual evaluation

MedleyDB (6 channels)
Gain + panning
+ EQ + comp. + reverb



We often outperform the baseline (Mono and Random) mixes.
Wave-U-Net completely fails on this task (outputs noise + distortion).

Conclusion

Our approach (**DMC**) is able to learn to produce mixes that exceed the baseline approaches (Mono & Random) directly from uncurated multitrack mix data and waveforms of mixes, without any knowledge of the underlying parameters.

What’s next?

Contact us!

Joseph Colonel
@josephtcolonel
josephtcolonel.com
j.t.colonel@qmul.ac.uk

Christian Steinmetz
@csteinmetz1
christiansteinmetz.com
cjstein@clemson.edu

References

Brecht De Man and Joshua D. Reiss, “A knowledge-engineered autonomous mixing system,” 135th Convention of the Audio Engineering Society, October 2013.

D. Moffat, and M. Sandler, "Machine Learning Multitrack Gain Mixing of Drums," Audio Engineering Society, Engineering Brief 527, (2019 October.)

Stoller, D., S. Ewert, and S. Dixon. "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation." ISMIR. 2018.

Gillet, Olivier, and Gaël Richard. "ENST-Drums: an extensive audio-visual database for drum signals processing." ISMIR. 2006.

M. Martinez, D. Stoller, and D. Moffat "A Deep Learning Approach to Intelligent Drum Mixing with the Wave-U-Net" Journal of the Audio Engineering Society, Accepted Manuscript

Engel, Jesse, Chenjie Gu, and Adam Roberts. "DDSP: Differentiable Digital Signal Processing." International Conference on Learning Representations. 2019.

Barchiesi, Daniele, and Joshua Reiss. "Reverse engineering of a mix." Journal of the Audio Engineering Society 58.7/8 (2010): 563-576.