AES NY 2023
INNOVATE. CREATE. RESONATE.

AES NYC 2023 Workshop • 25 October 2023

AI for Multitrack Music Mixing

Soumya Sai Vanka[1] Christian J. Steinmetz[1] Gary Bromham[1] Marco A. Martínez-Ramírez[2]
Junghyun Koo[3] Brecht De Man[4] Angeliki Mourgela[5]

[1] Centre for Digital Music, Queen Mary University of London
[2] Sony Research, Tokyo, Japan
[3]Music and Audio Research Group, Department of Intelligence and Information, Seoul National University
[4]PXL-Music, Hasselt, Belgium
[5]RoEx

Queen Mary University of London · AIM AI + MUSIC · UKRI UK Research and Innovation · SONY · PXL MUSIC-RESEARCH · roex

Presenters

Soumya Sai Vanka 

 Christian J. Steinmetz

Gary Bromham 

 Marco A. Martínez-Ramírez

Junghyun (Tony) Koo 

 Brecht De Man

Angeliki Mourgela

This session is brought to you by

Technical Committee on Machine Learning and Artificial Intelligence

https://www.aes.org/technical/mlai/

Introduction and Background

Brecht De Man

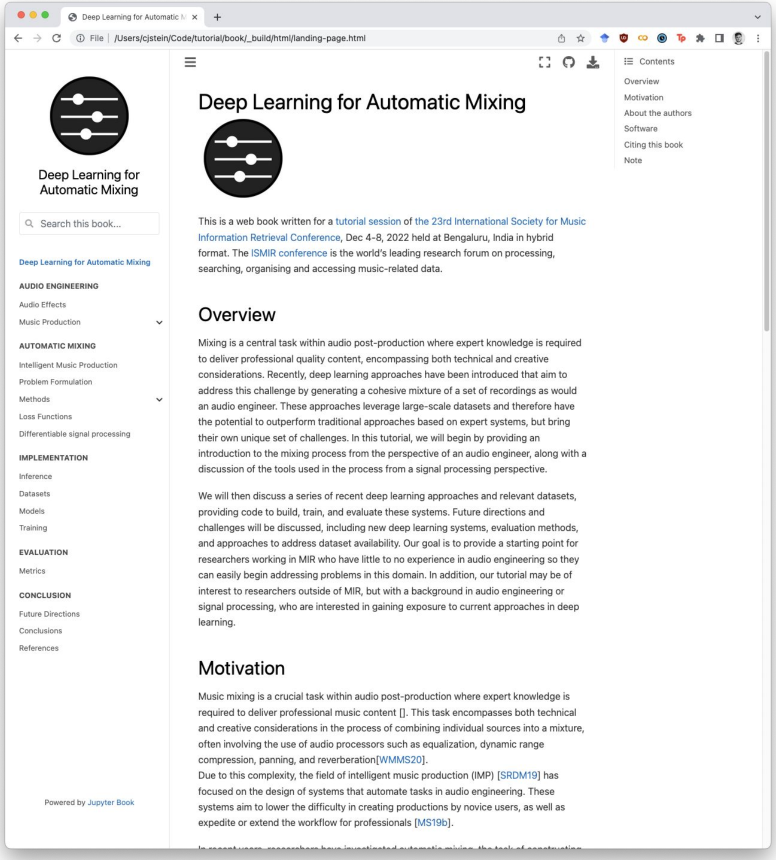
“Hey!” “Hi!”

AI for

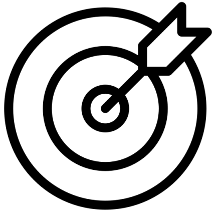
- Multitrack
- Music
- Mixing

Book

https://dl4am.github.io/tutorial



Goals



- Recent advances in large-scale deep learning
  - Differentiable mixing consoles
  - Mixing style transfer
- Importance of
  - Context in mixing
  - Interpretable systems
  - Interactive systems
- Challenges in system design
- Exchange and collaboration

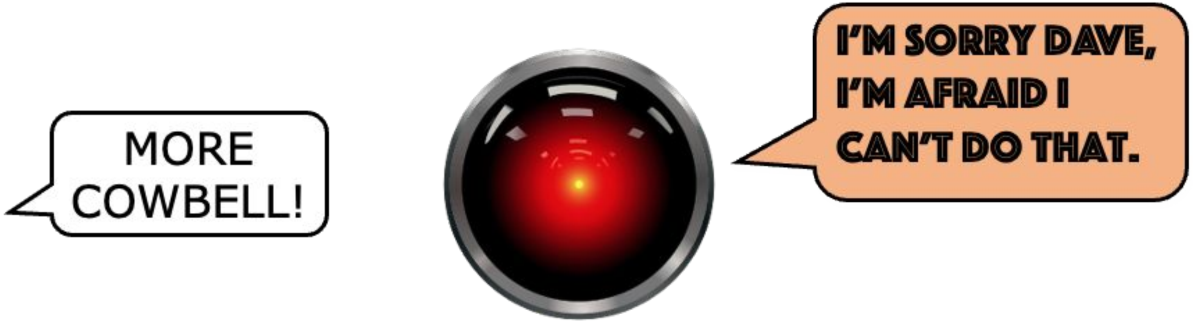
Outline

Context and challenges	Gary
System components	Soumya
Methods	Marco, Tony, Christian
Automixing As Technology	Angeliki
Conclusion and Demonstrations	
Questions	You!

Y tho

Not so fast

Resistance is ~~futile~~ COMMON



- Job security
- Sameness
- Copyright
- Ownership
- Lack of control
- …

PES (Photography Engineering Society)

Learn all about:

- Auto-focus
- Auto-exposure
- Auto-flash
- Stabiliser
- Face detection
- Smile detection
- …

PES (Photography Engineering Society)

- Amateur: No expertise required
- Professional: Increase productivity

Focus on creative aspects

Increased demand

- Man-made, linear, recorded music
- Live music
- Interactive music
- Generative music

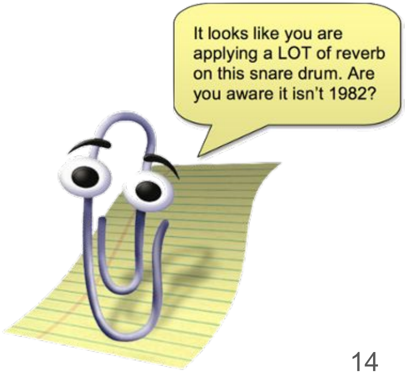
AI comes in many forms

"The Black Box"

"The Assistant"

"The Smart Interface"

"The Diagnostician"



History

Dan Dugan, “Automatic Microphone Mixing,”
Journal of the Audio Engineering Society, vol. 23, July/August 1975.

Automatic Microphone Mixing*

DAN DUGAN

San Francisco, Calif. 94108

A method of analysis of sound reinforcement problems by means of active and passive speech zones is outlined. The need for automatic control of multimicrophone systems is defined, along with the problems associated with the use of voice-operated switches (VOX). Adaptive threshold gating is proposed as the best solution to the problem of active microphone detection. The development and performance of two effective automatic control systems is described.

Enrique Perez Gonzalez and Joshua D. Reiss, “Automatic Mixing: Live Downmixing Stereo Panner,” 10th Int. Conf. on Digital Audio Effects, 10–15 September 2007.

Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France, September 10-15, 2007

AUTOMATIC MIXING: LIVE DOWNMIXING STEREO PANNER

Enrique Perez Gonzalez and Joshua Reiss

Centre for Digital Music,
Queen Mary University of London, Electronic Engineering,
Mile End Road, E1 4NS
London, United Kingdom
enrique.perez@elec.qmul.ac.uk
josh.reiss@elec.qmul.ac.uk

ABSTRACT

An automatic stereo panning algorithm intended for live multi-track downmixing has been researched. The algorithm uses spectral analysis to determine the panning position of sources. The method uses filter bank quantitative channel dependence, priority channel architecture and constrained rules to assign panning criteria. The algorithm attempts to minimize spectral masking by allocating similar spectra to different panning spaces. The algorithm has been implemented; results on its convergence, automatic panning space allocation, and left-right inter-channel phase relationship are presented.

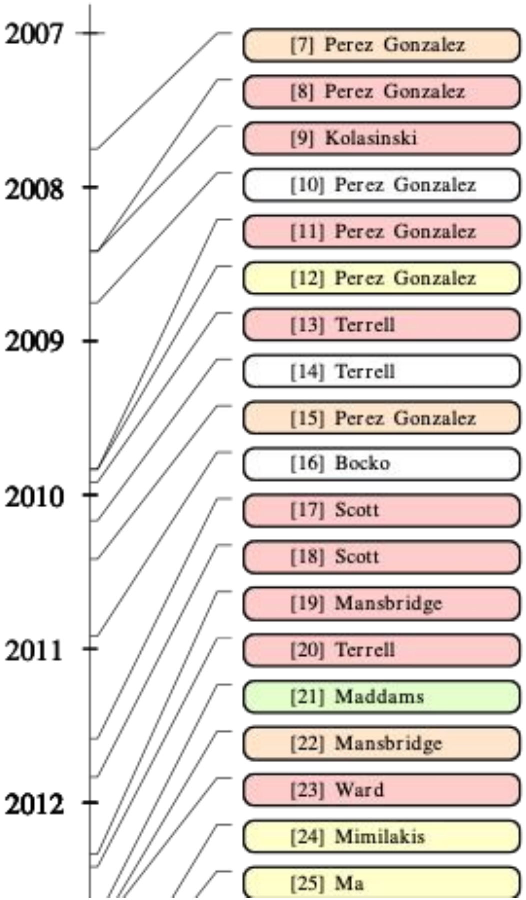
1. INTRODUCTION

An audio engineer carefully handcrafts the characteristics of mul-

This autonomous process can be treated as a constrained rule problem in which the design of the control rules determines the process to be applied to the input signals. The automated process, on the other hand, is the result of playing back in sequence a series of user recorded actions. This involves playing back previously recorded and stored actions, regardless of whether automatically or manually generated.

A common task in live mixing is downmixing a series of mono inputs into a two channel stereo mix. For doing this the input channels get summed into a Left (L) and a Right (R) channel. The proportion at which these multiple mono inputs are added to each L and R channels are responsible for the perceived stereo image. Previous related work on downmixing for spatial audio coding, from 5.1 surround to 2.0 stereo, has been attempted by [4]. Proc-

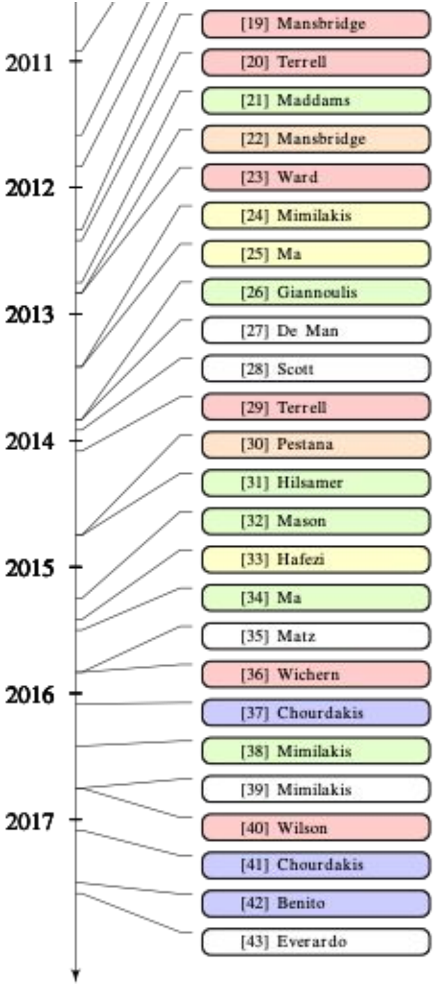
History 2007-2012



(14)

Brecht De Man, Ryan Stables and Joshua D. Reiss, "Ten Years of Automatic Mixing," Proceedings of the 3rd Workshop on Intelligent Music Production, Salford, UK, 15 September 2017.

History 2012-2017

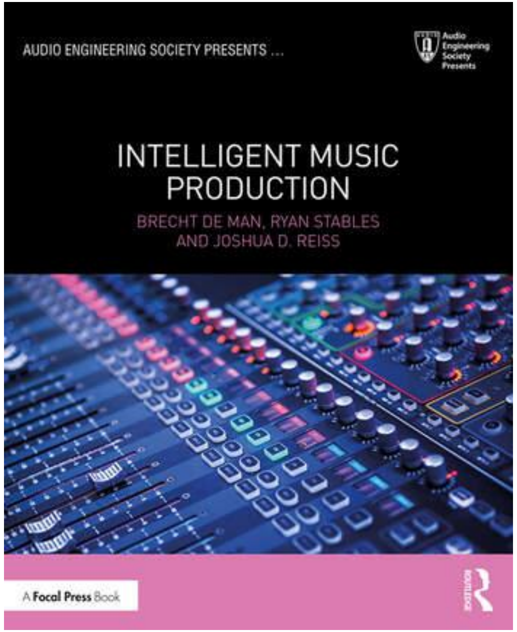


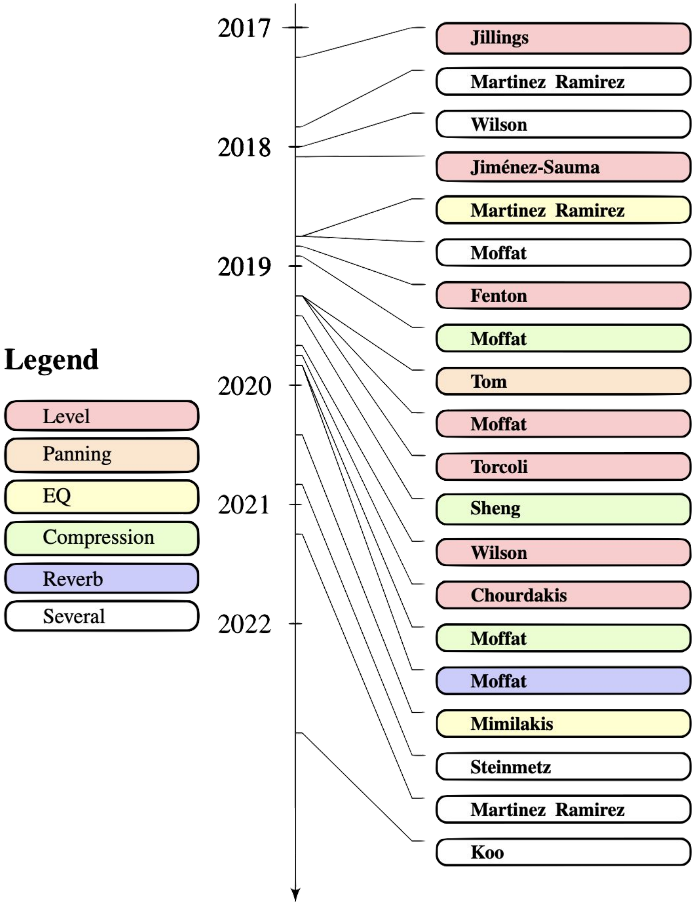
(23)

Brecht De Man, Ryan Stables and Joshua D. Reiss, “Ten Years of Automatic Mixing,” Proceedings of the 3rd Workshop on Intelligent Music Production, Salford, UK, 15 September 2017.

History 2017-2023

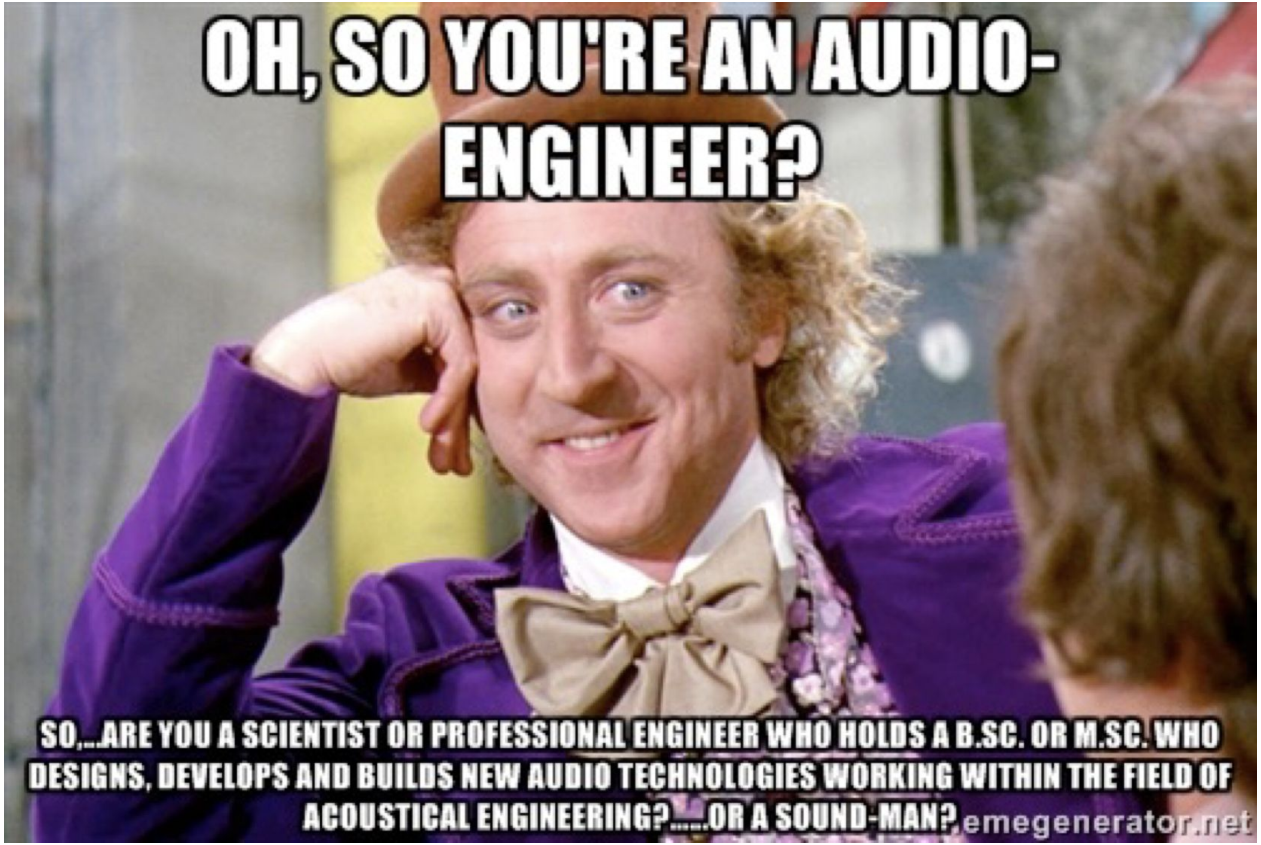
https://csteinmetz1.github.io/AutomaticMixingPapers/





Context and Challenges

Gary Bromham


OH, SO YOU'RE AN AUDIO-ENGINEER?
SO,...ARE YOU A SCIENTIST OR PROFESSIONAL ENGINEER WHO HOLDS A B.SC. OR M.SC. WHO DESIGNS, DEVELOPS AND BUILDS NEW AUDIO TECHNOLOGIES WORKING WITHIN THE FIELD OF ACOUSTICAL ENGINEERING?.....OR A SOUND-MAN?
emegenerator.net

What is Mixing?

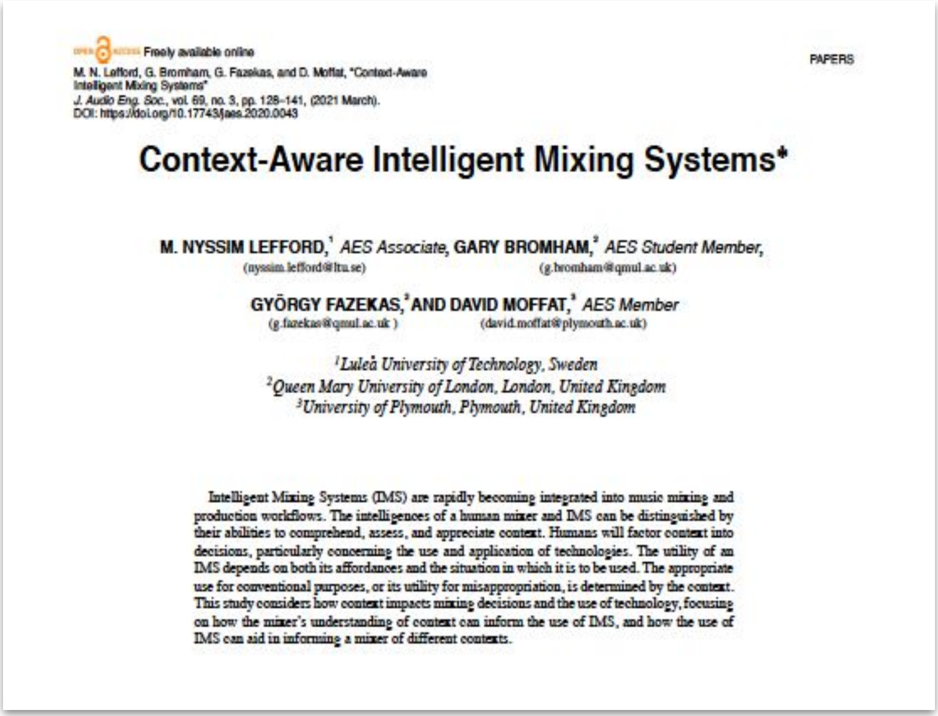
Technical

… a process in which multitrack material – whether recorded, sampled or synthesized–is balanced, treated and combined into a multichannel format.

Artistic

… a less technical definition, one that does justice to music, is that a mix is a sonic presentation of emotions, creative ideas and performance.

Context-Aware Intelligent Mixing Systems (IMS)



Freely available online

M. N. Lefford, G. Bromham, G. Fazekas, and D. Moffat, "Context-Aware Intelligent Mixing Systems"
J. Audio Eng. Soc., vol. 69, no. 3, pp. 128–141, (2021 March).
DOI: https://doi.org/10.17743/jaes.2020.0043

PAPERS

Context-Aware Intelligent Mixing Systems*

M. NYSSIM LEFFORD,[1] *AES Associate*, **GARY BROMHAM,**[2] *AES Student Member*,
(nyssim.lefford@ltu.se) (g.bromham@qmul.ac.uk)

GYÖRGY FAZEKAS,[2] **AND DAVID MOFFAT,**[3] *AES Member*
(g.fazekas@qmul.ac.uk) (david.moffat@plymouth.ac.uk)

[1]*Luleå University of Technology, Sweden*
[2]*Queen Mary University of London, London, United Kingdom*
[3]*University of Plymouth, Plymouth, United Kingdom*

Intelligent Mixing Systems (IMS) are rapidly becoming integrated into music mixing and production workflows. The intelligences of a human mixer and IMS can be distinguished by their abilities to comprehend, assess, and appreciate context. Humans will factor context into decisions, particularly concerning the use and application of technologies. The utility of an IMS depends on both its affordances and the situation in which it is to be used. The appropriate use for conventional purposes, or its utility for misappropriation, is determined by the context. This study considers how context impacts mixing decisions and the use of technology, focusing on how the mixer's understanding of context can inform the use of IMS, and how the use of IMS can aid in informing a mixer of different contexts.

Lefford, M. Nyssim, Gary Bromham, Gyorgy Fazekas, and David Moffat. "Context aware intelligent mixing systems." Journal of the Audio Engineering Society, 2021.

Context and Intelligent Mixing Systems (IMS)

- Technical vs. aesthetic.
- Level of experience? **Am**ateur <> **Pro**fessional-**Am**ateur <> **Pro**fessional.
- Style, genre & taste in mixing.
- Mixing is essentially emotional.
- **IMS** struggles to communicate this.

Experience

Professional <-> **Pro**fessional - **Am**ateur <-> **Am**ateur (Hobbyist)

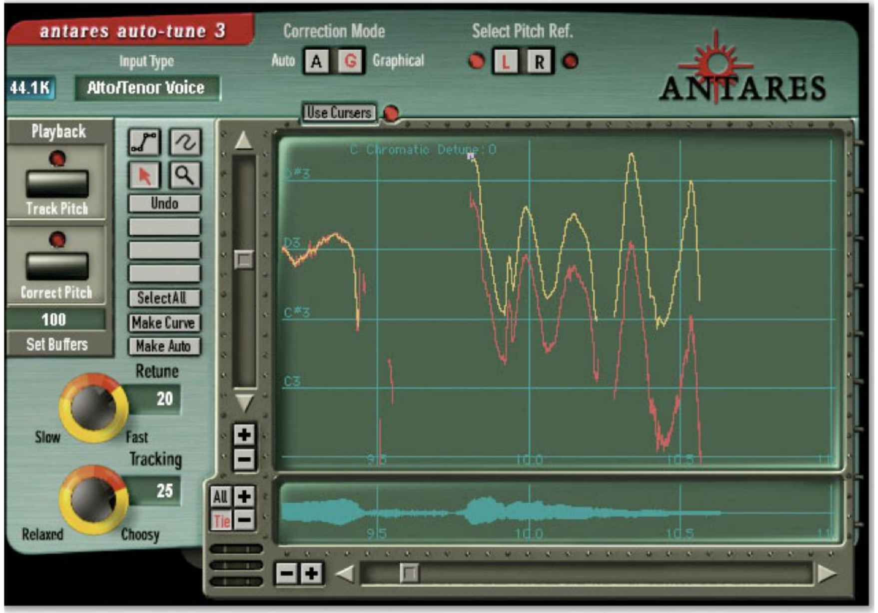
- Three distinct groups in the music production chain. Sandler, M. et al. 2019.
- All three groups have different motivations as mix engineers and producers.
- Intelligent music productions tools are often designed for those with less experience.
- Pro-Am’s who are looking to attain professional-sounding results without much concern for how the goal is achieved.

Conventions and traditional paradigms

- Established conventions and existing workflows
- *“I know what I like and I like what I know”*
- Nostalgia as a motivation for developing tools in a DAW

Misappropriation of Music Production Tools

‘Happy accidents’



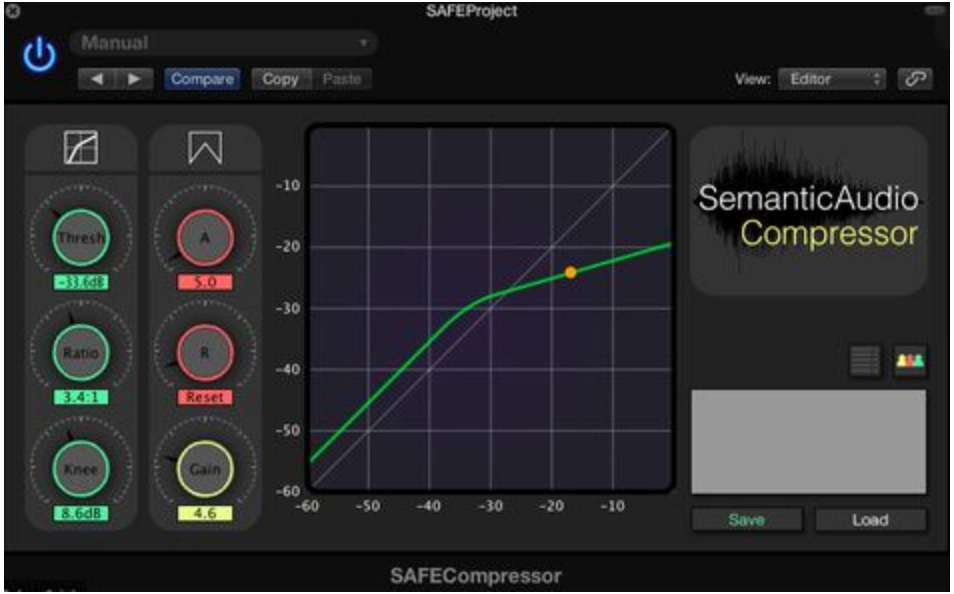
Antares Autotune

The Language of Mixing - Semantics

- ‘Studio Speak’
  - Cross-modal perception.
  - Semantic cross-talk. *Is it warmth or is it muddiness?* Wallmark 2019.
- Connects user input with machine functionality.
- Need for an ontology of audio descriptors which define musical and technical meaning. How can this help IMS? (Intelligent Music Systems)
  - http://www.semanticaudio.co.uk
  - SAFE Plugins. https://somagroup.co.uk/applications/safe-plugins

Waves Parallel Particles

SAFE Compressor



Challenges

- Resistance and aversion to AI-based tools & IMS with mix engineers and producers. Changing mindset.
    - Misconception that it is there to replace rather than assist and augment creative process.
- Limited datasets.
- Controllability
- Musical output can be homogenized and repetitive.

How can we reconcile?

Pros

- Speeds up workflow!
- Takes care of mundane tasks such as editing and labelling
- Presets! We’ve been using them forever anyway!
- Can assist creativity by offering suggestions when engineer lacks inspiration or ideas
- There has always been a resistance to adopt new technology! Get over it!

Cons

- Largely ignores context.
- Creativity often in the outliers in data. ‘Creep’ by Radiohead.
- Mixing is essentially an emotional response or reaction to a piece of music.

Context in Mixing

- Context in mixing could be something as obvious as style or genre or an emotional reaction to a piece of music.

- Mixing is essentially about delivering the emotional context of a musical piece and so far IMS cannot convey this.

Antares Autotune

Context and Intelligent Mixing Systems (IMS)

- Negotiating and reconciling the technical vs. aesthetic domains
- What is the role of experience? Amateur to professional and the emergence of the Pro-am (Professional amateur).
- How do we legislate for style, genre & taste in mixing? Two engineers will hear a mix very differently!
  - Agency, intention and tacit knowledge play a key role.
- Mixing is essentially about delivering the emotional context of a musical piece and so far IMS struggles to communicate this.

Context in Mixing

- Because mixing is a combination of technical and artistic (aesthetic) creative practice and decision-making it attempts to reconcile these two spaces.

- The technical part is much easier to replicate than the latter as it most often doesn’t conform to strict rule sets.

- Intelligent Mixing Systems (IMS) are good at performing perfunctory tasks which adhere to established practices and acquired tacit knowledge but are less good at recognising context which is essentially a human-centric function.

Experience

Professional <-> **Pro**fessional - **Am**ateur <-> **Am**ateur (Hobbyist)

- Three distinct groups in the music production chain.
- All three groups have different motivations as mix engineers and producers.
- Which groups are intelligent tools targeting?
- The interesting case of the Pro-Am’s!

The Language of Mixing

- Semantics - Is it warmth or is it muddiness?
- Language used in a studio has always been confusing.
- Need for descriptors to define musical and technical meaning.
- http://www.semanticaudio.co.uk/

Loudness

- The **average loudness** (LUFS) is computed, then each stem is loudness normalized

EQ

- The **average frequency magnitude spectrum** is computed, then we normalized each stem by performing EQ matching

Panning

- The **average spectral-panning position** is computed, and then we re-pan accordingly

Dynamic Range Compression

- The **average onset peak level** is computed, and we apply a compressor to upper bound the peak levels of the stems

Reverberation

- -A data augmentation approach where **we stochastically add reverberation to already reverberated stems**

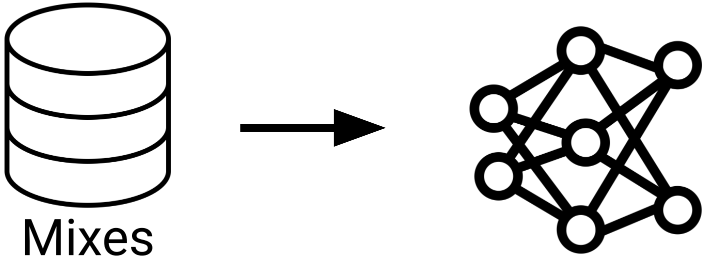
- -Then, the process of learning “the right amount of reverb” is carried out by the network by learning to **filter out the additional reverberation**

Part 2

System Components

Soumya Sai Vanka

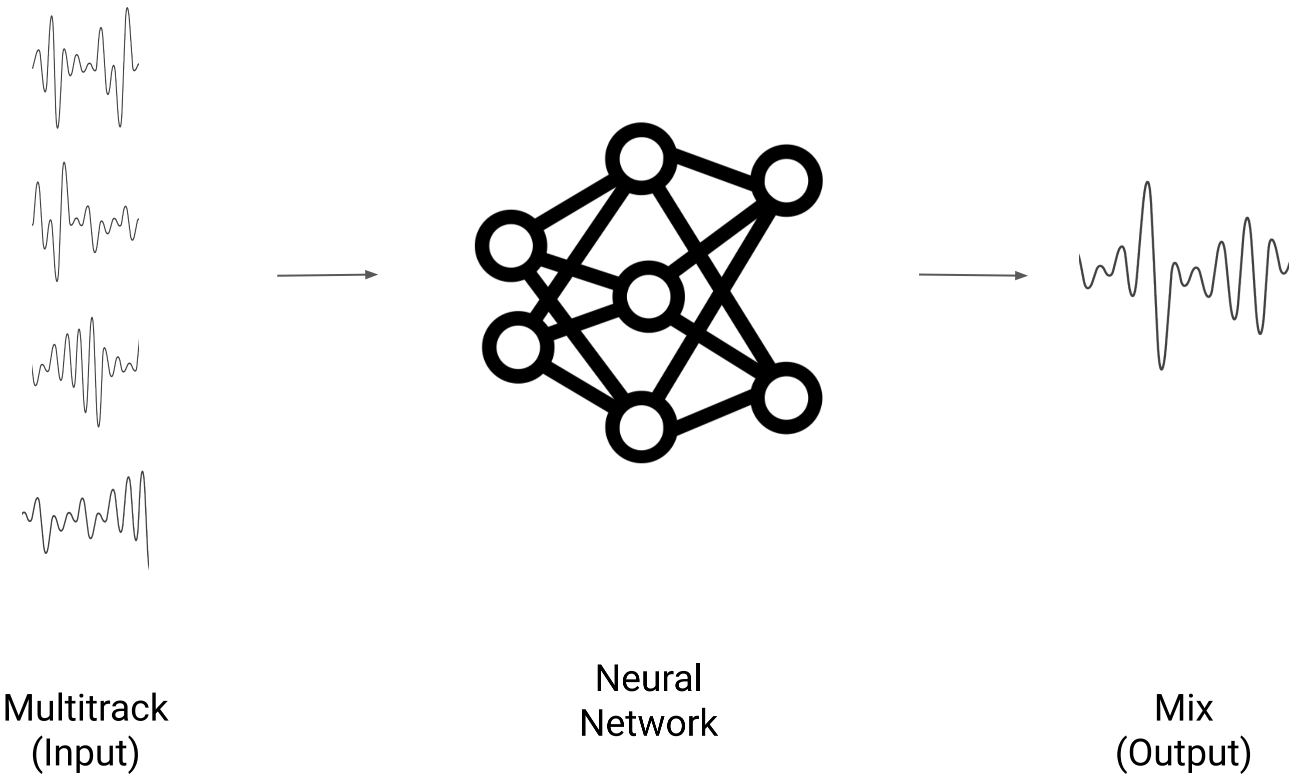
Deep Learning



Mixes

*Can we **learn** to produce mixes directly from data?*

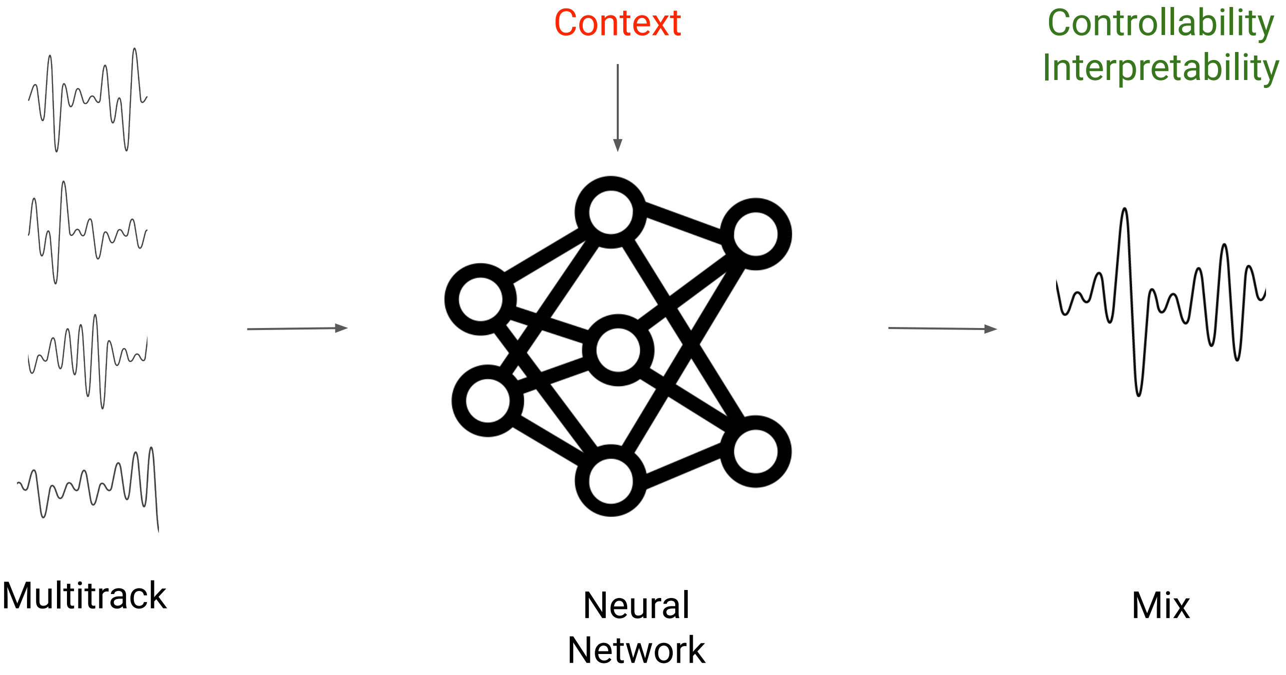
What we want? (at Inference)



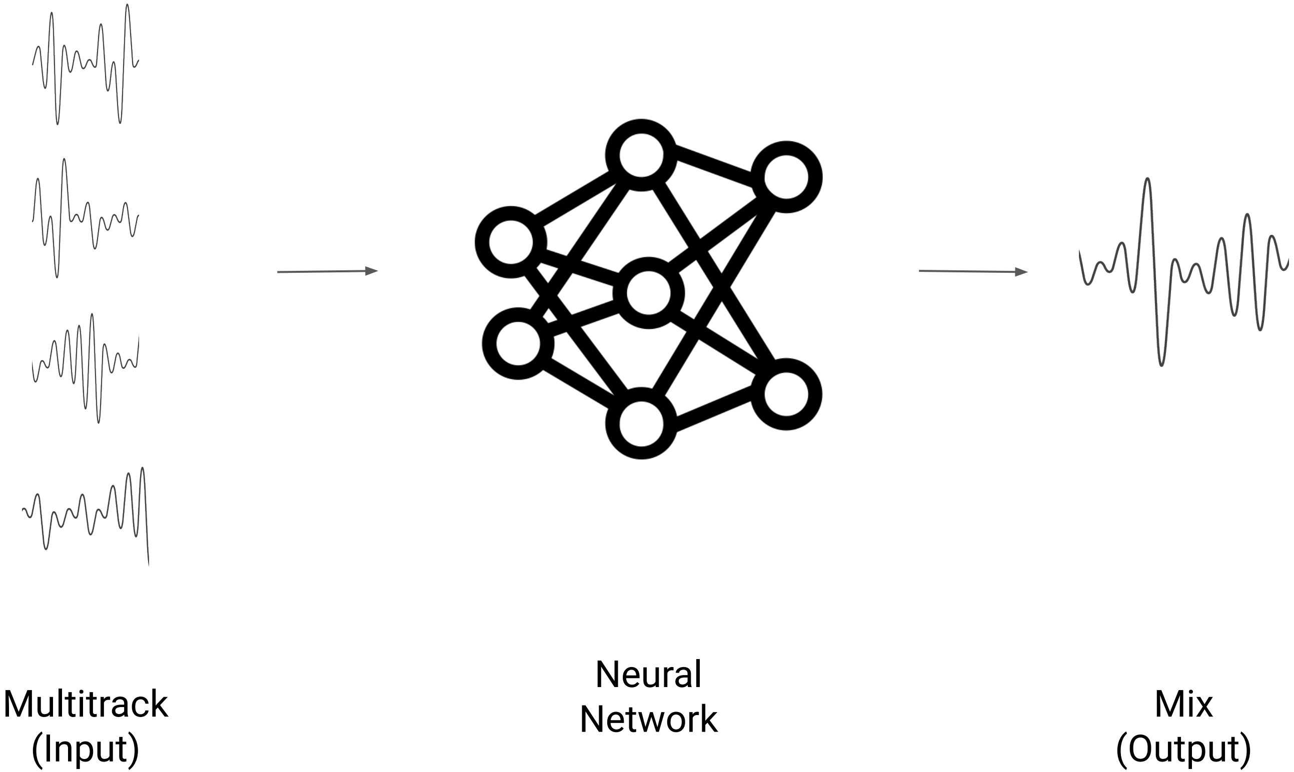
Considerations

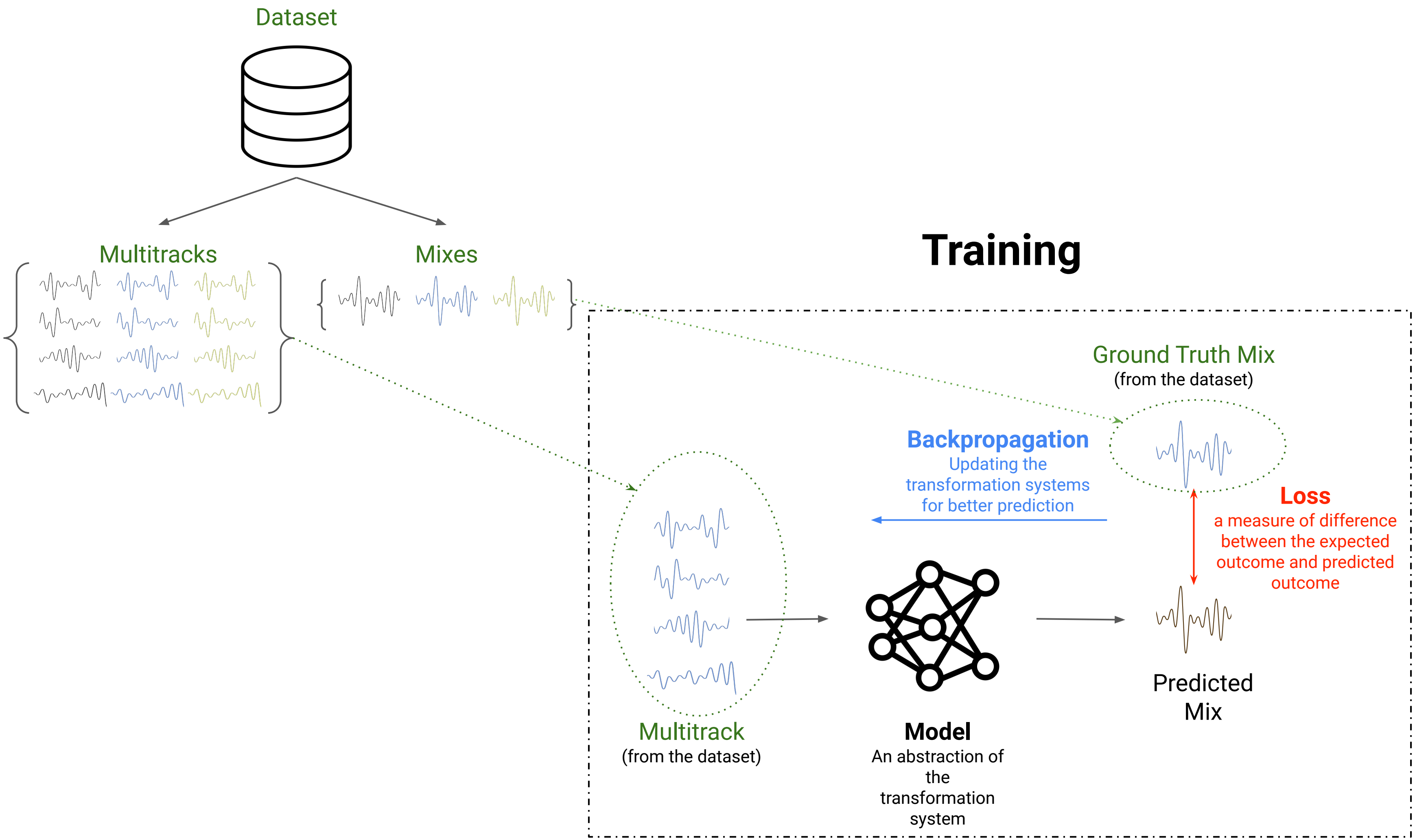
- Interpretability
- Input Taxonomy
- Controllability
- Fidelity
- Context
- Interaction

What we want?


Context
Controllability
Interpretability
Multitrack
Neural
Network
Mix

Let’s begin with simple case




Dataset
Multitracks
Mixes
Training
Ground Truth Mix
(from the dataset)
Backpropagation
Updating the transformation systems for better prediction
Loss
a measure of difference between the expected outcome and predicted outcome
Predicted Mix
Multitrack
(from the dataset)
Model
An abstraction of the transformation system

Popular Multitrack Datasets

ENST-Drums

- 8 channels of drum components
- Recordings by 3 drummers
- Accessible on request
- Size: 1.25 hrs

MedleyDB and Mixing Secrets

- Complete songs with varied number of channels and instruments
- Different Genres
- Medley (7.2hrs) + Mixing Secrets (~50hrs)

MuseDB

- Stems have audio effects applied
- Four stems: Vocals, Bass, Drums, and Others
- Mostly rock, pop, and metal
- ~10hrs

We have very limited open source, time-aligned, real multi-track data capturing various genres and types of music.

Speech recognition: >300 hrs data
Music sequence classification: 280 GB worth data

MoisesDB

MoisesDB is a comprehensive multitrack dataset for source separation beyond 4-stems, comprising 240 previously unreleased songs by 47 artists spanning twelve high-level genres. The total duration of the dataset is 14 hours, 24 minutes and 46 seconds, with an average recording length of 3:36 seconds. MoisesDB is offered free of charge for non-commercial research use only and includes baseline performance results for two publicly available source separation methods.

More datasets

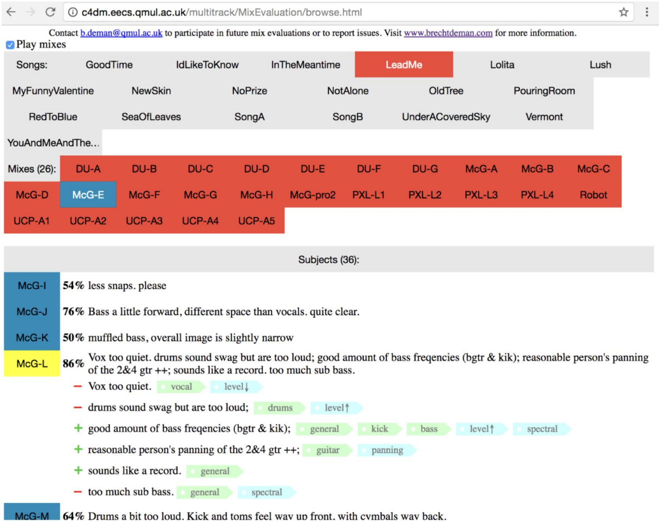
Slakh2100

Manilow, Ethan [1]; Wichern, Gordon [2]; Seetharaman, Prem [1]; Le Roux, Jonathan [2]

Show affiliations

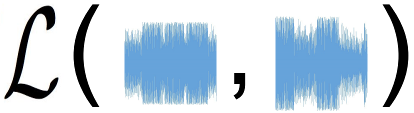
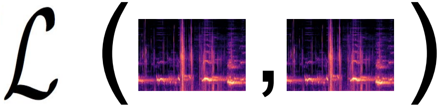
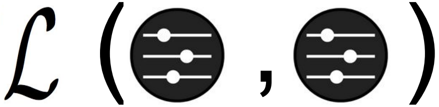
Introduction:

The Synthesized Lakh (Slakh) Dataset is a dataset of multi-track audio and aligned MIDI for music source separation and multi-instrument automatic transcription. Individual MIDI tracks are synthesized from the Lakh MIDI Dataset v0.1 using professional-grade sample-based virtual instruments, and the resulting audio is mixed together to make musical mixtures. This release of Slakh, called Slakh2100, contains 2100 automatically mixed tracks and accompanying, aligned MIDI files, synthesized from 187 instrument patches categorized into 34 classes, totaling 145 hours of mixture data.

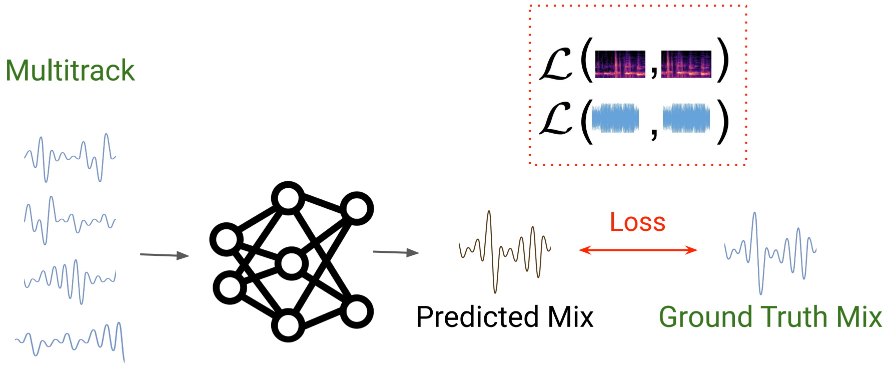


Open Multitrack testbed

Loss functions

Time domain (Audio Loss)	Frequency domain (Audio Loss)	Parameter Loss
		
Audio needs to be time aligned	Need to choose proper scaling that can capture perceptual qualities of sound	Multiple parameter combinations can lead to same result, may penalise the model unnecessarily

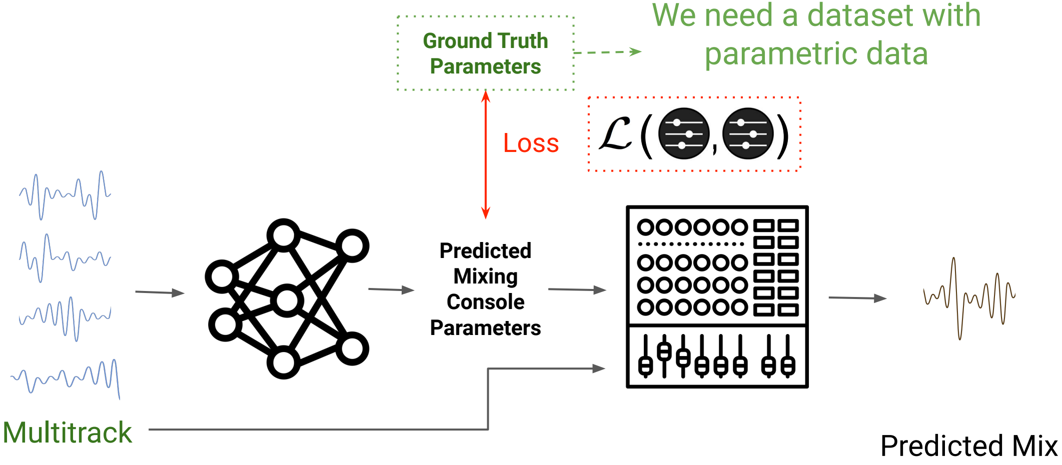
Model Types



Direct Transformation

Black box system that lacks interpretability and controllability (context not incorporated)

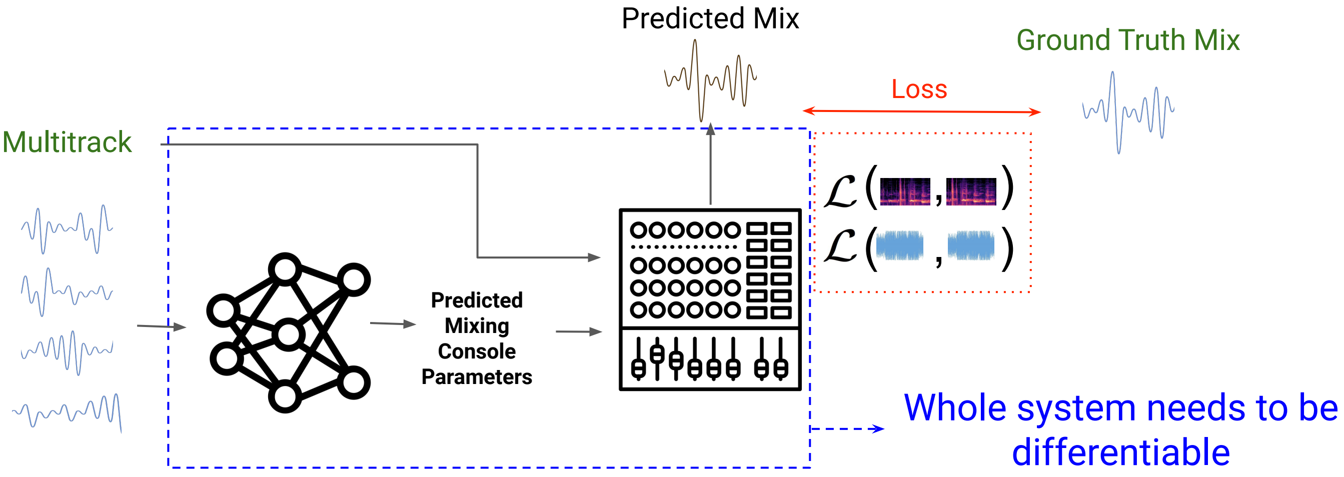
Model Types



Parameter Estimation
(Parameter Loss)

Black box system that allows interpretability and controllability (context not incorporated)

Model Types

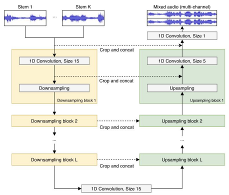


Parameter Estimation
(Audio Loss)

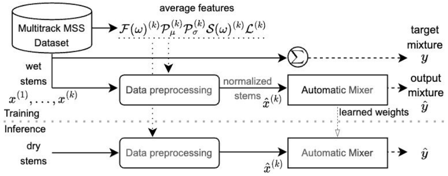
Black box system that allows interpretability and controllability (context not incorporated)

State of the Art

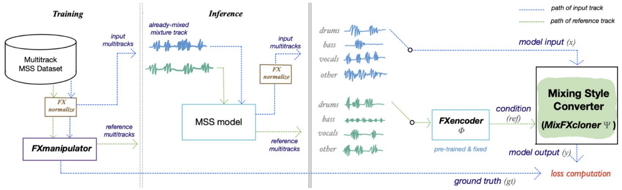
Direct Transformation



Wave-U-Net for drum mixing [a]

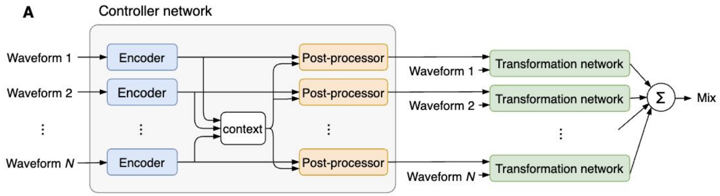


Mixing with out-of-domain data [c]



Mixing style transfer [d]

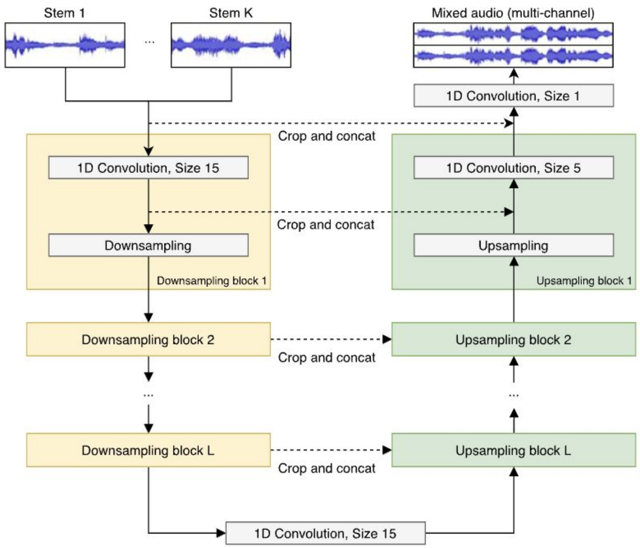
Parameter Estimation

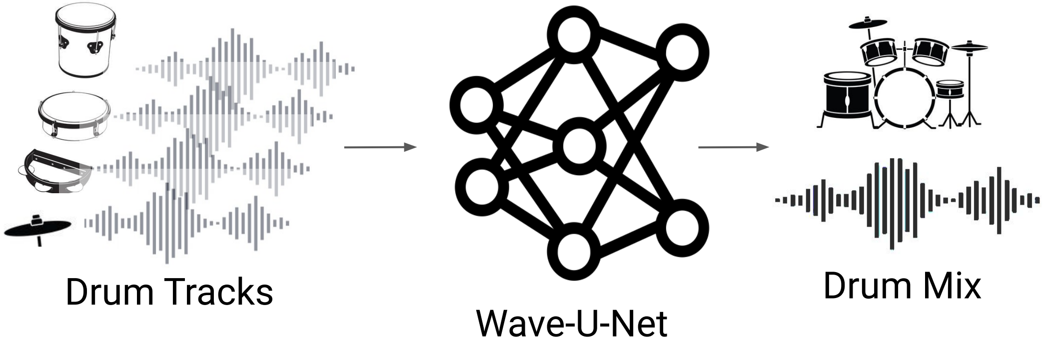


Mixing with neural mixing console [b]

[a] A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net, Martinéz et al. (JAES Mar, 2021)
[b] Automatic multitrack mixing with a differentiable mixing console of neural audio effects, Steinmetz et al. (ICASSP 2021)
[c] Automatic music mixing with deep learning and out-of-domain data, Martinéz et al. (ISMIR 2022)
[d] Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects, Koo et al. (ICASSP 2023)

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

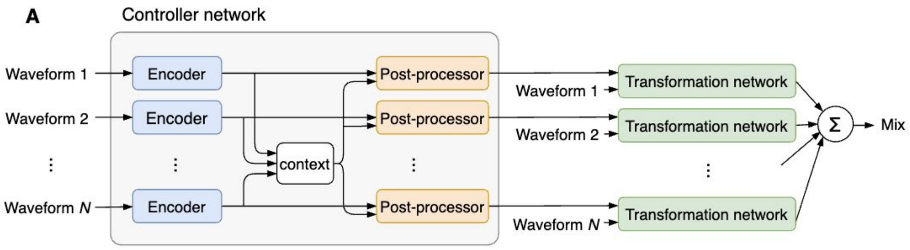




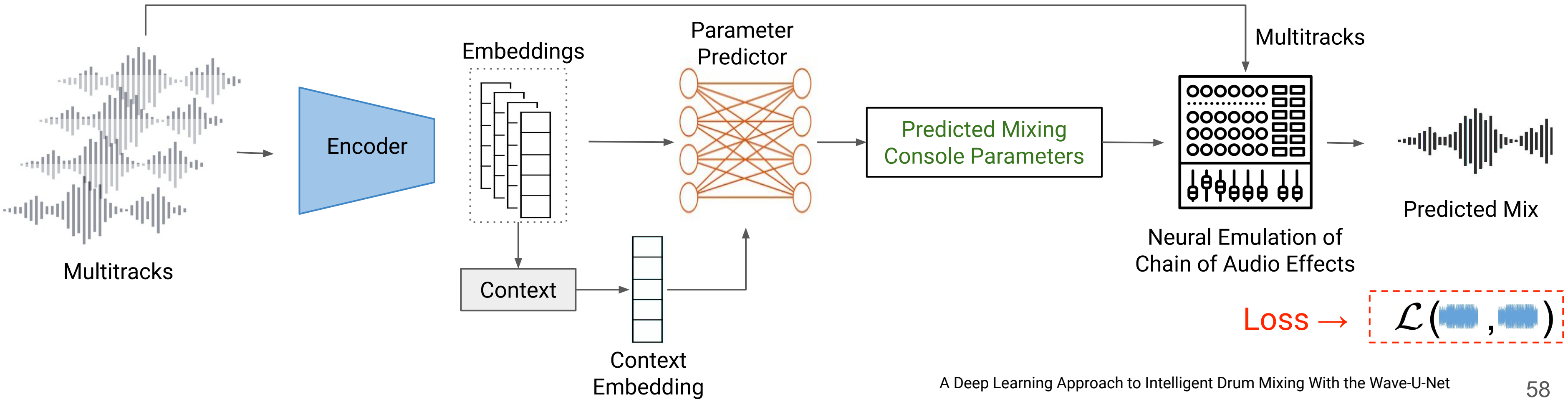
Loss → $\mathcal{L}(\ ,\ )$

- Pros: directly learns the audio transformation
- Limitations: **Only drum mixing**, number of tracks is fixed

Automatic multitrack mixing with a differentiable mixing console of neural audio effects

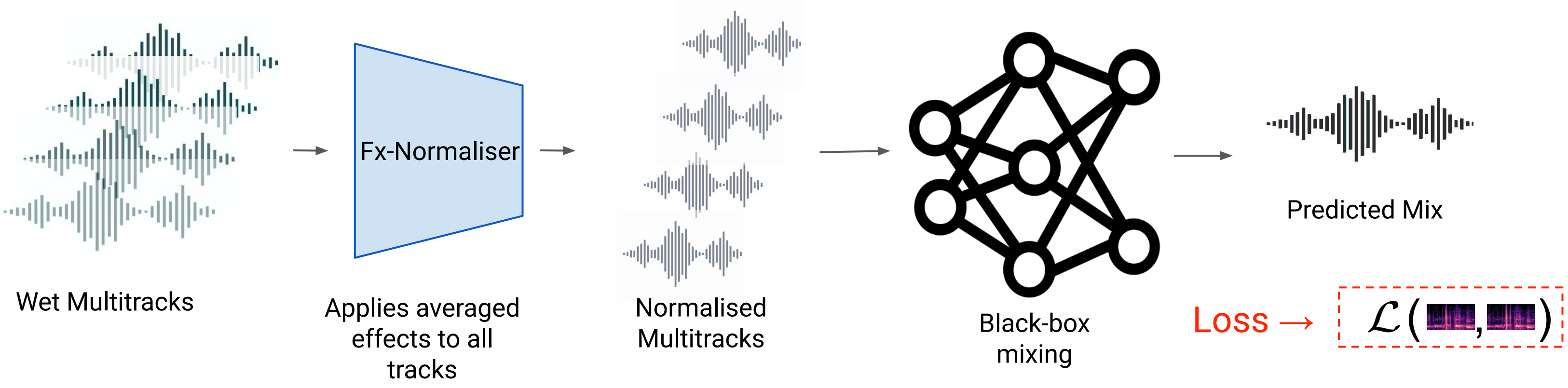


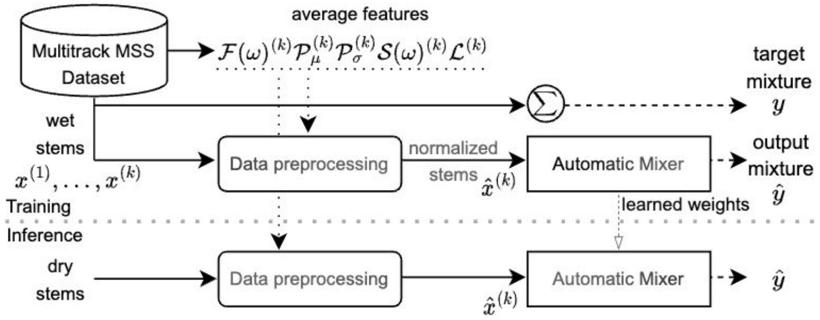
- Pros: Permutation invariant, works for any number of tracks, allows multitrack mixing
- Limitations: neural emulation of effects are difficult to train, **doesn't work well for all cases (Could be due to lack of enough data)**



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

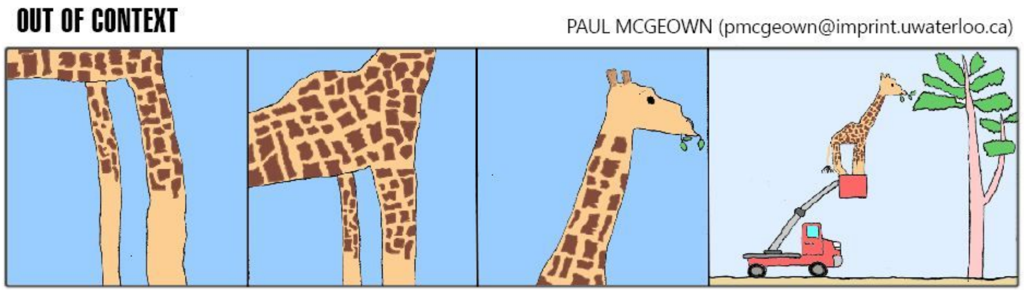
Automatic music mixing with deep learning and out-of-domain data



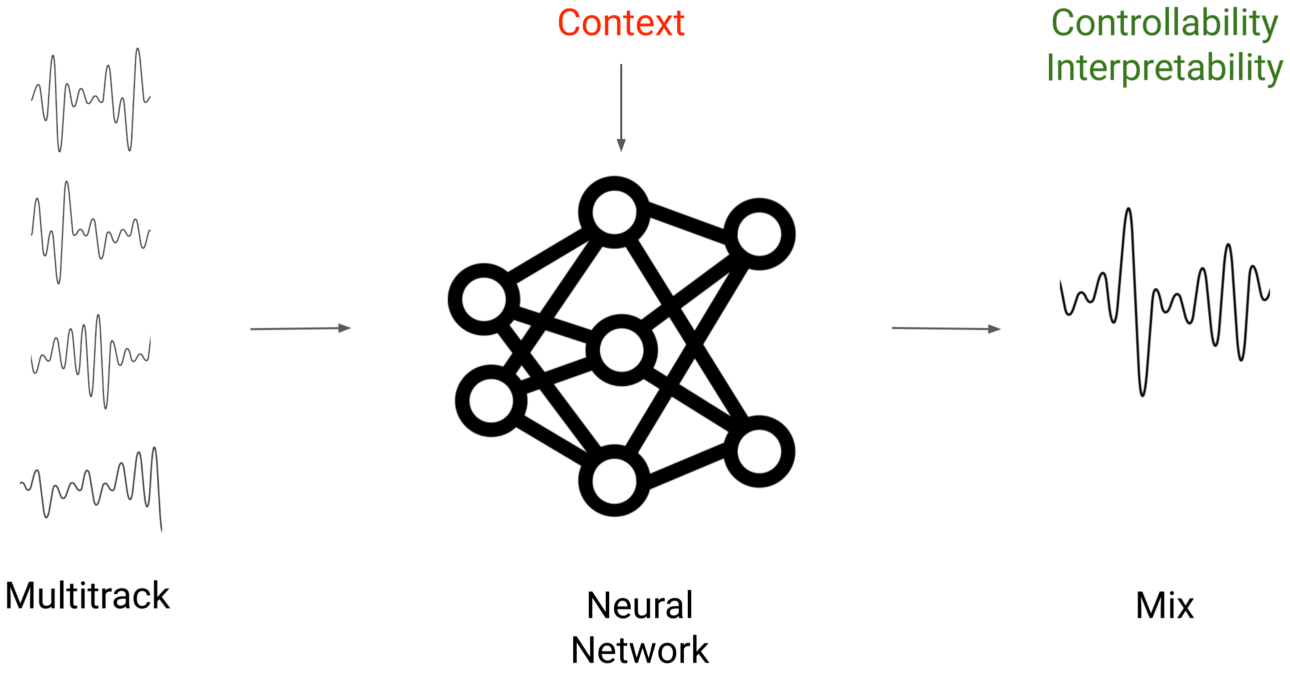


- Pros: uses of wet/processed stems to train, creates possibility for using extensive source separation datasets with wet stems
- Limitations: lacks interpretability and controllability, works for 4 stems

Limitations

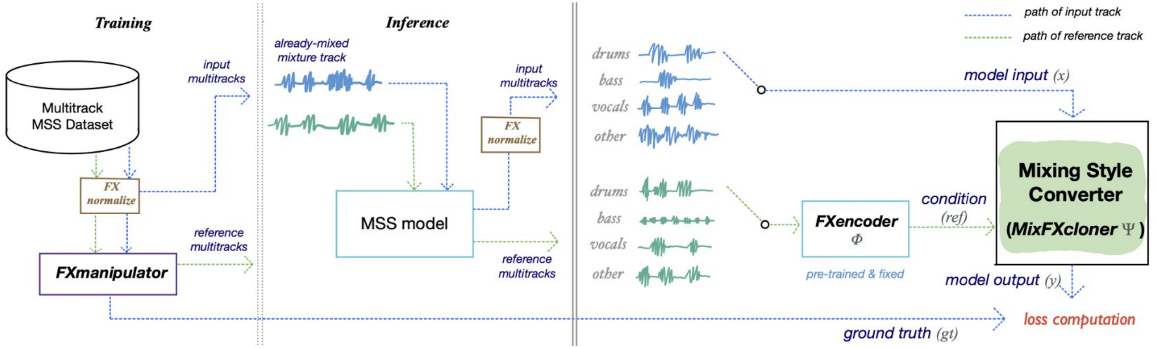


What we want?

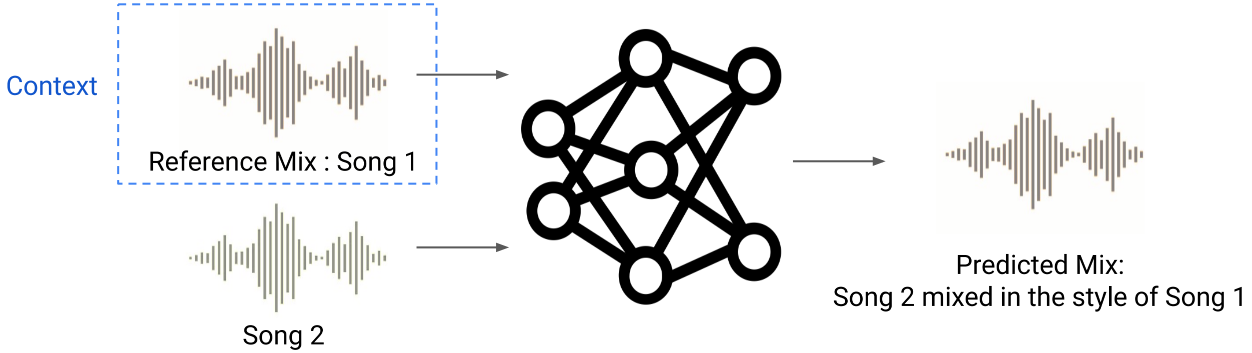


Context
Controllability
Interpretability
Multitrack
Neural
Network
Mix

Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects



- Pros: incorporates context through reference
- Limitations: mix to mix transfer, lacks interpretability

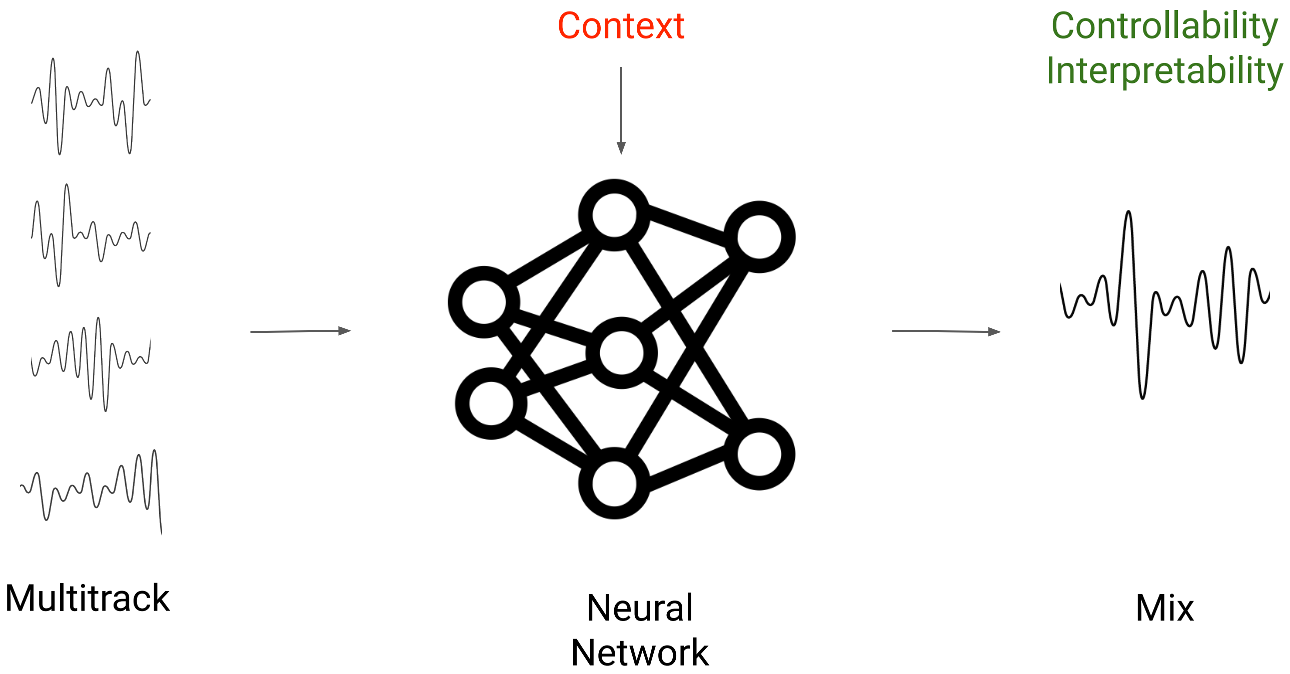


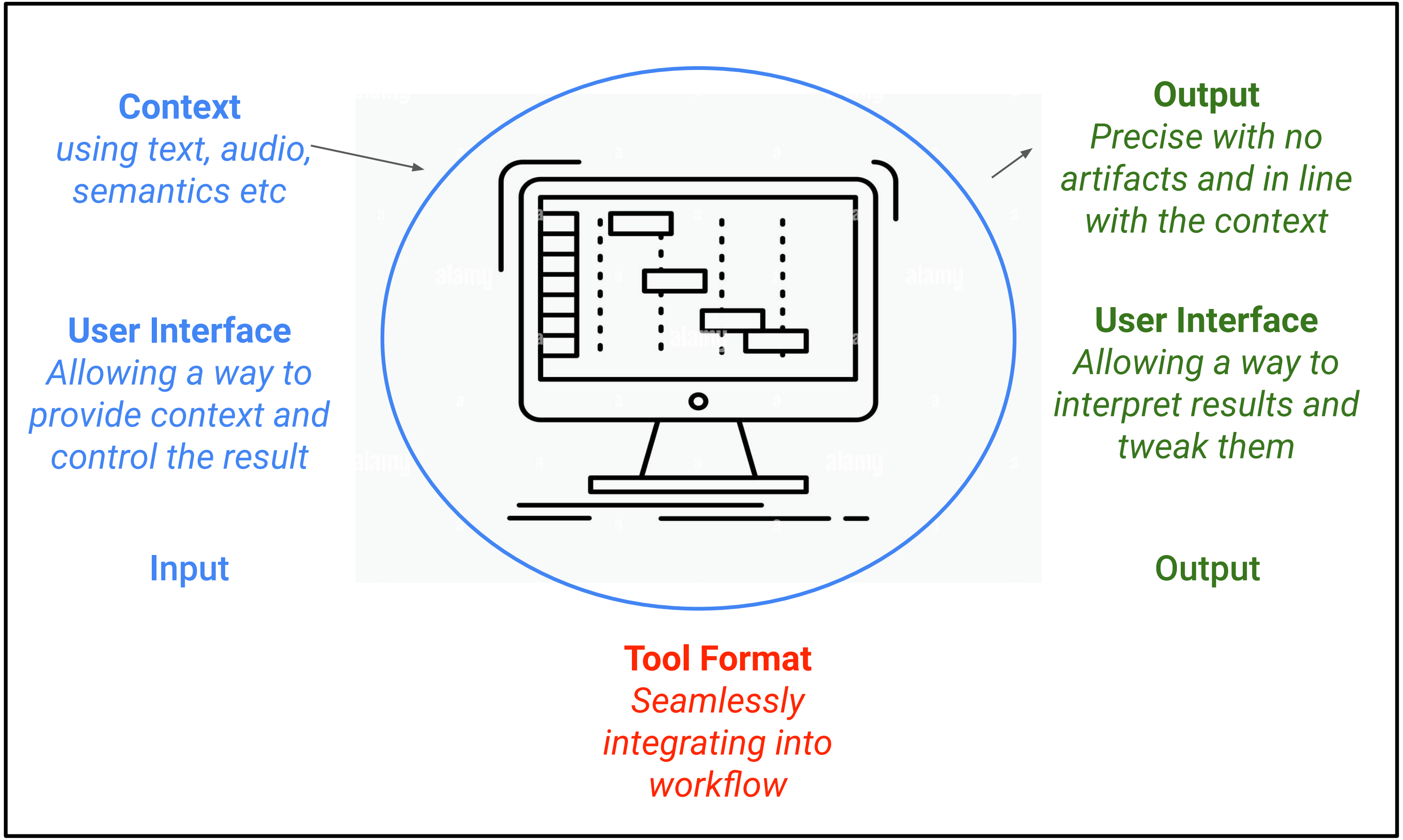
Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects

Summary

Model	System Type	Controllability	Context	Interpretability	Input Taxonomy
Wave-U-Net for drum mixing	Direct transformation	No	No	No	Drums only
Mixing with neural mixing console	Parameter estimation	Yes	No	Yes	Multitrack, permutation and number of tracks invariant
Mixing with out-of-domain data	Direct transformation	No	No	No	Wet stems, limited on number of tracks
Mixing style transfer	Direct transformation	No	Yes (reference song)	Yes	Mix and style reference mix

What’s next?


Context
Controllability
Interpretability
Multitrack
Neural
Network
Mix



Ideal design for an automatic mixing system

Part 3

Methods

 Marco A. Martínez-Ramírez

 Junghyun (Tony) Koo

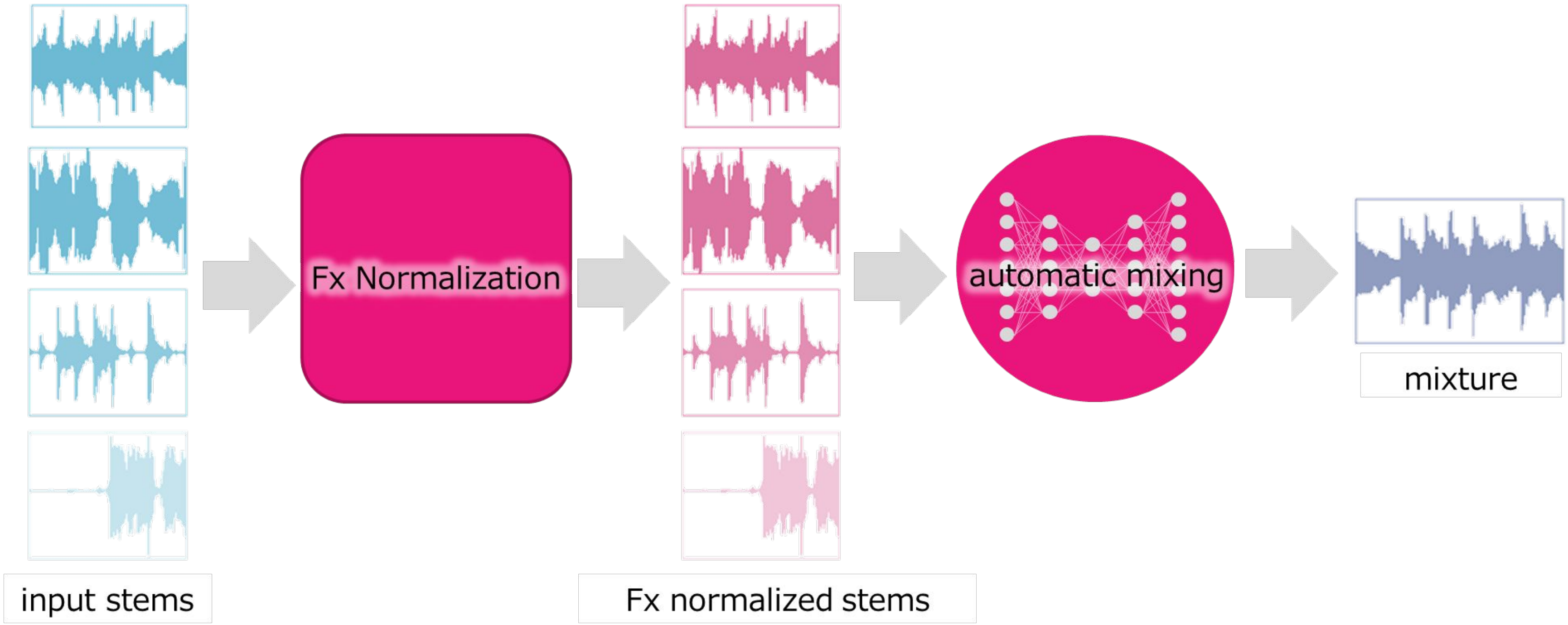
 Christian J. Steinmetz

FX Normalization

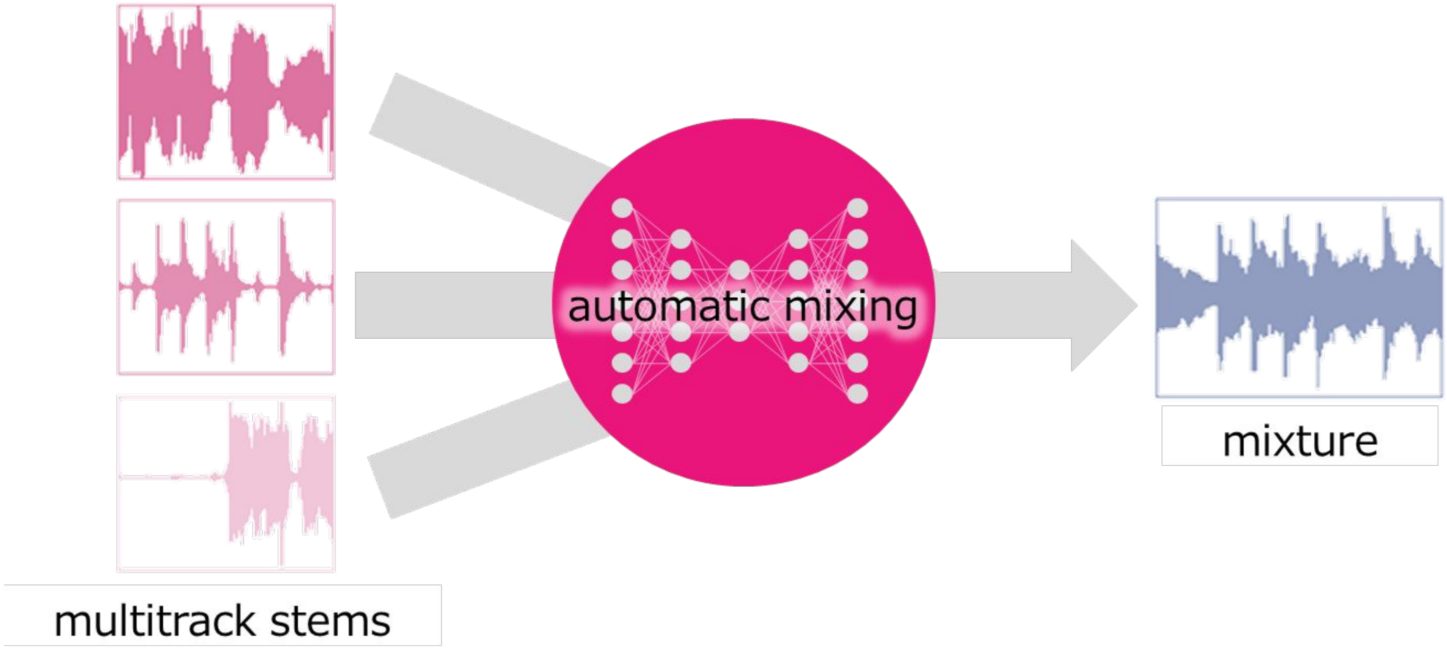
Automatic music mixing with deep learning and out-of-domain data
ISMIR 22 Paper

Marco A. Martínez-Ramírez

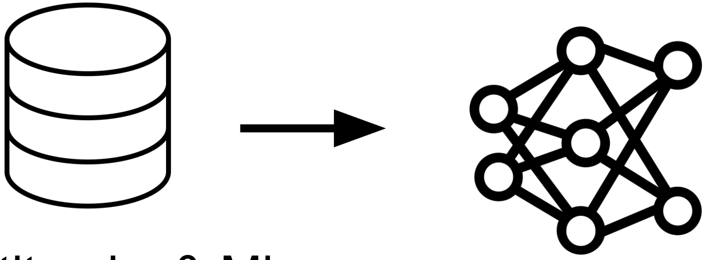
Fx Normalization



Supervised Learning Approach



Challenging



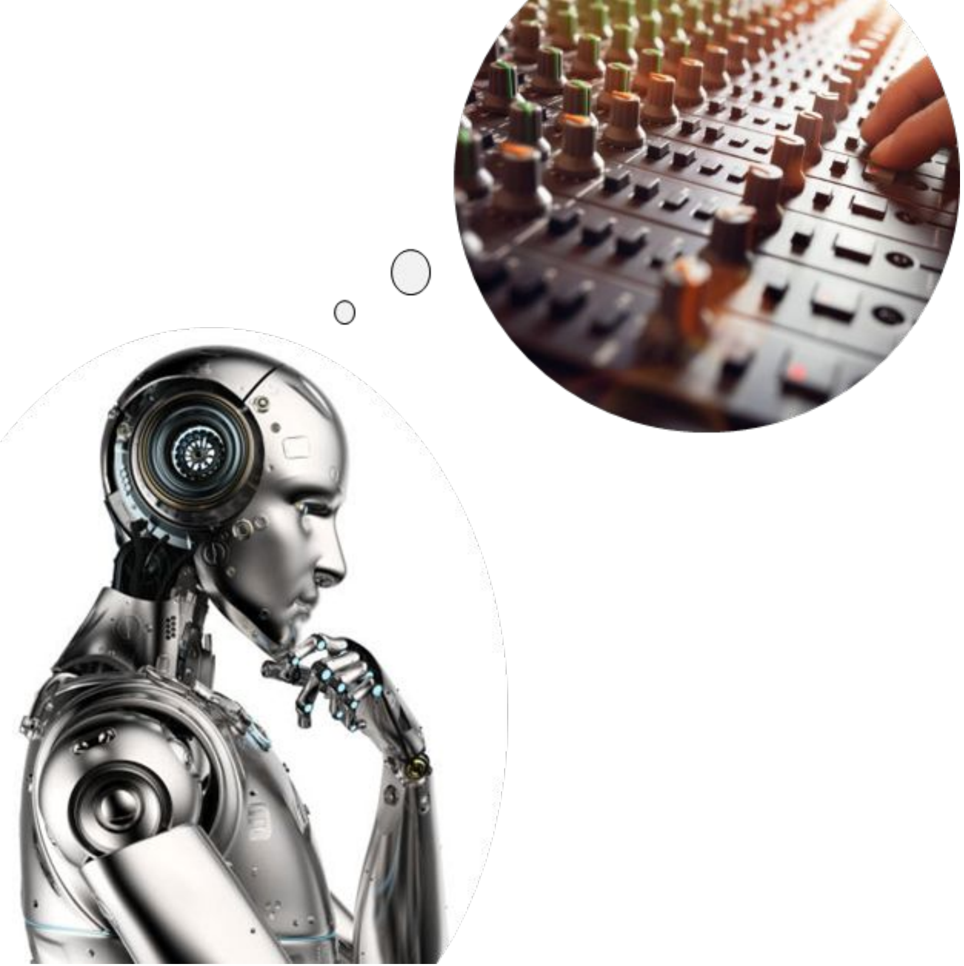
Dry multitracks & Mixes

Data driven approaches need data,
however, ***collecting dry data is difficult***

Previous works

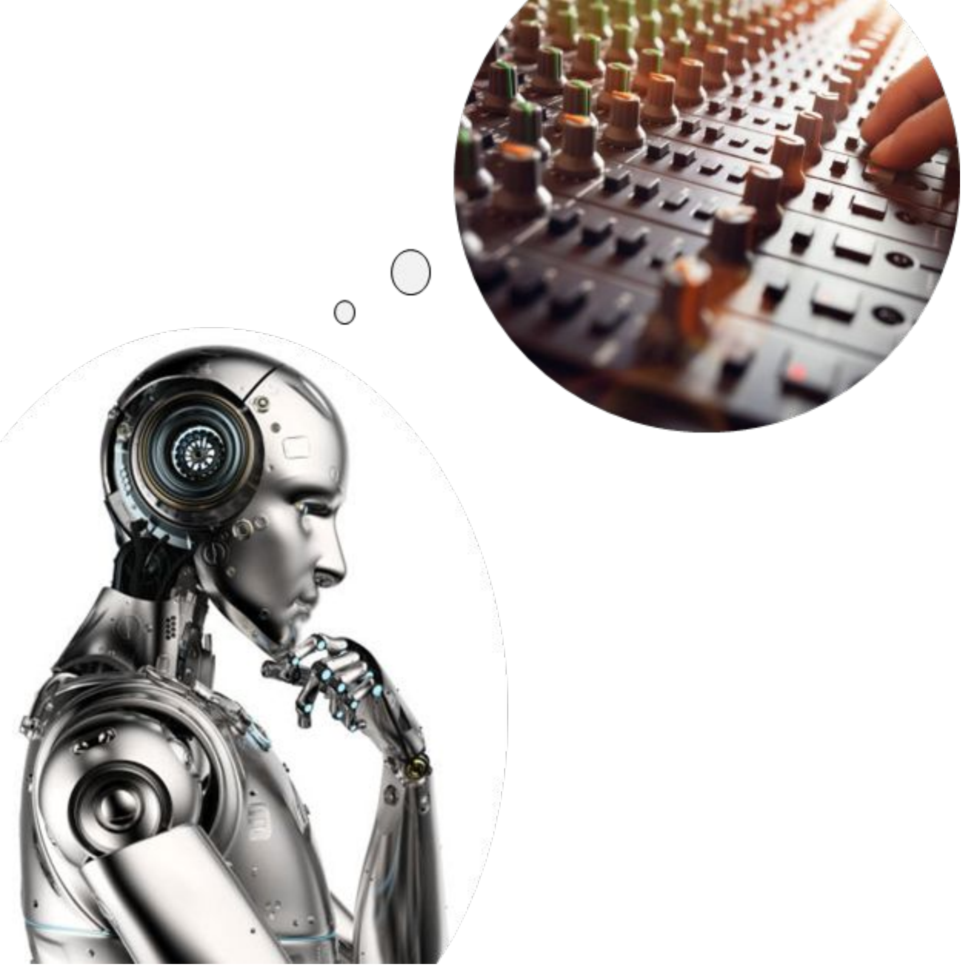
- Previous methods have not yet achieved the level of professional audio engineers mixes

- It has been hypothesized that the **bottleneck of performance can be resolved with a large enough dataset**



Research Question

- ***Can we use wet multitrack music data*** *and* ***repurpose it*** *to train deep learning models that perform automatic music mixing?*

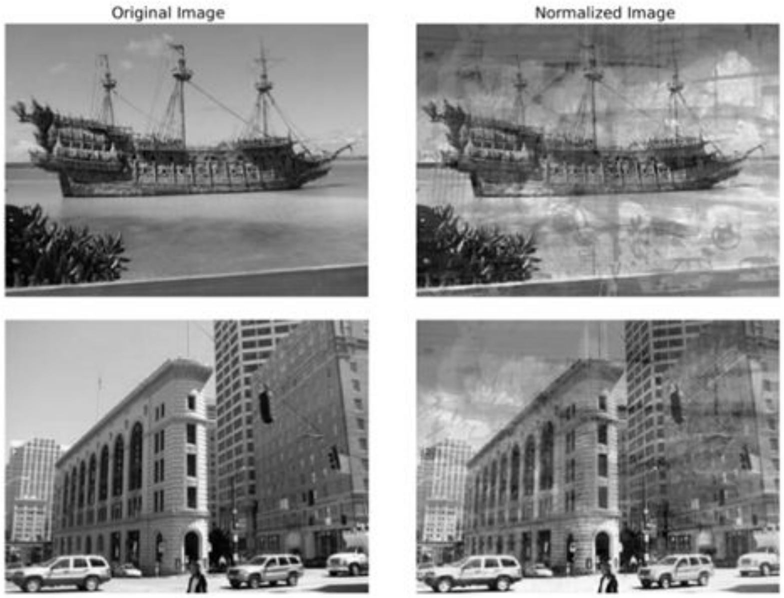


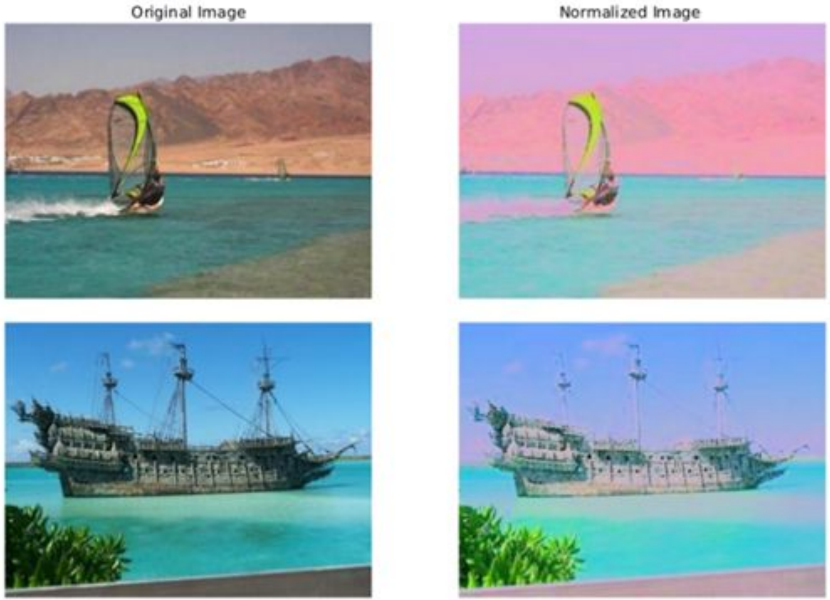
How ?

➢ ***Wet multitracks already contain the desired mixing effects****, which are what the networks need to learn*

Fx Normalization !

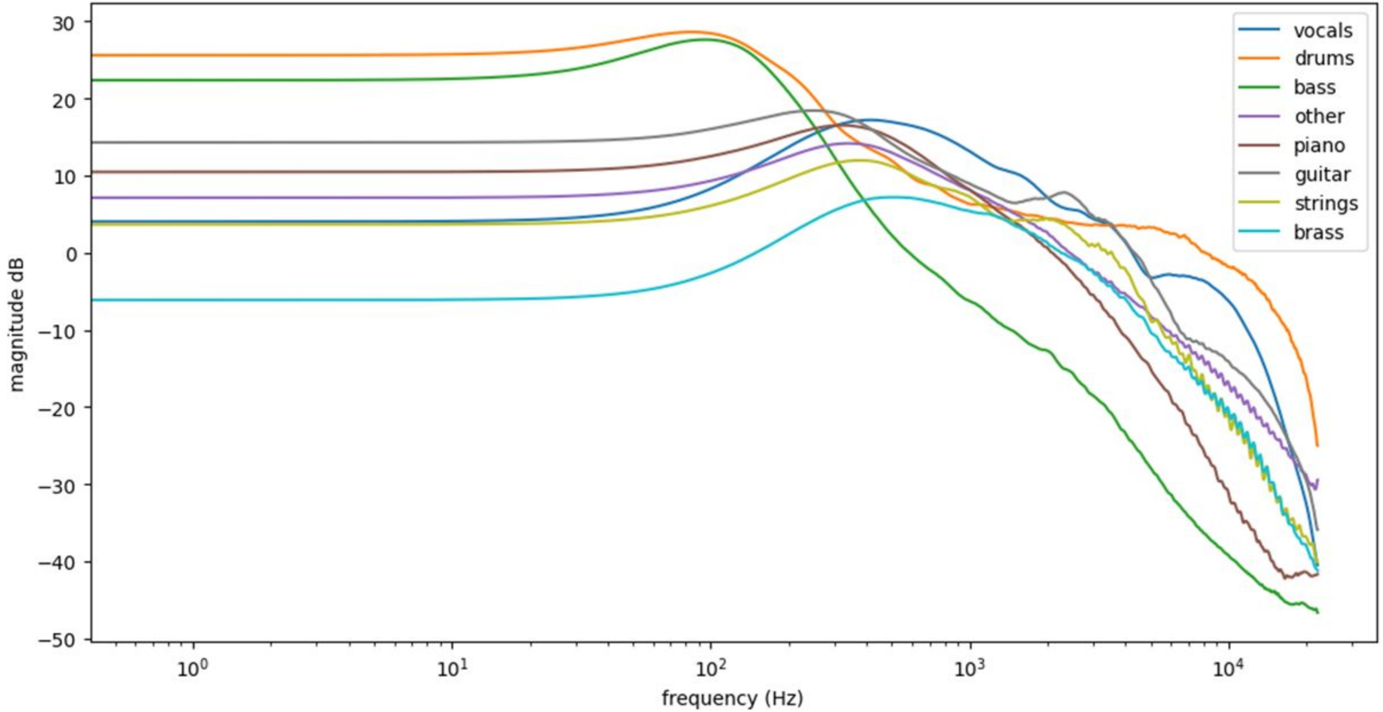
Data Normalization



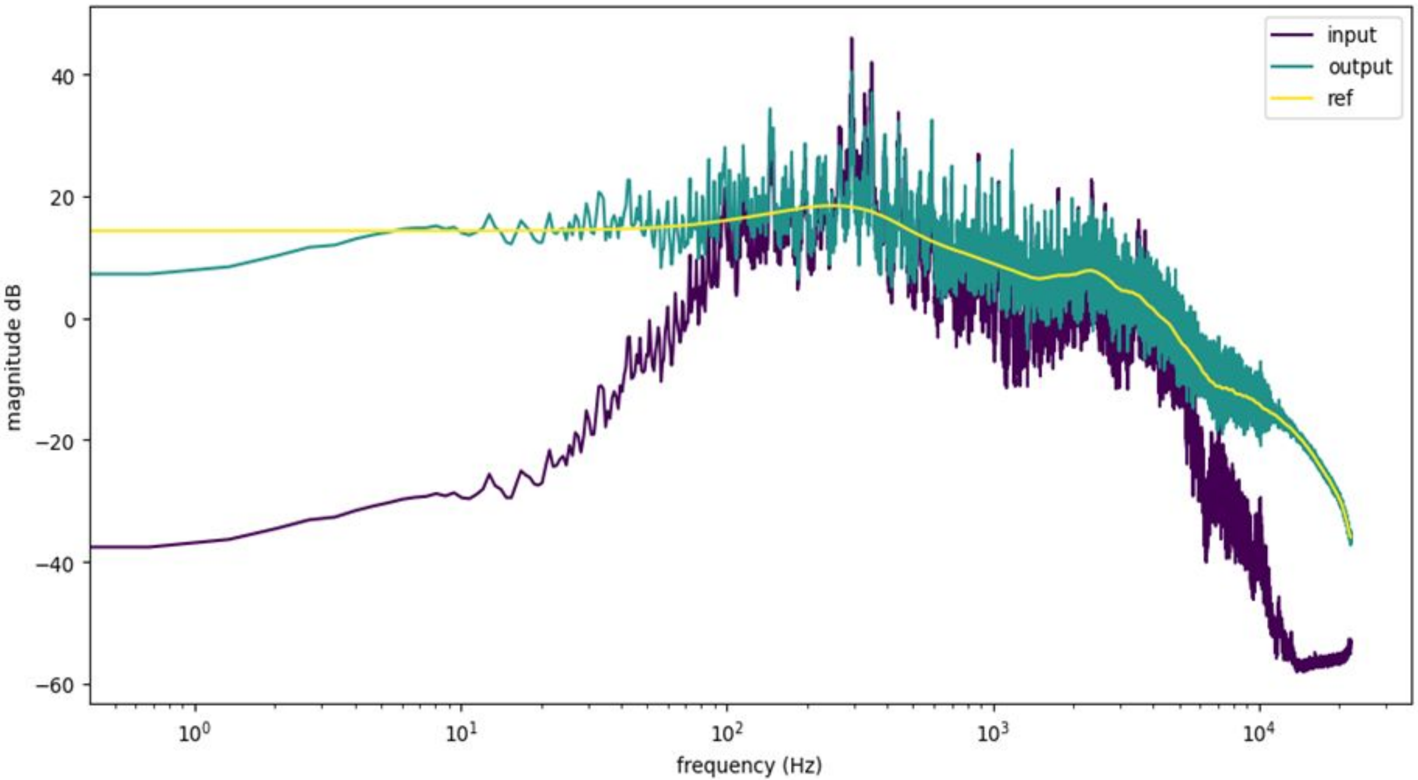


We apply the same to audio effects !

Fx Normalization–EQ average features

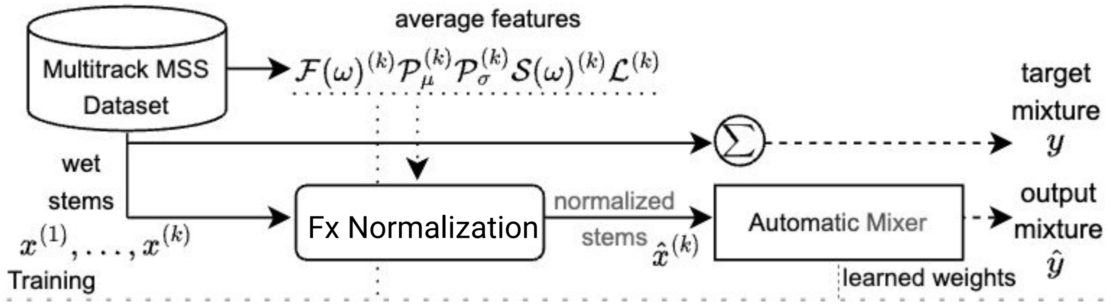


EQ Normalization



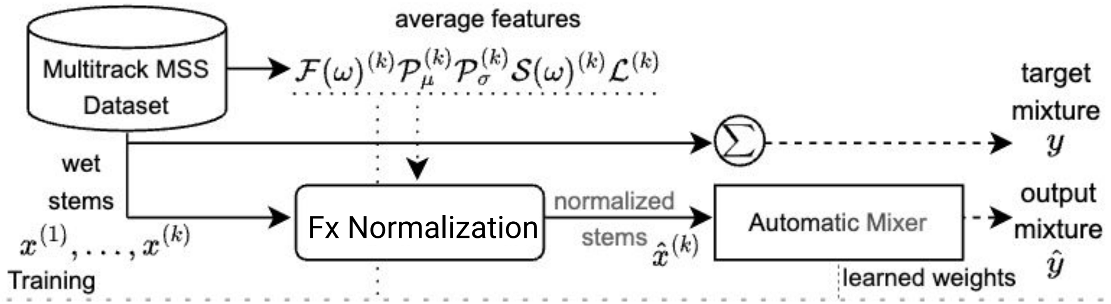
We propose loudness, EQ, panning, compression and reverberation normalization procedures

Method



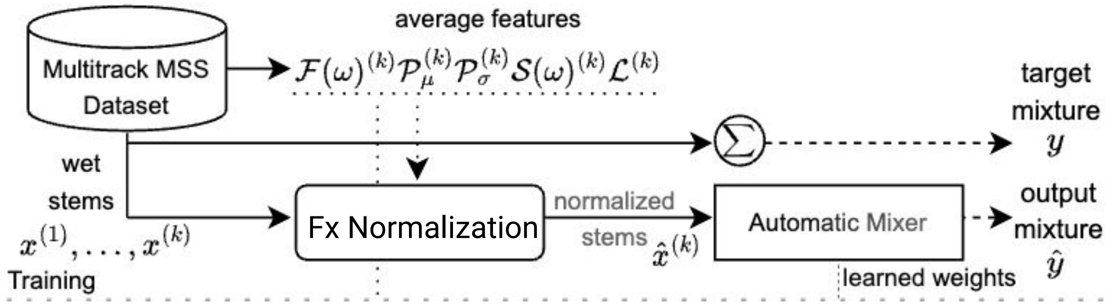
- We use data preprocessing that calculates average features related to audio effects on a music source separation dataset

Method



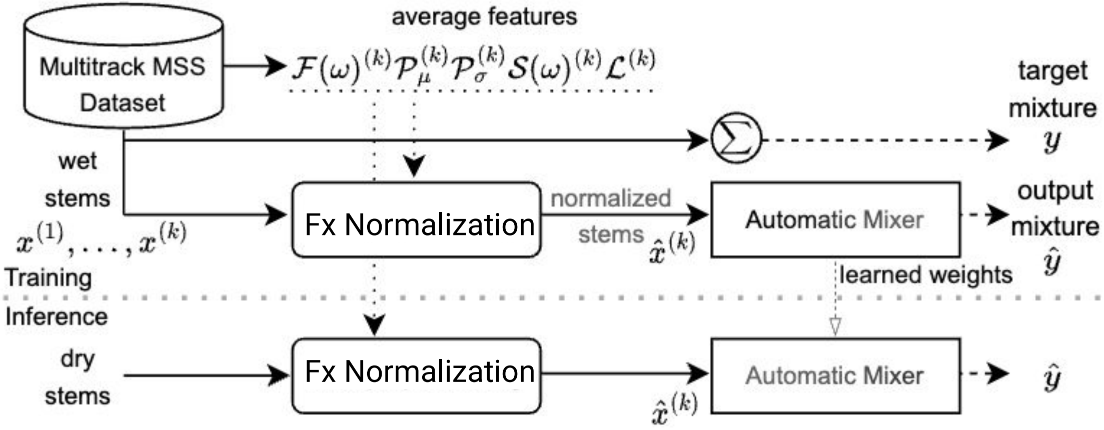
- Based on these features, we “effect-normalize” the wet stems and then train an automatic mixing network

Method



- During training, the model learns how to denormalize the input stems and thus approximate the original mix

Method



- At inference, t**he same preprocessing is applied to dry data**

Evaluation

Listening Test

Perceptual listening tests have become the conventional way to evaluate these systems

There is no standardized test type or platform

We can design tests based on a set of best practices

Adjust them to the specific characteristics of the automatic mixing system

Nicholas Jillings, Brecht De Man, David Moffat and Joshua D. Reiss, "Web Audio Evaluation Tool: A Browser-Based Listening Test Environment," 12th Sound and Music Computing Conference, July 2015.

Listening Test

Criteria

Production Value

- Technical quality of the mix
- Subjective preferences related to the overall technical quality of the mix

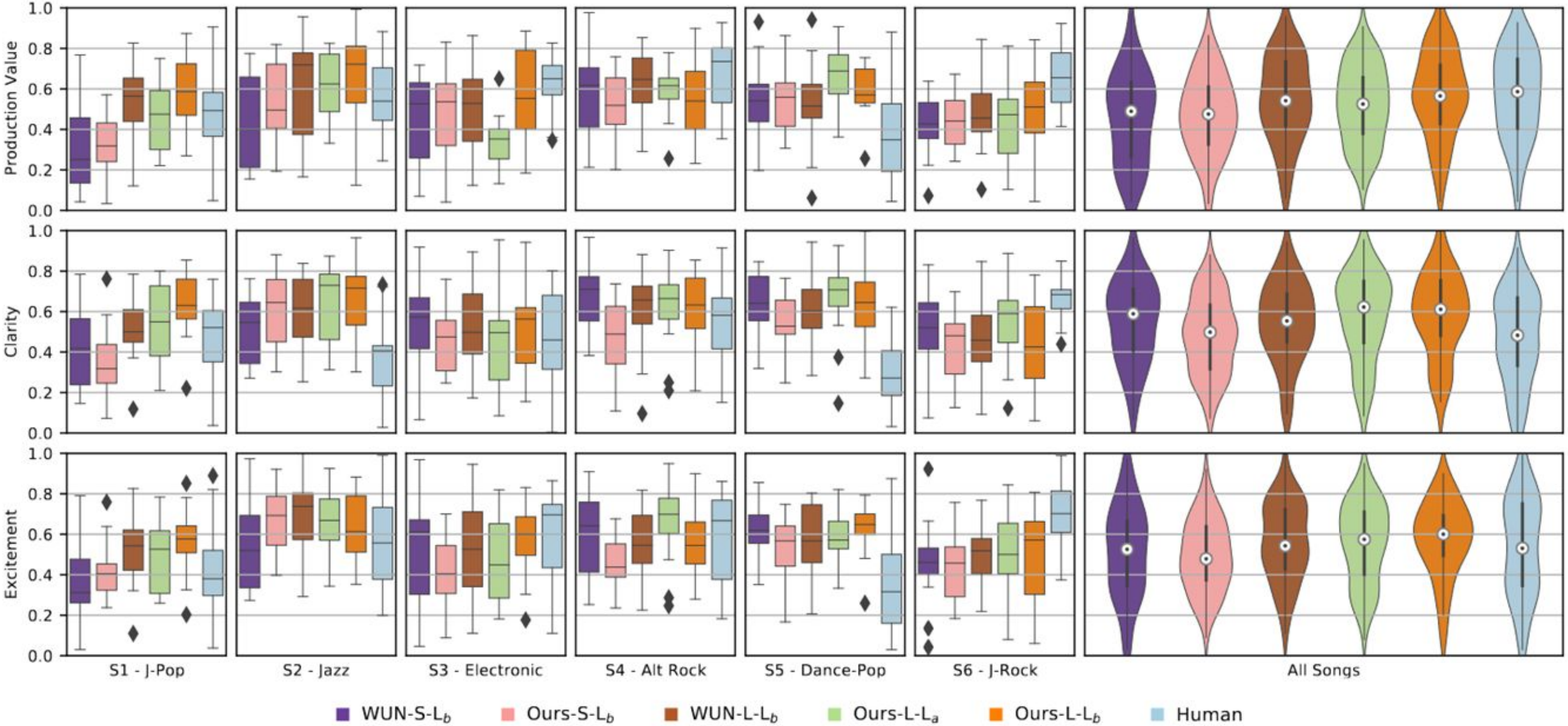
Clarity

- Ability to differentiate musical sources
- This is entirely objective

Excitement

- A non-technical subjective reaction to the mix
- Not related to an evaluation of quality, but to a more personal perception of novelty

Results


Production Value
Clarity
Excitement
1.0
0.8
0.6
0.4
0.2
0.0
S1 - J-Pop
S2 - Jazz
S3 - Electronic
S4 - Alt Rock
S5 - Dance-Pop
S6 - J-Rock
All Songs
WUN-S-L_b
Ours-S-L_b
WUN-L-L_b
Ours-L-L_a
Ours-L-L_b
Human

Conclusion

- We developed a method that performs automatic loudness, EQ, panning, compression and reverberation music mixing

- Fx Normalization works !—Our approach leverages on wet data

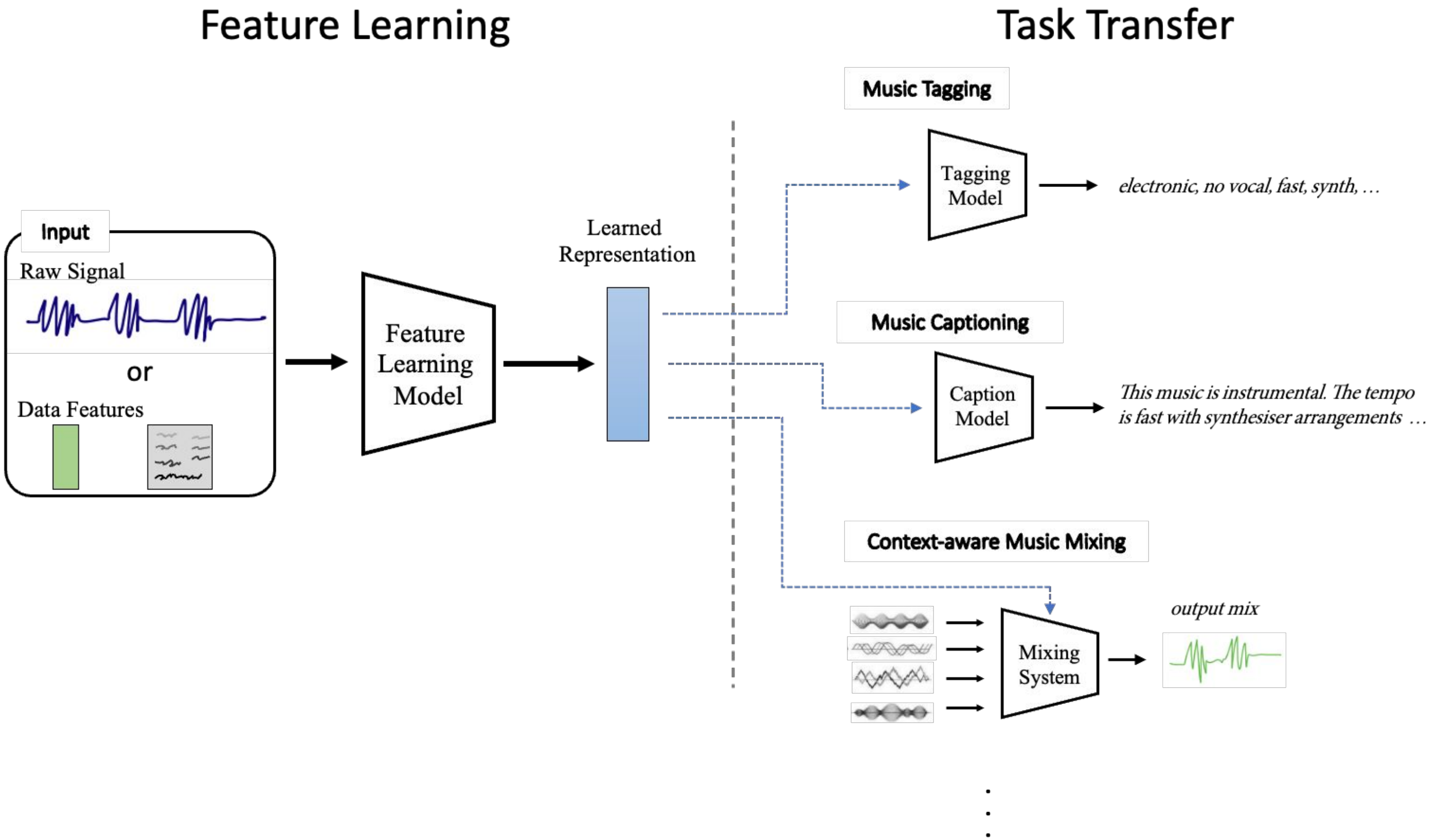
- Resulting mixes compared to professional mixes scored higher in terms of Clarity and are indistinguishable in terms of Production Value and Excitement

Audio Effects Feature Learning

Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects ICASSP 23 Paper

Junghyun (Tony) Koo

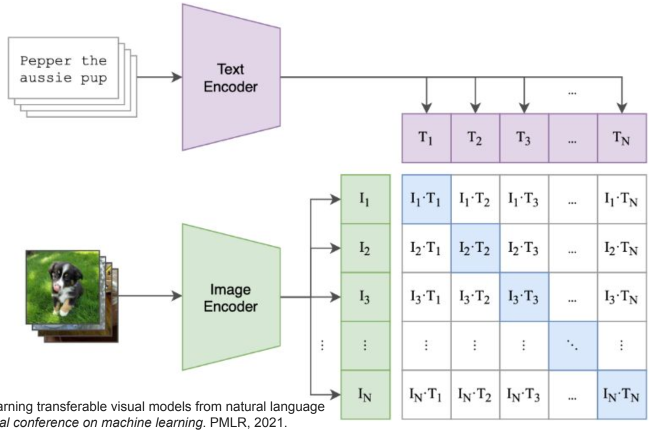
What is Feature Learning?



Contrastive Learning - Recent Applications

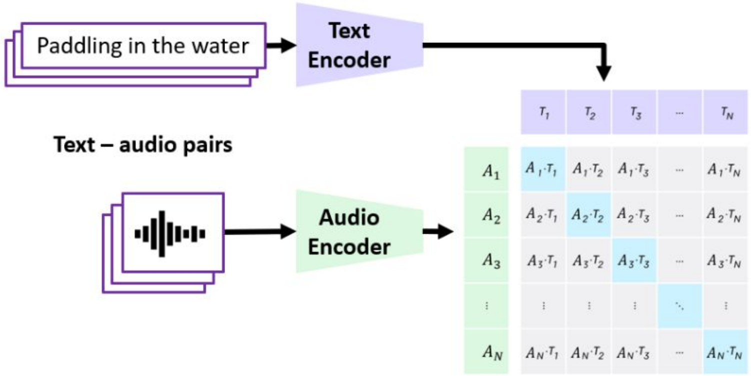
Contrastive Pre-training

Image



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Audio



Elizalde, Benjamin, et al. "Clap learning audio concepts from natural language supervision." *ICASSP 2023*. IEEE, 2023.

Text Prompt Generative Models

Text-to-Image

DALL·E

Stable Diffusion

Text-to-Audio/Music

Google MusicLM

Stable Audio

AudioLDM

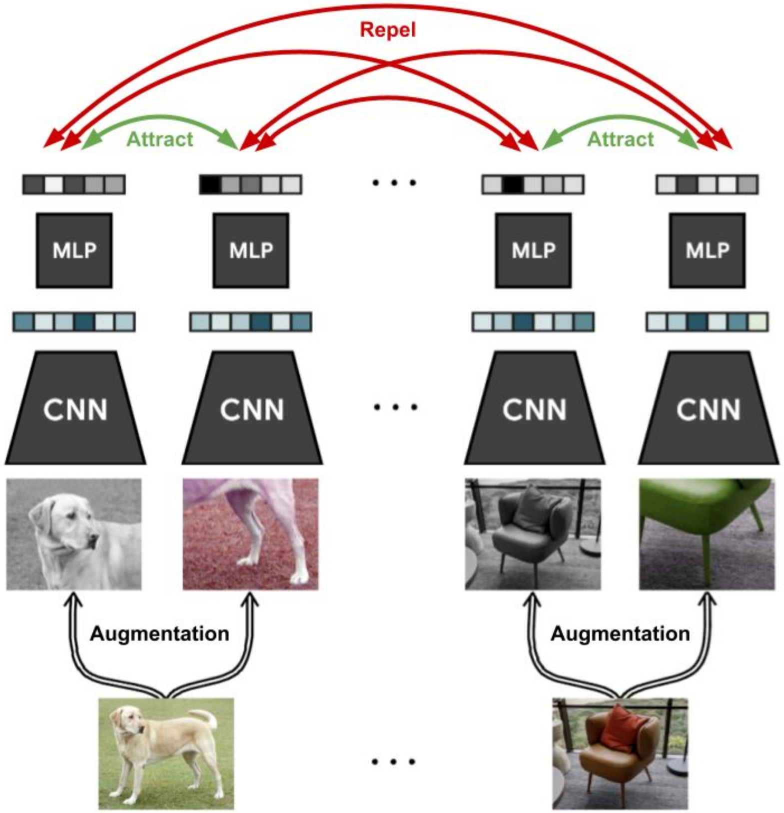
AudioGen

Meta MusicGen AI

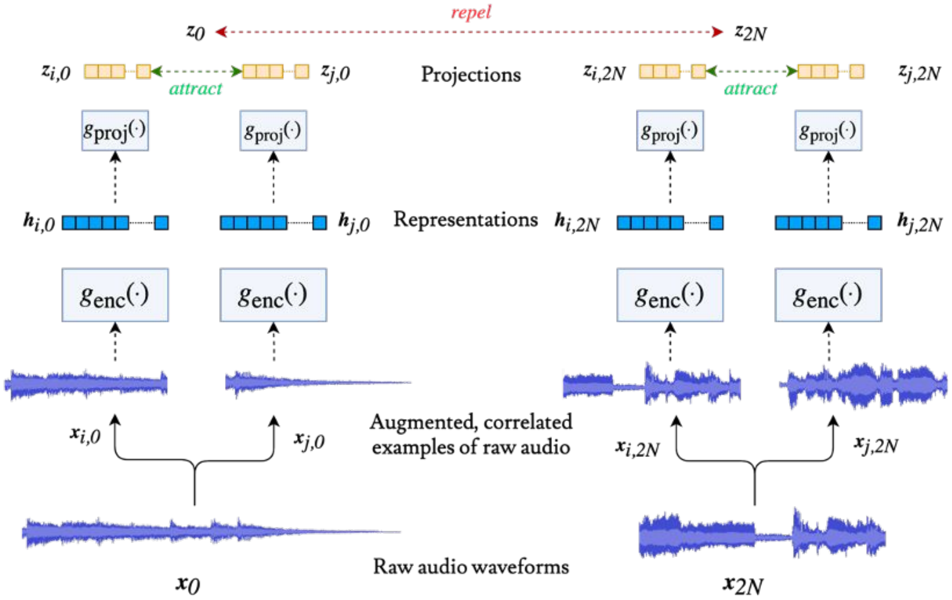
Mubert

Contrastive Learning - Training Method

SimCLR



CLMR



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

Spijkervet, Janne, and John Ashley Burgoyne. "Contrastive learning of musical representations." *ISMIR* 2021.

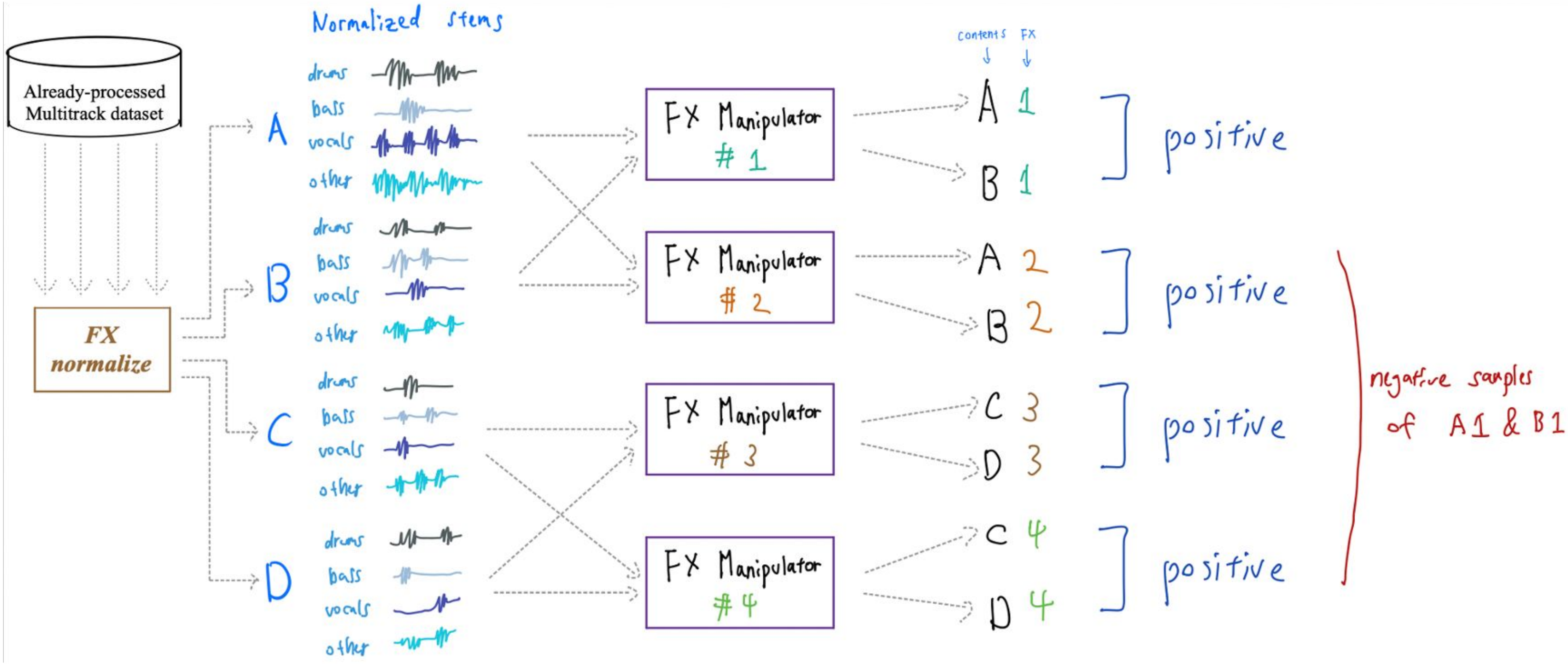
Contrastive Learning on Audio Effects

- Utilizes contrastive learning to understand audio effects.

- Objective: to disentangle mixing styles from musical content.

- Apply learnt representation to downstream task such as mixing style transfer.

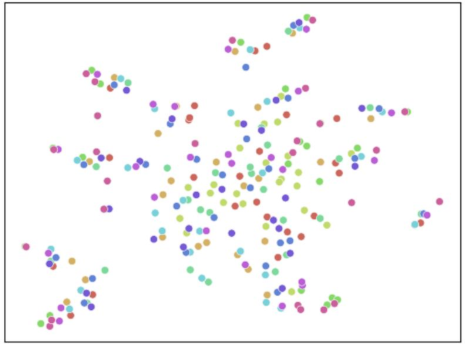
Training Procedure of the FXencoder



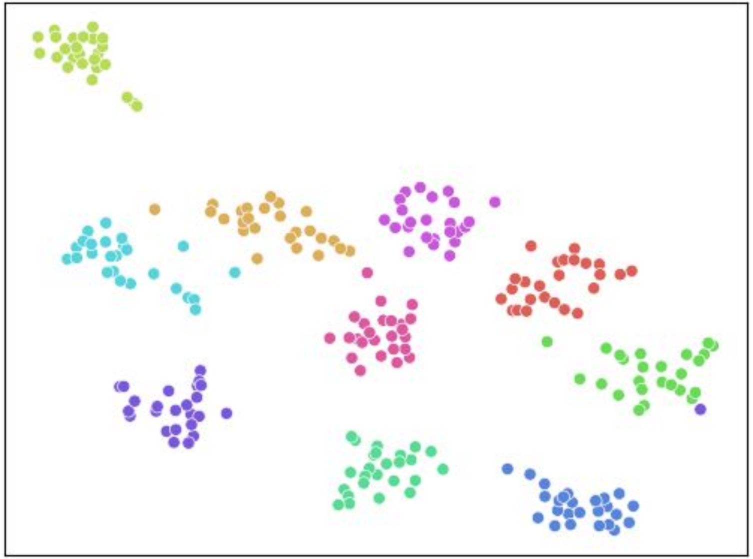
Koo, Junghyun, et al. "Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects." *ICASSP 2023*. IEEE, 2023.

Disentangled Representation

- t-SNE visualization on FXencoder
  - dimensional reduction on feature space
- 10 different random FX manipulation (color) on 25 different songs (point dot)



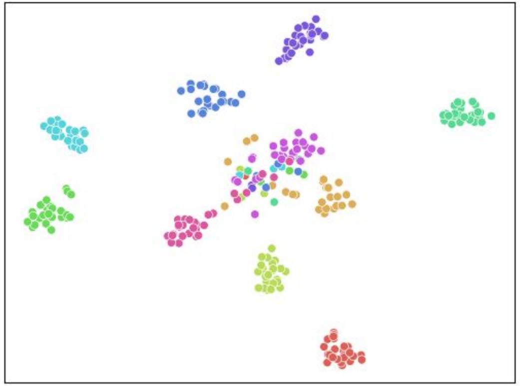
MEE
(model trained with standard approach)



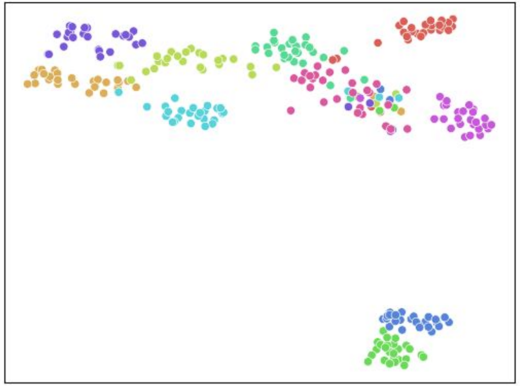
FXencoder

Disentangled Representation - Individual Instrument

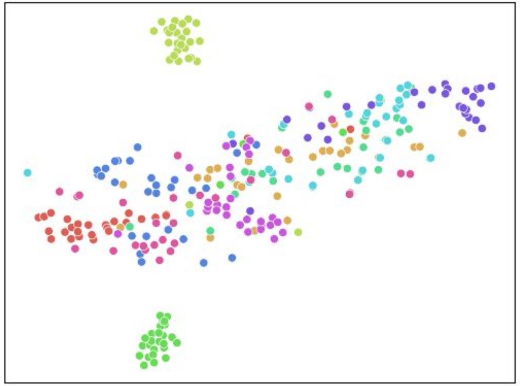
drums



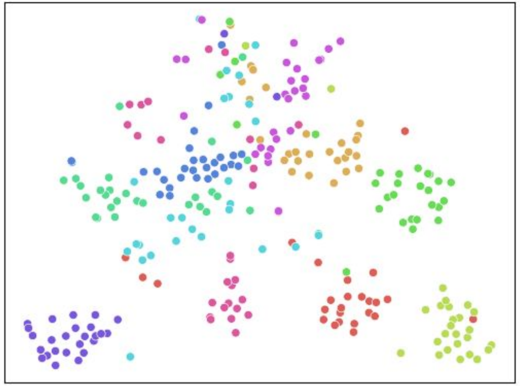
vocals



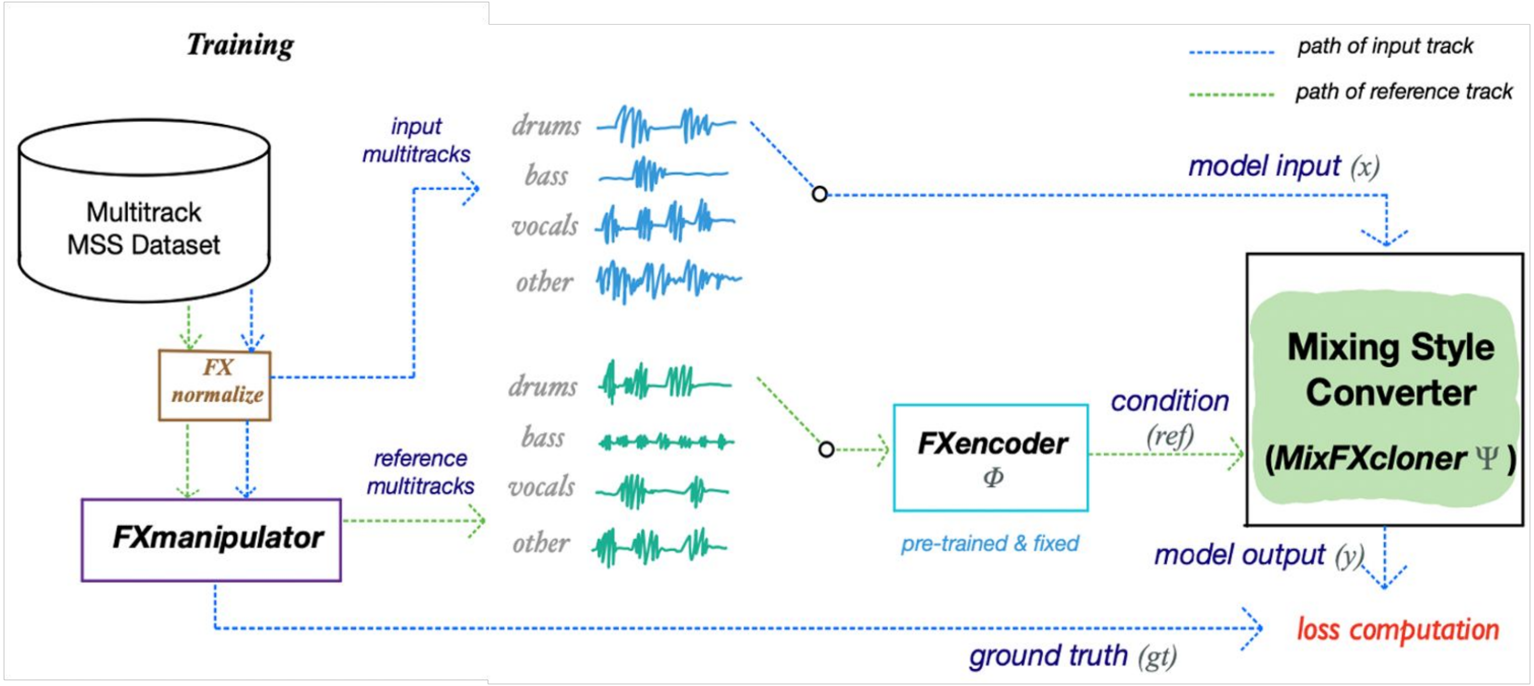
bass



other

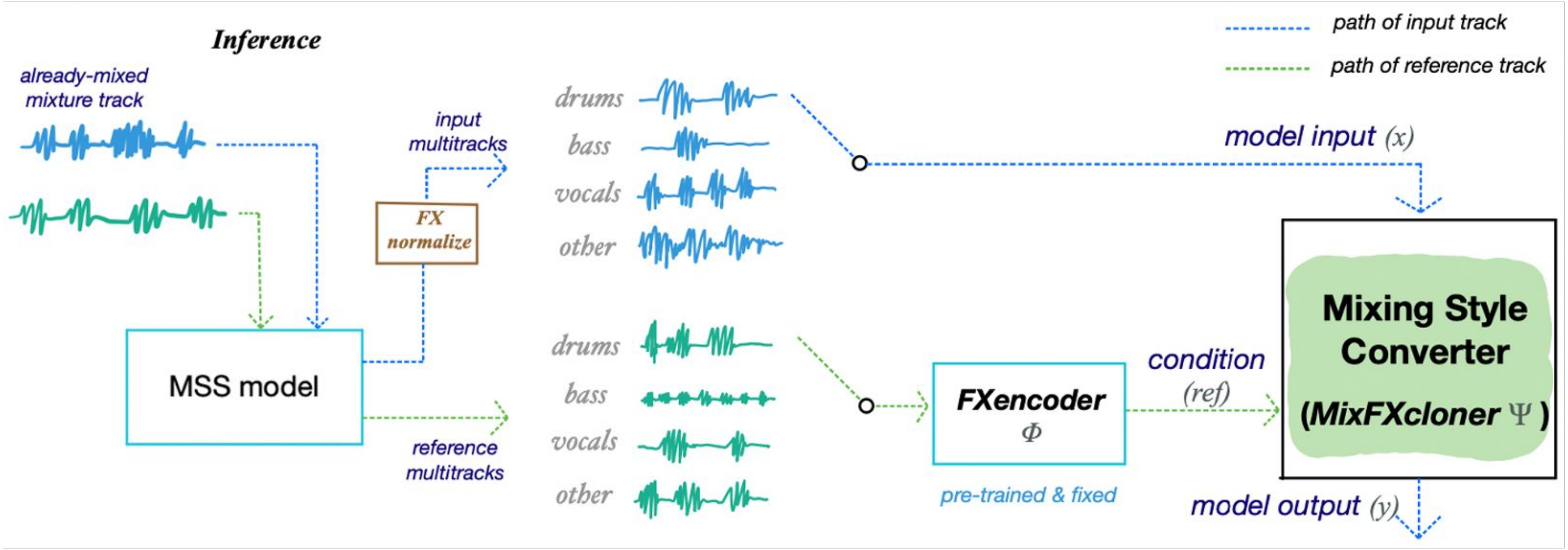


Music Mixing Style Transfer with FXencoder



- Training the mixing style converter is performed by utilizing the representation extracted with already-trained FXencoder

Music Mixing Style Transfer with FXencoder



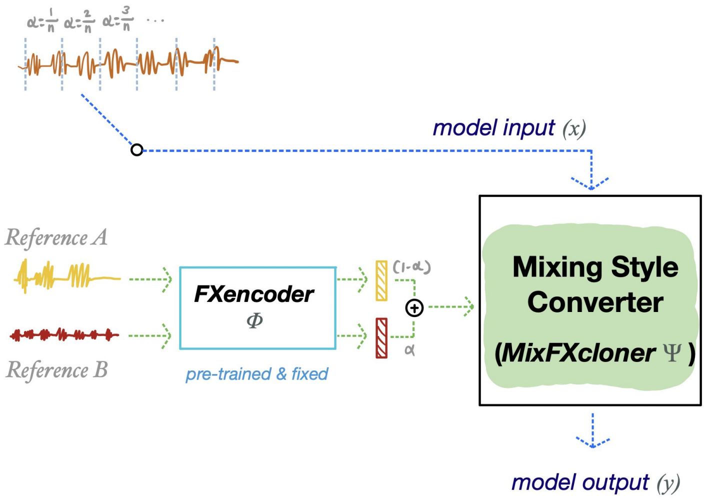
- During inference stage, we can transfer mixing style of mixture-wise inputs using a music source separation (MSS) model

Demo - Mixing Style Transfer

Input Mix:

	Reference A	Reference B
Target Style Mix		
Individual Output		

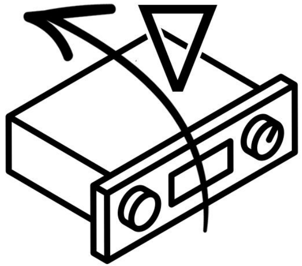
Interpolated Output



Try with your samples!

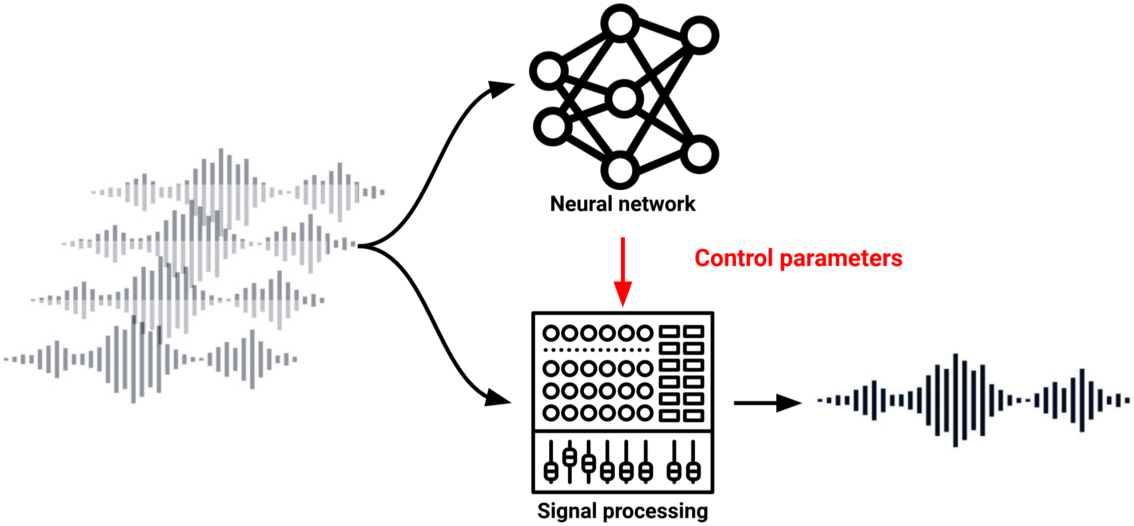
Differentiable signal processing

for automatic mixing



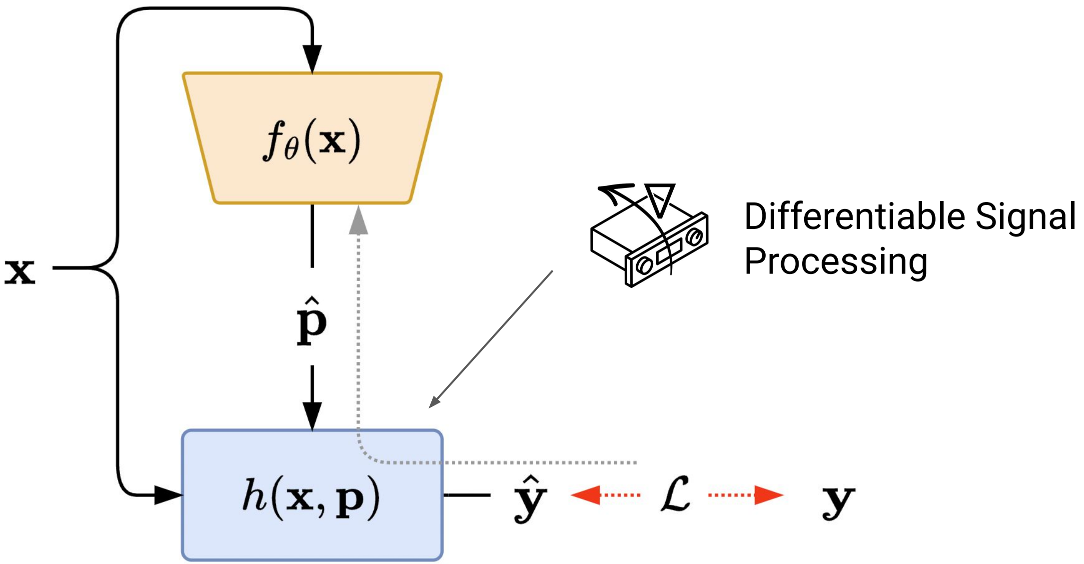
Christian Steinmetz

Neural networks that control DSP



- High-fidelity with minimal risk of introducing artifacts
- Audio processing is visible and controllable by end users
- Significantly more efficient enabling operation on CPU

Neural networks that control DSP



...but this requires haromization of signal processing and **gradient-based learning**

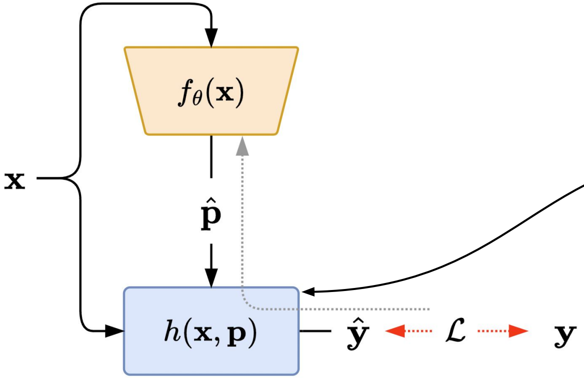
Techniques

1. **Automatic differentiation (AD)**
   Engel et al. 2020

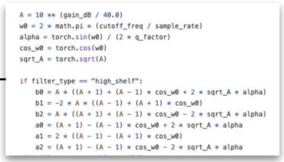
2. **Neural proxies and hybrids (NP)**
   Steinmetz et al. 2020, Steinmetz et al. 2022

3. **Numerical gradient approximation (NGA)**
   Martínez Ramírez et al. 2021

Automatic Differentiation



PyTorch TensorFlow



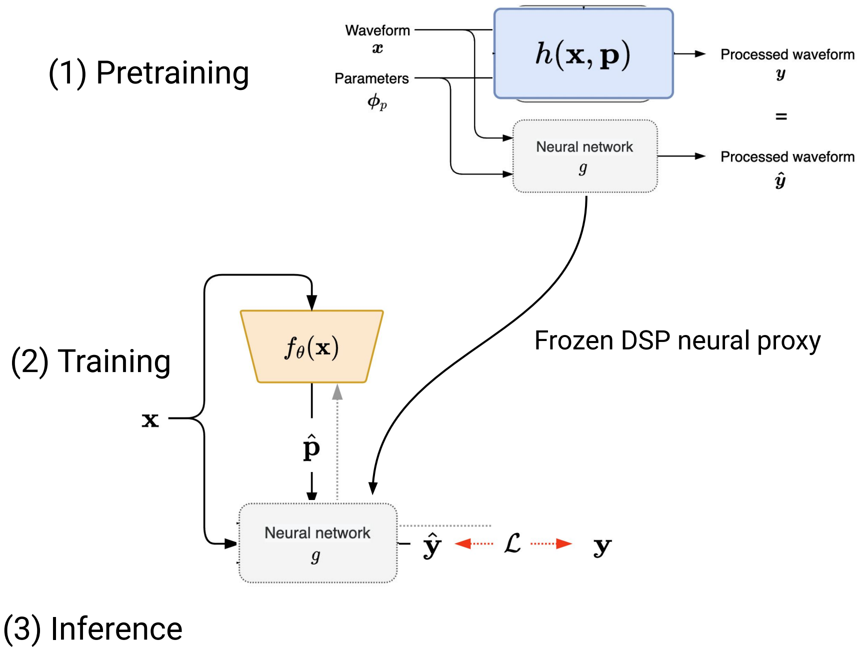
```
A = 10 ** (gain_dB / 40.0)
w0 = 2 * math.pi * (cutoff_freq / sample_rate)
alpha = torch.sin(w0) / (2 * q_factor)
cos_w0 = torch.cos(w0)
sqrt_A = torch.sqrt(A)

if filter_type == "high_shelf":
    b0 = A * ((A + 1) + (A - 1) * cos_w0 + 2 * sqrt_A * alpha)
    b1 = -2 * A * ((A - 1) + (A + 1) * cos_w0)
    b2 = A * ((A + 1) + (A - 1) * cos_w0 - 2 * sqrt_A * alpha)
    a0 = (A + 1) - (A - 1) * cos_w0 + 2 * sqrt_A * alpha
    a1 = 2 * ((A - 1) - (A + 1) * cos_w0)
    a2 = (A + 1) - (A - 1) * cos_w0 - 2 * sqrt_A * alpha
```

Explicitly define signal processing operations in autodiff framework

Engel, Jesse, et al. "DDSP: Differentiable digital signal processing." *ICLR* (2021).

Neural Proxy



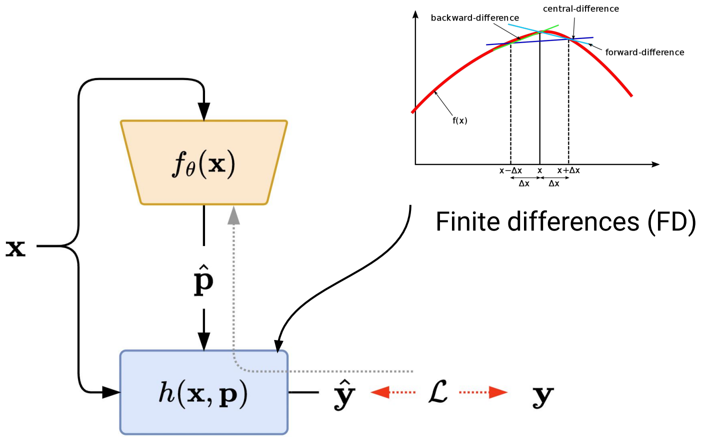
Steinmetz, Christian J., et al. "Automatic multitrack mixing with a differentiable mixing console of neural audio effects." ICASSP, 2021.

Gradient Approximation

$$\frac{\hat{h}(x, p_i)}{p_i} = \frac{h(x, p + \varepsilon\Delta^P) - h(x, p - \varepsilon\Delta^P)}{2\varepsilon\Delta_i^P}, \qquad (2)$$

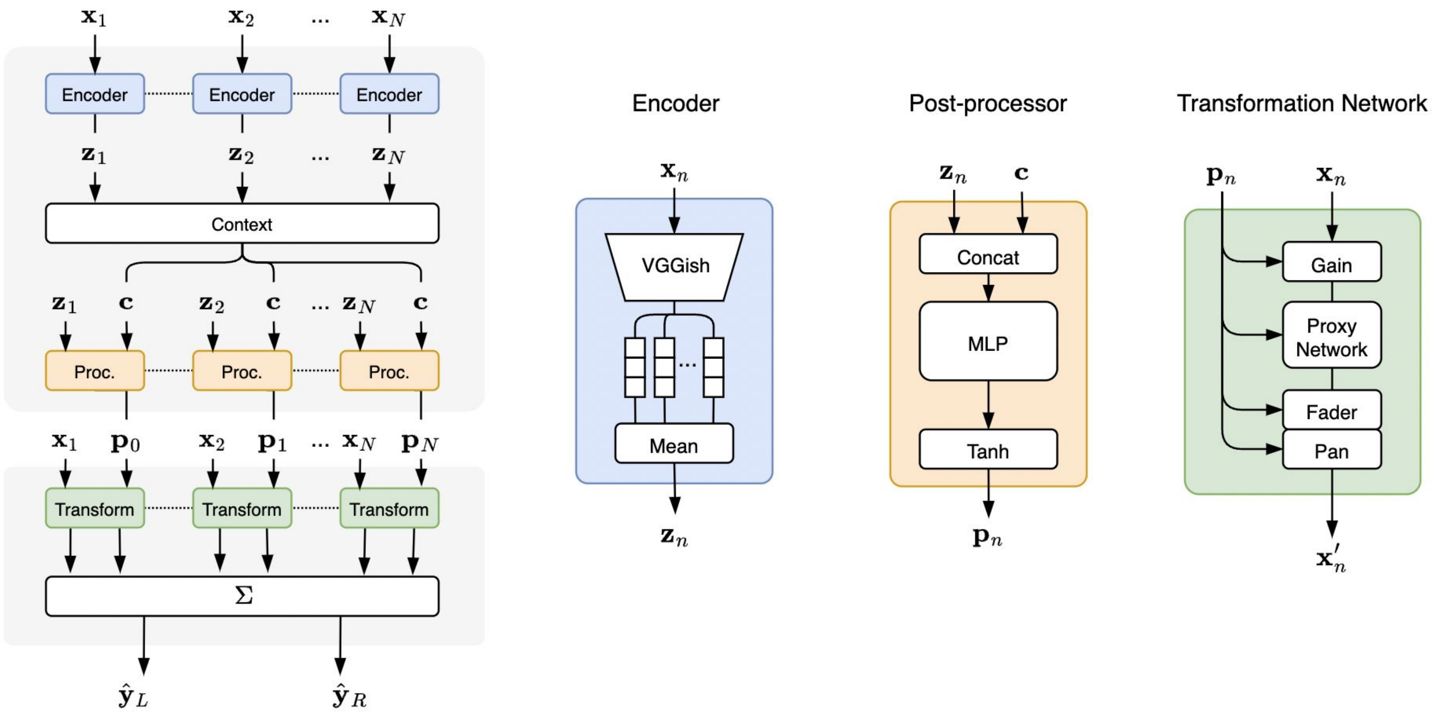
where ε is a small, non-zero value and $\Delta^P \in \mathbb{R}^P$ is a random vector sampled from a symmetric Bernoulli distribution ($\Delta_i^P = \pm 1$) [46].

Simultaneous perturbation stochastic approximation (SPSA)



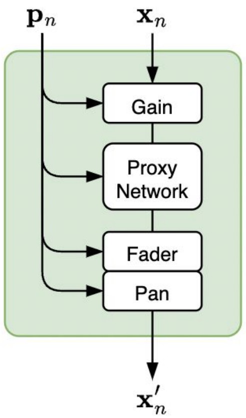
Martínez Ramírez, Marco A., et al. "Differentiable signal processing with black-box audio effects." ICASSP, 2021.

Creating a differentiable mixing console

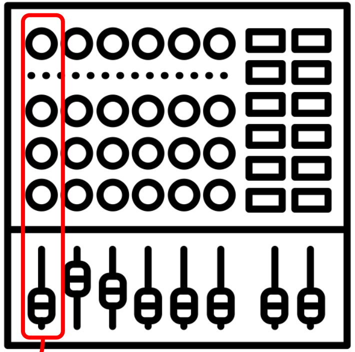


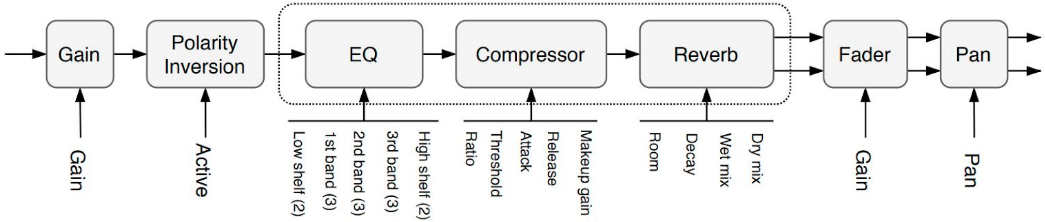
Steinmetz, Christian J., et al. "Automatic multitrack mixing with a differentiable mixing console of neural audio effects." ICASSP, 2021.

Creating a differentiable mixing console



Proxy network



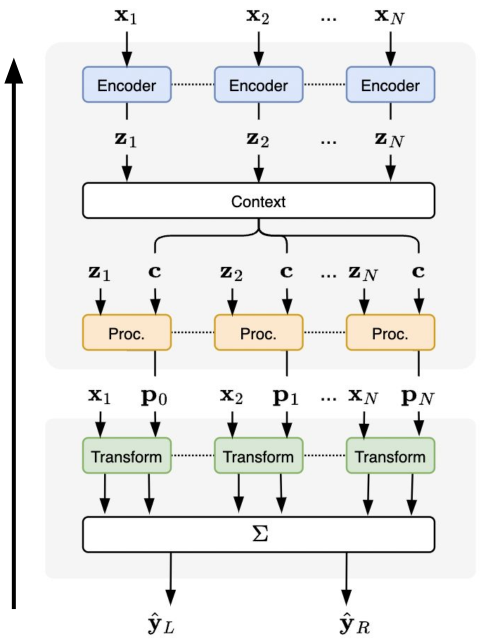


Differentiable channel strip

Steinmetz, Christian J., et al. "Automatic multitrack mixing with a differentiable mixing console of neural audio effects." ICASSP, 2021.

Creating a differentiable mixing console

Backpropagation
(Training)



Input tracks

Latent features

Parameter estimators

Differentiable audio effects

Predicted mix

Steinmetz, Christian J., et al. "Automatic multitrack mixing with a
differentiable mixing console of neural audio effects." ICASSP, 2021.

Coming soon

DASP

Differentiable audio signal processors

in PyTorch

Reverberation

Compressor / Expander

Parametric Equalizer

Distortion

Stereo Widener

Stereo Panner

with more coming...

Coming soon

DASP

Differentiable audio signal processors

in PyTorch

Pure functional interface for each audio processor

Differentiable implementations enable backprop

Can target CPU or GPU with support for batching

Permissive open source license (Apache 2.0)

Questions

Commercialising Audio Research

Angeliki Mourgela

Meet RoEx

William Trevis
Full-stack Engineer
Previously at Boeing and is an ex-founder
3 years of experience

Dr David Ronan
CEO/CTO
Former Head of Research at AI Music (Acquired by Apple)
14 years of experience

Dr Angeliki Mourgela
Research Engineer
Professional sound engineer by trade
13 years of experience

Research to product - Key Challenges

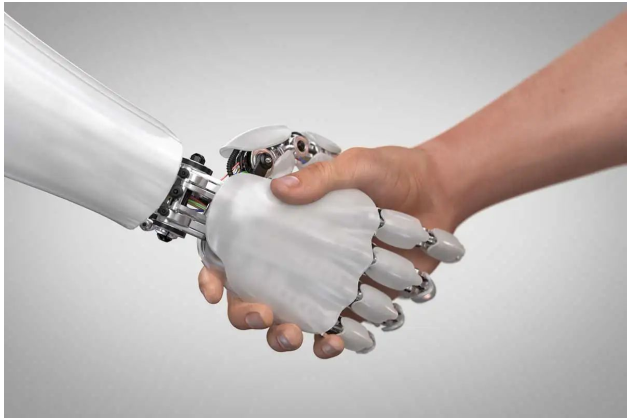
- What is a good mix? **Definition** and **target**
- **Complexity** and **variety** of genres
- **Balance** between user control and automation
- **Quality of input** audio is most likely not ideal

Current Market

- 14.6 million music creators online
- Most creators lack audio engineering skills
- User target group - amateurs, pro-amateurs

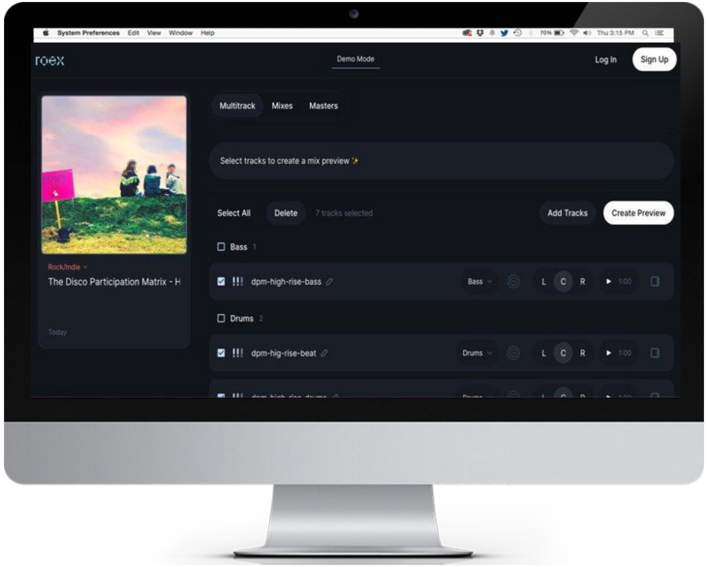
Our technology

- Combination of **machine learning** and **traditional audio engineering** methods
- **Genre-specific** mixing and mastering
- User has **choice of** how much **control** they want to have both before and after the processing



User workflow - tackling the challenges

- Combination of machine learning models for **corrective processing** of the input audio to ensure quality
- Research-driven **subgroup mixing** approach (artificial limit of 8 tracks)
- Choice of **priority, pan and reverb** settings prior to mixing
- **Mix preview** and gain adjustments

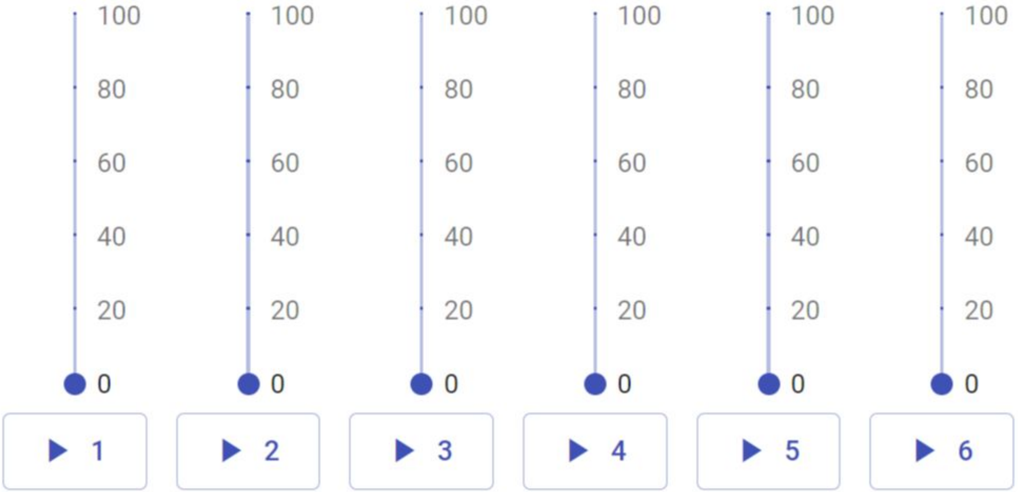


Roex Automix Demo

Demos

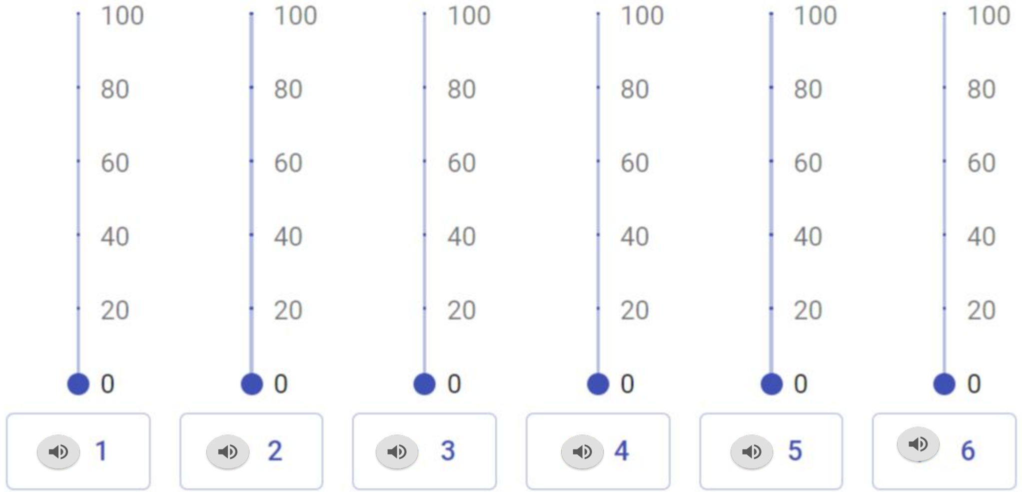
Mixes

Please rate each mix based on your overall preference



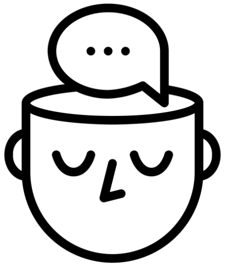
Mixes

Please rate each mix based on your overall preference



Mixes

1. (Koo et al., 2022a) - Music Mixing Style Transfer with reference from MUSDB18
2. Mono mix
3. Gary Bromham - Professional audio engineer mix
4. (Steinmetz et al., 2021) - DMC mix trained with MedleyDB - Gain and Panning
5. (Martinez-Ramirez et al., 2022) - Fx Normalization
6. RoEx



Future Directions

Generative AI

Functional art

Text prompt

Outpainting

Style transfer

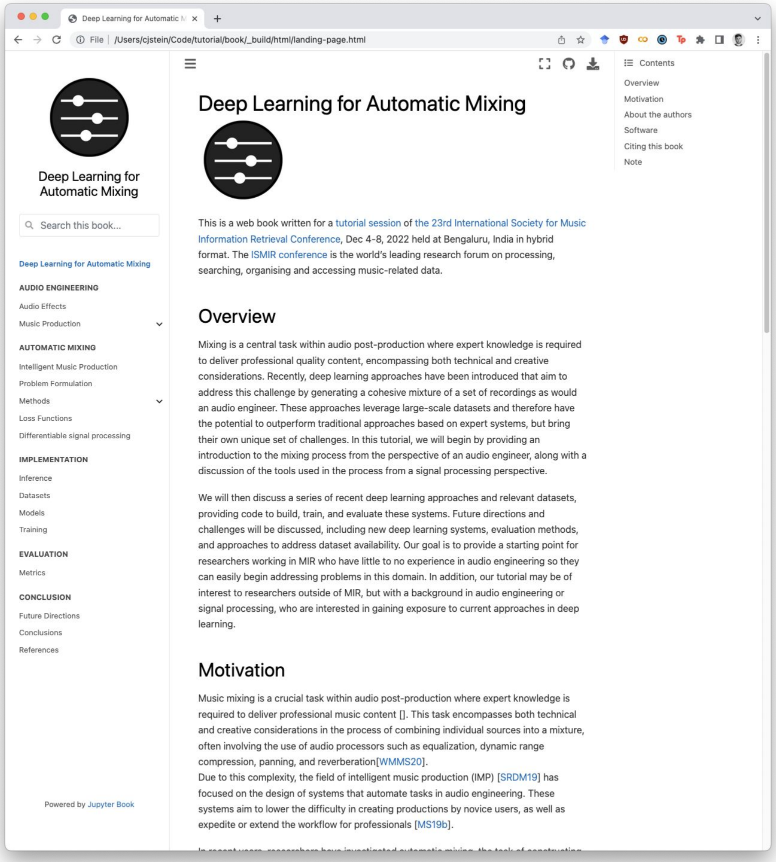
+

=

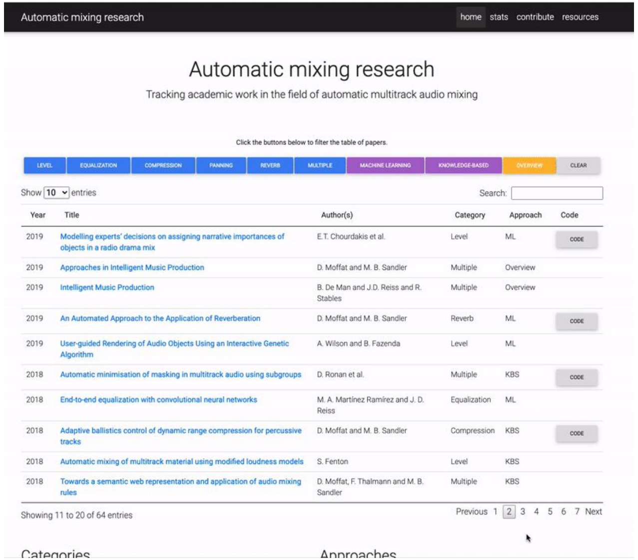
Resources

Book

https://dl4am.github.io/tutorial



More works on automatic mixing research


Automatic mixing research home stats contribute resources

Automatic mixing research

Tracking academic work in the field of automatic multitrack audio mixing

Click the buttons below to filter the table of papers.

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

Show 10 entries Search:

Year	Title	Author(s)	Category	Approach	Code
2019	Modelling experts' decisions on assigning narrative importances of objects in a radio drama mix	E.T. Chourdakis et al.	Level	ML	CODE
2019	Approaches in Intelligent Music Production	D. Moffat and M. B. Sandler	Multiple	Overview	
2019	Intelligent Music Production	B. De Man and J.D. Reiss and R. Stables	Multiple	Overview	
2019	An Automated Approach to the Application of Reverberation	D. Moffat and M. B. Sandler	Reverb	ML	CODE
2019	User-guided Rendering of Audio Objects Using an Interactive Genetic Algorithm	A. Wilson and B. Fazenda	Level	ML	
2018	Automatic minimisation of masking in multitrack audio using subgroups	D. Ronan et al.	Multiple	KBS	CODE
2018	End-to-end equalization with convolutional neural networks	M. A. Martínez Ramírez and J. D. Reiss	Equalization	ML	
2018	Adaptive ballistics control of dynamic range compression for percussive tracks	D. Moffat and M. B. Sandler	Compression	KBS	CODE
2018	Automatic mixing of multitrack material using modified loudness models	S. Fenton	Level	KBS	
2018	Towards a semantic web representation and application of audio mixing rules	D. Moffat, F. Thalmann and M. B. Sandler	Multiple	KBS	

Showing 11 to 20 of 64 entries Previous 1 2 3 4 5 6 7 Next

Categories Approaches

Searchable/filterable table of relevant papers and stats

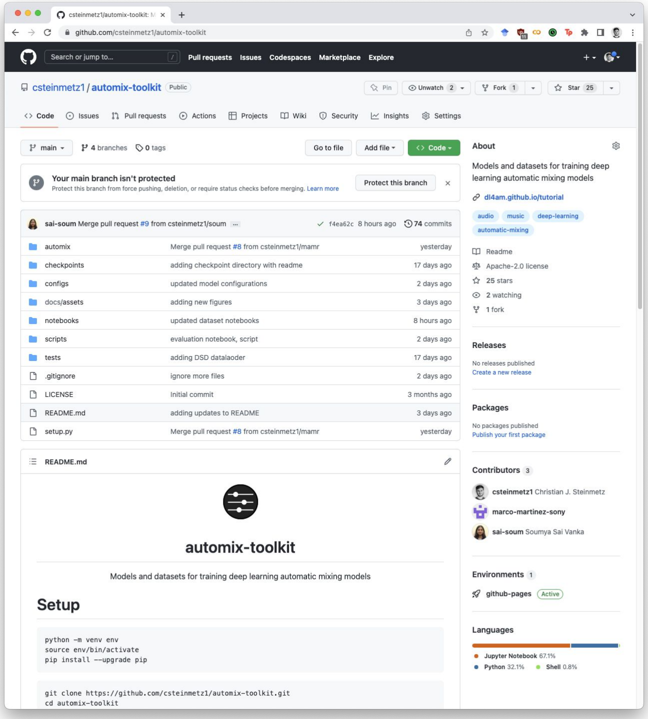
https://csteinmetz1.github.io/AutomaticMixingPapers

automix-toolkit

https://github.com/csteinmetz1/automix-toolkit

Star

Star it on GitHub



Thank You

Question?