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

Gary Bromham

Junghyun (Tony) Koo

Christian J. Steinmetz

Marco A. Martínez-Ramírez

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

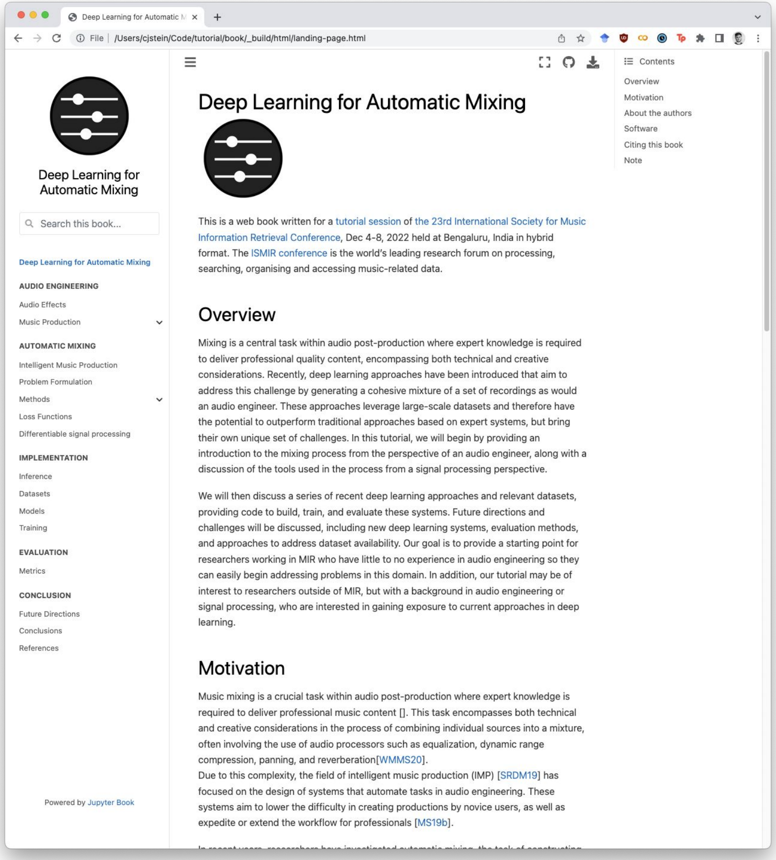
“Hey!” “Hi!”

AI for

- Multitrack
- Music
- Mixing

Book

https://dl4am.github.io/tutorial



Deep Learning for Automatic Mixing

This is a web book written for a tutorial session of the 23rd International Society for Music Information Retrieval Conference, Dec 4-8, 2022 held at Bengaluru, India in hybrid format. The ISMIR conference is the world's leading research forum on processing, searching, organising and accessing music-related data.

Overview

Mixing is a central task within audio post-production where expert knowledge is required to deliver professional quality content, encompassing both technical and creative considerations. Recently, deep learning approaches have been introduced that aim to address this challenge by generating a cohesive mixture of a set of recordings as would an audio engineer. These approaches leverage large-scale datasets and therefore have the potential to outperform traditional approaches based on expert systems, but bring their own unique set of challenges. In this tutorial, we will begin by providing an introduction to the mixing process from the perspective of an audio engineer, along with a discussion of the tools used in the process from a signal processing perspective.

We will then discuss a series of recent deep learning approaches and relevant datasets, providing code to build, train, and evaluate these systems. Future directions and challenges will be discussed, including new deep learning systems, evaluation methods, and approaches to address dataset availability. Our goal is to provide a starting point for researchers working in MIR who have little to no experience in audio engineering so they can easily begin addressing problems in this domain. In addition, our tutorial may be of interest to researchers outside of MIR, but with a background in audio engineering or signal processing, who are interested in gaining exposure to current approaches in deep learning.

Motivation

Music mixing is a crucial task within audio post-production where expert knowledge is required to deliver professional music content []. This task encompasses both technical and creative considerations in the process of combining individual sources into a mixture, often involving the use of audio processors such as equalization, dynamic range compression, panning, and reverberation[WMMS20].
Due to this complexity, the field of intelligent music production (IMP) [SRDM19] has focused on the design of systems that automate tasks in audio engineering. These systems aim to lower the difficulty in creating productions by novice users, as well as expedite or extend the workflow for professionals [MS19b].

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

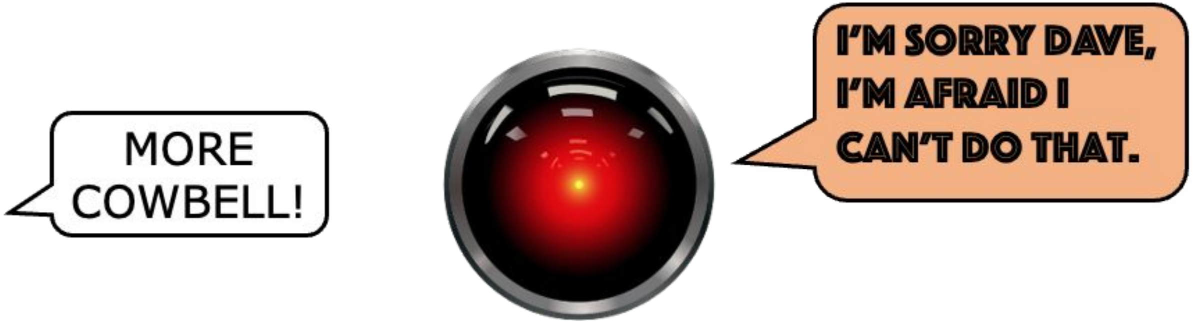
Outline

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

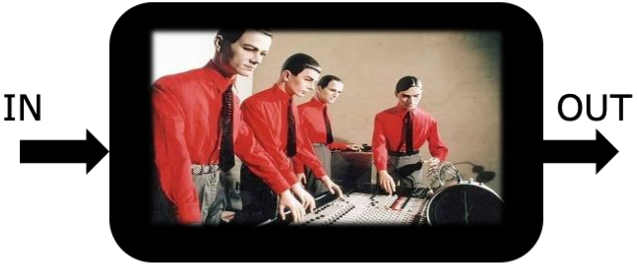
Focus on creative aspects

Increased demand

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

AI comes in many forms

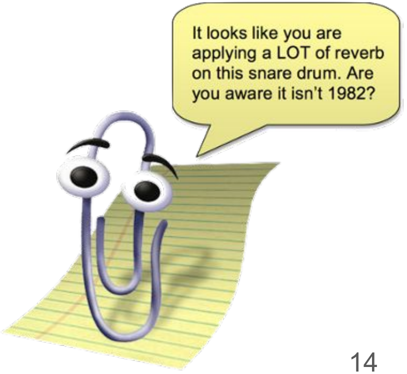
“The Black Box”



“The Assistant”

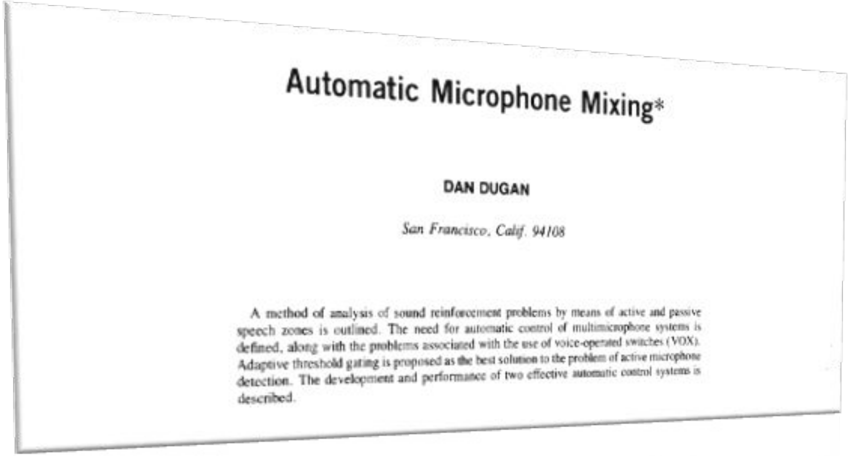
“The Smart Interface”

“The Diagnostician”


It looks like you are applying a LOT of reverb on this snare drum. Are you aware it isn't 1982?

History

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



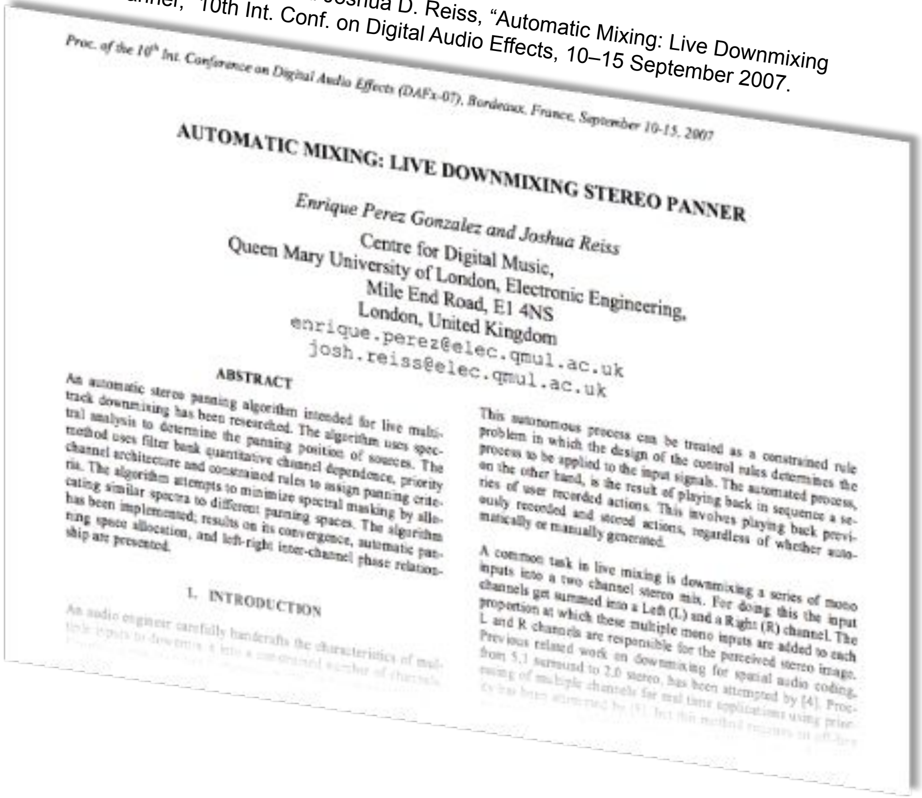
Automatic Microphone Mixing*

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 multitrack 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 multiple inputs to downmix it into a constrained number of channels.

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]. Processing of multiple channels for real time applications using prior

History 2007-2012

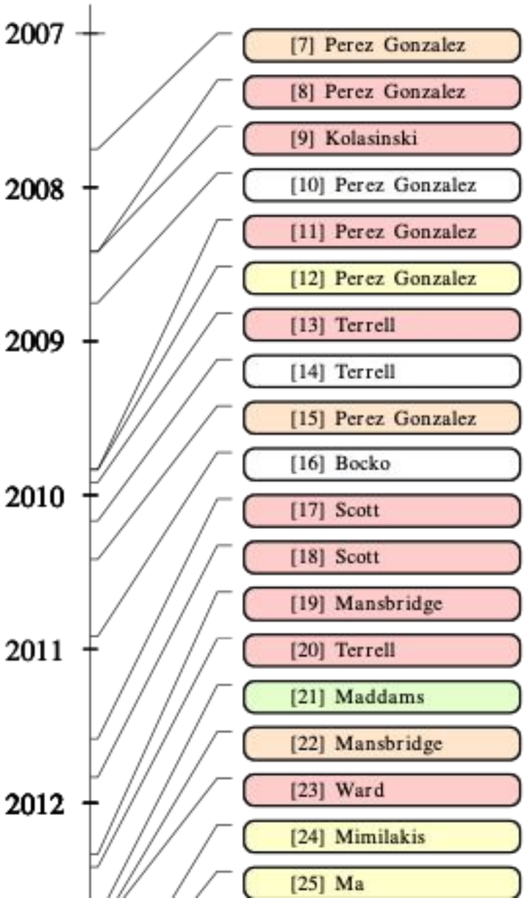
Legend

🟥 Level

🟧 Panning

🟨 EQ

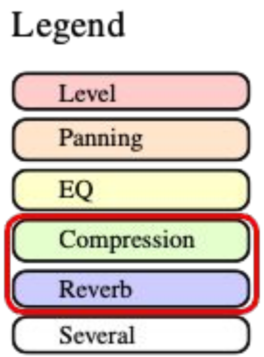
⬜ Several

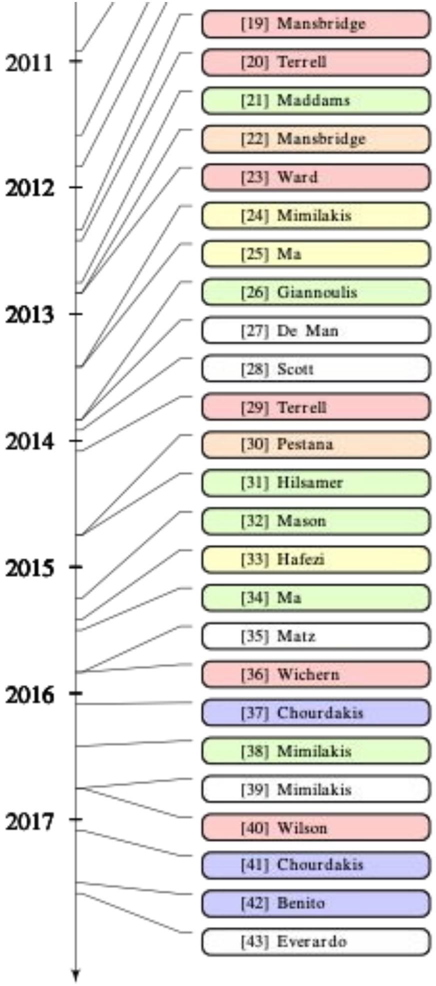


(14)

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



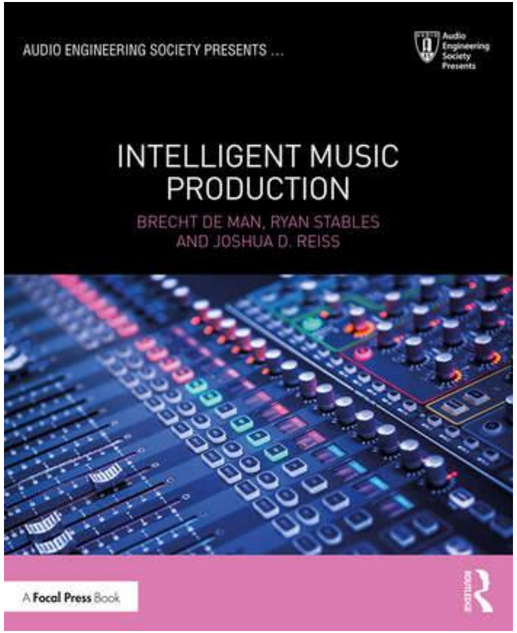


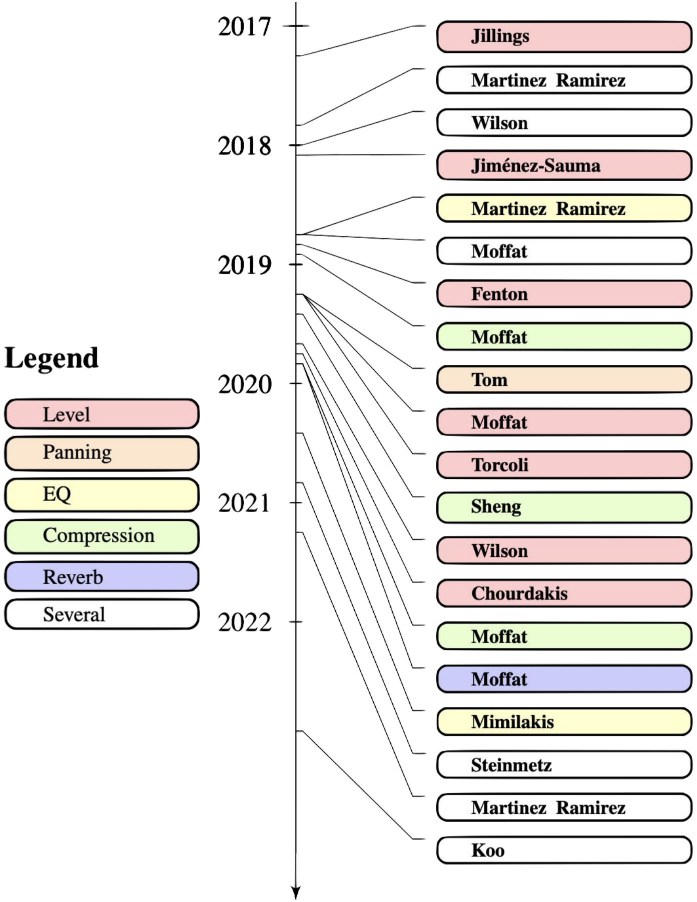
(23)

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

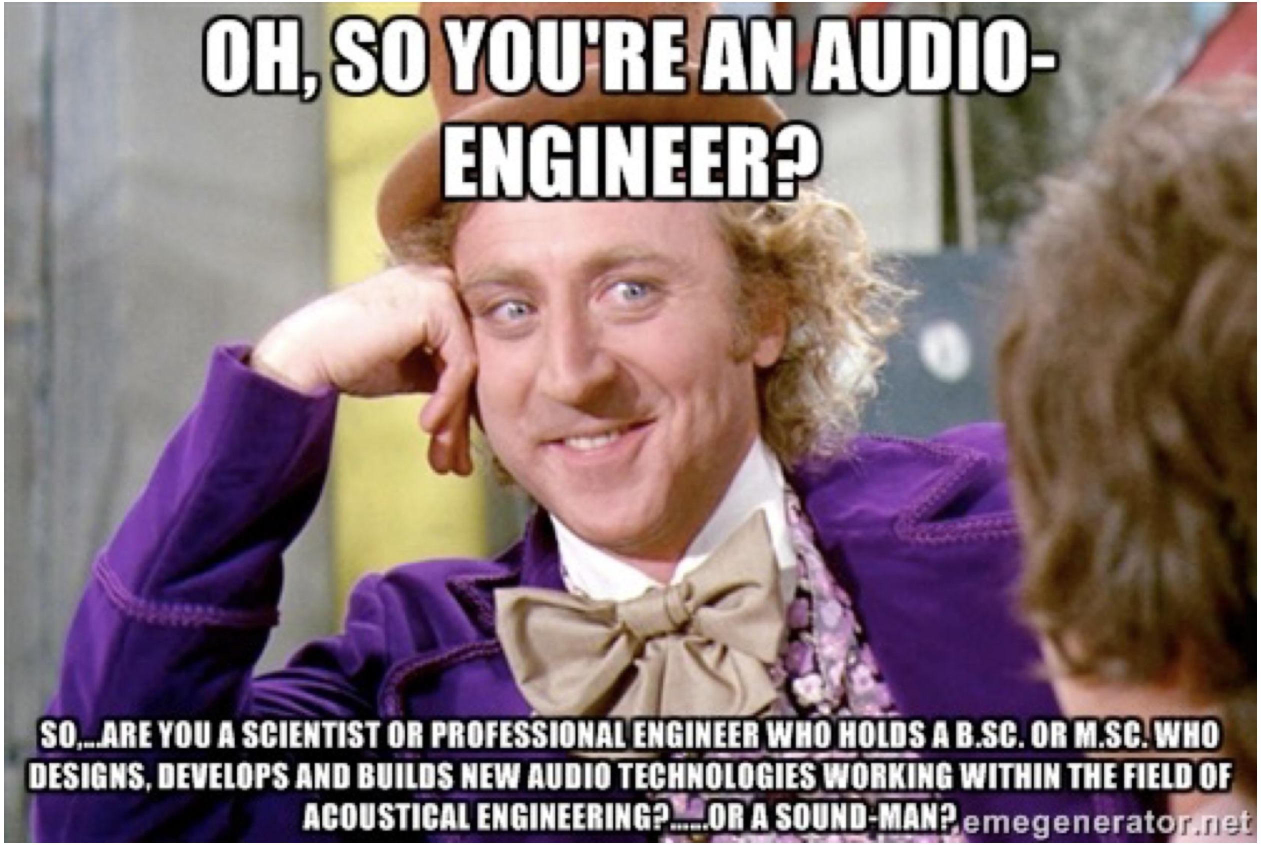
https://csteinmetz1.github.io/AutomaticMixingPapers/


AUDIO ENGINEERING SOCIETY PRESENTS ...
Audio Engineering Society Presents
INTELLIGENT MUSIC PRODUCTION
BRECHT DE MAN, RYAN STABLES AND JOSHUA D. REISS
A Focal Press Book



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?

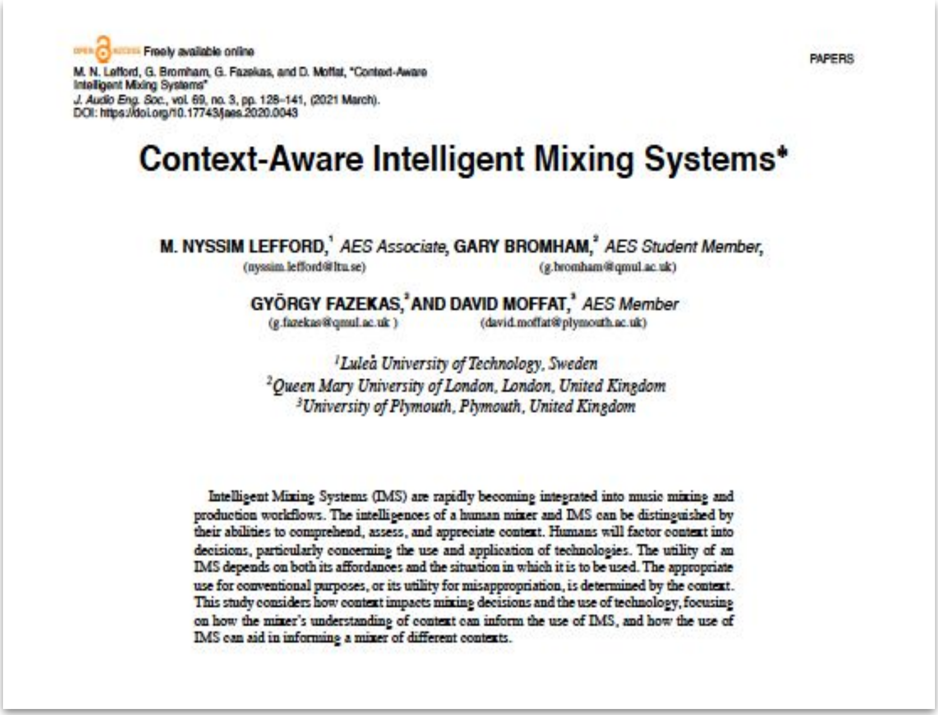
Technical

… 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)



OPEN ACCESS 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)

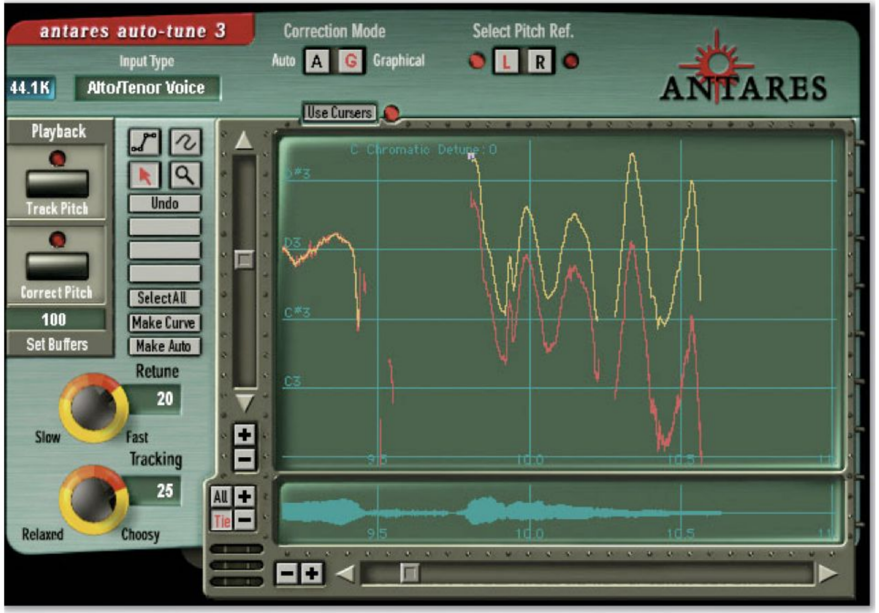
- Three distinct groups in the music production chain. Sandler, M. et al. 2019.
- All three groups have different motivations as mix engineers and producers.
- Intelligent music productions tools are often designed for those with less experience.
- Pro-Am’s who are looking to attain professional-sounding 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’



44.1K

Input Type: Alto/Tenor Voice

C Chromatic Detune: 0

100

Retune: 20

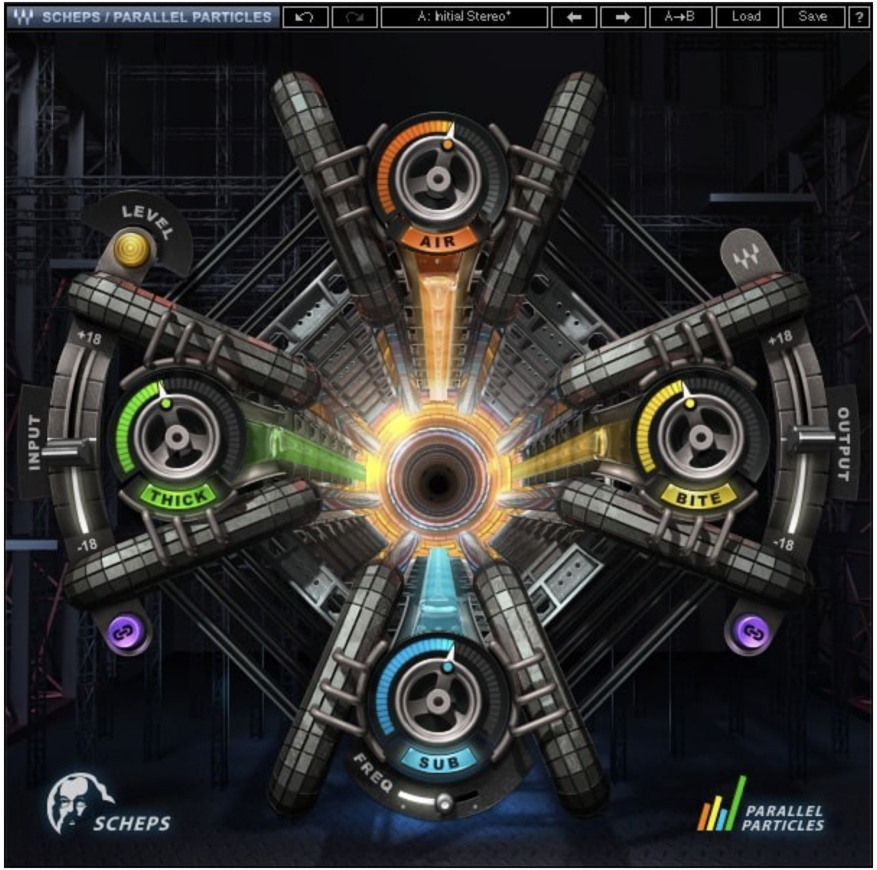
Tracking: 25

Antares Autotune

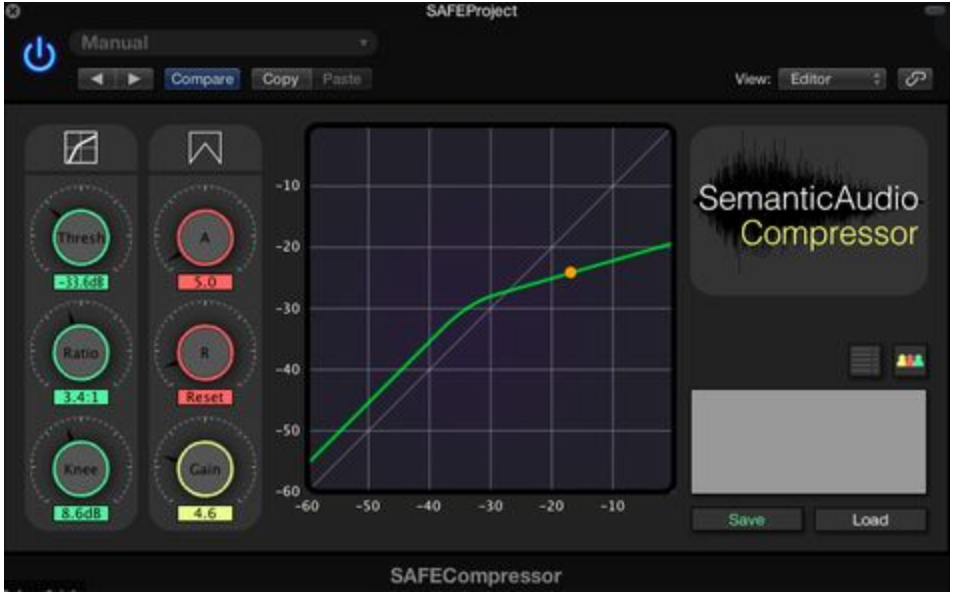
The Language of Mixing - Semantics

- ‘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



Manual

View: Editor

Thresh -33.6dB

Ratio 3.4:1

Knee 8.6dB

A 5.0

R Reset

Gain 4.6

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

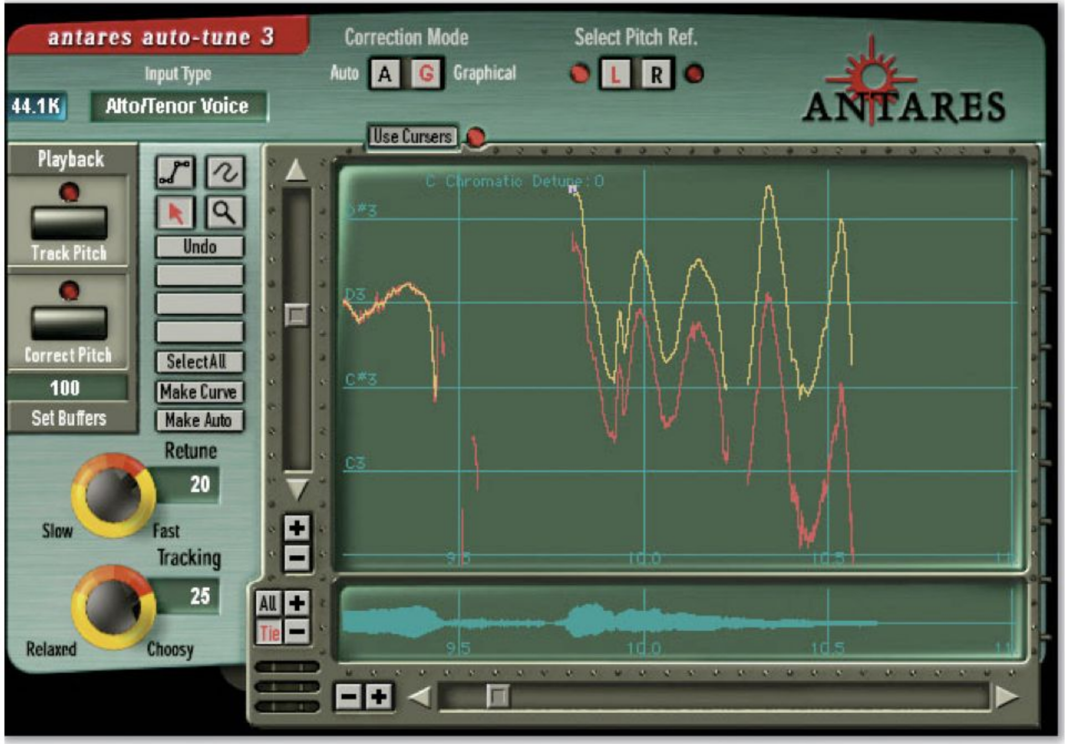
- Largely ignores context.
- Creativity often in the outliers in data. ‘Creep’ by Radiohead.
- Mixing is essentially an emotional response or reaction to a piece of music.

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



antares auto-tune 3

Input Type

Alto/Tenor Voice

44.1K

Correction Mode

Auto A G Graphical

Select Pitch Ref.

L R

C Chromatic Detune: 0

Retune

20

Tracking

25

100

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**

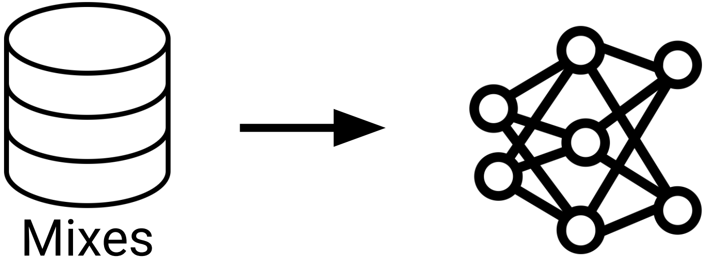
- -Then, the process of learning “the right amount of reverb” is carried out by the network by learning to **filter out the additional reverberation**

Part 2

System Components

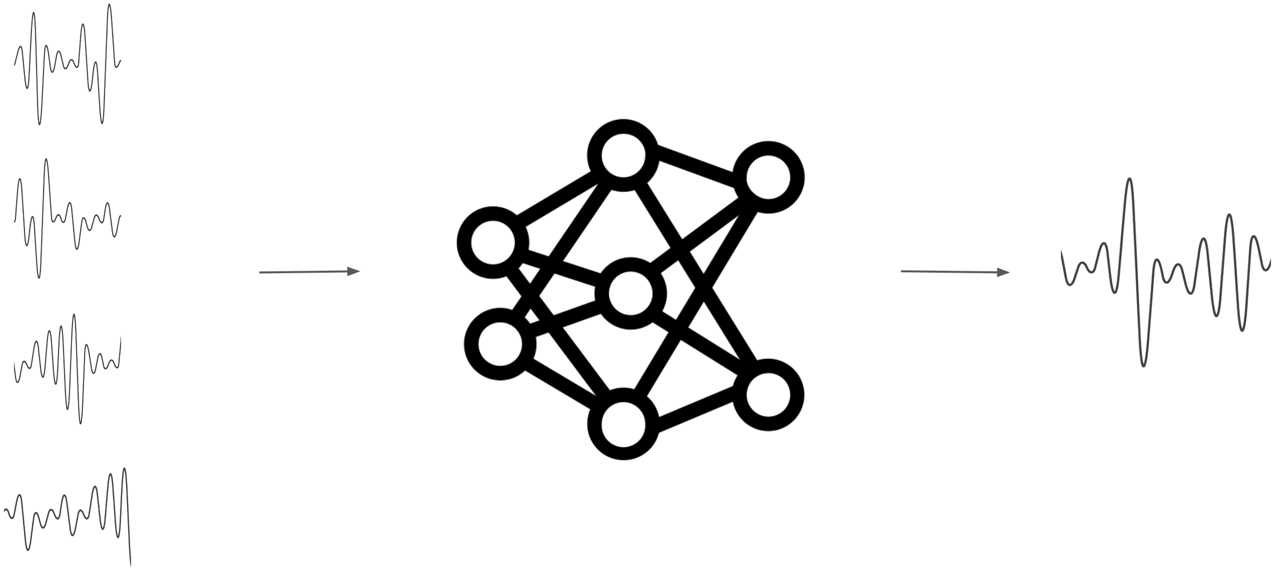
Soumya Sai Vanka

Deep Learning



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

What we want? (at Inference)



Multitrack
(Input)

Neural
Network

Mix
(Output)

Considerations

Interpretability

Input Taxonomy

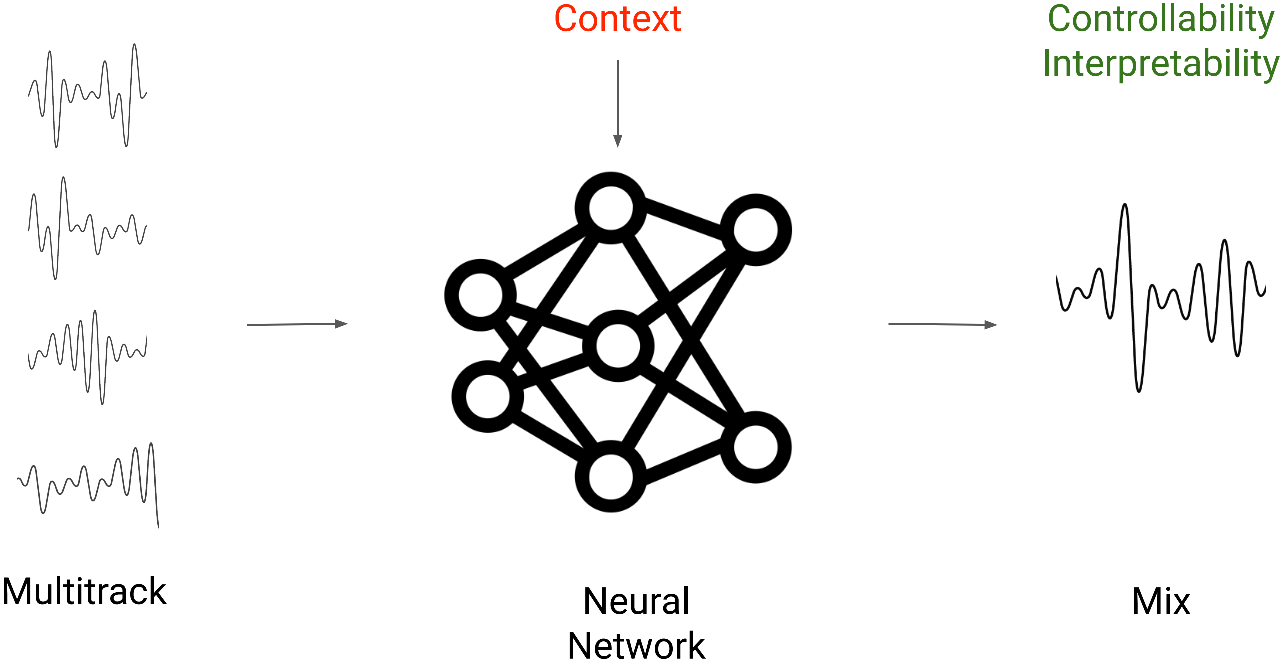
Controllability

Fidelity

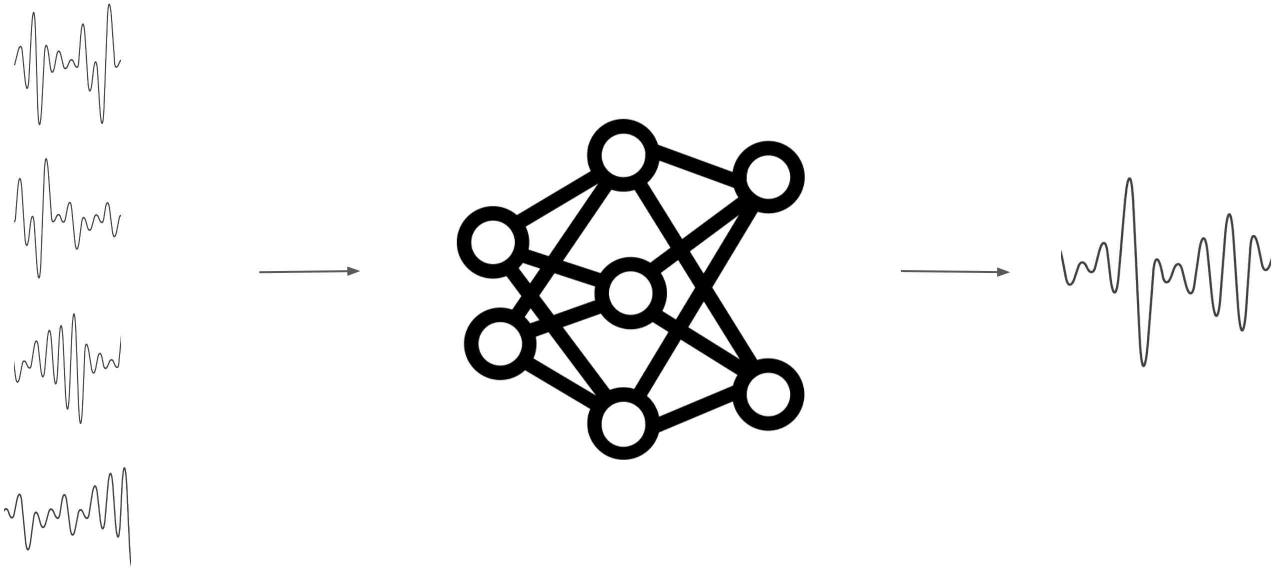
Context

Interaction

What we want?


Context
Controllability
Interpretability
Multitrack
Neural
Network
Mix

Let’s begin with simple case

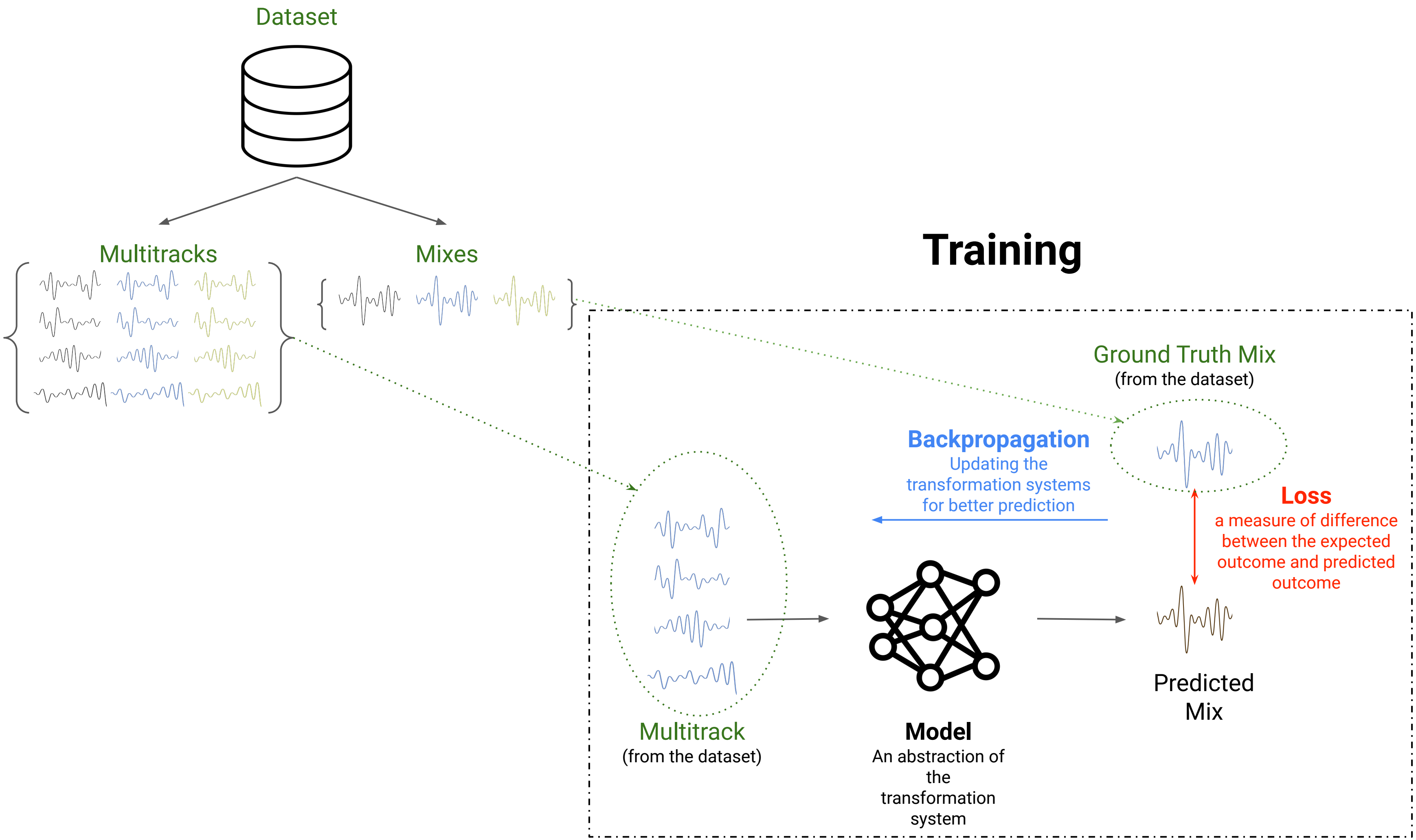


Multitrack
(Input)

Neural
Network

Mix
(Output)

Training



Dataset

Multitracks

Mixes

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

Multitrack
(from the dataset)

Model
An abstraction of the transformation system

Predicted Mix

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)

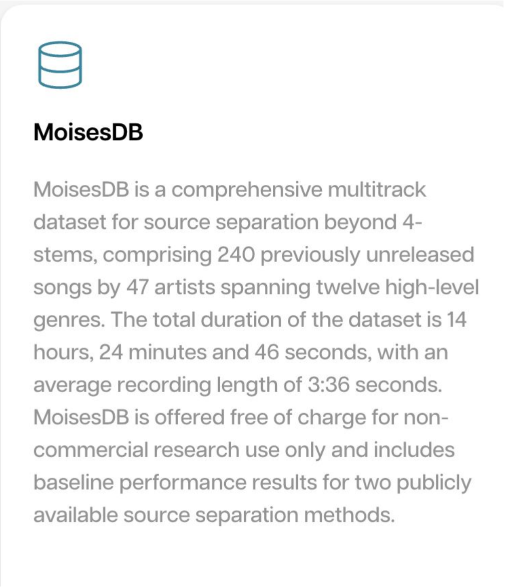
MuseDB

- 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.

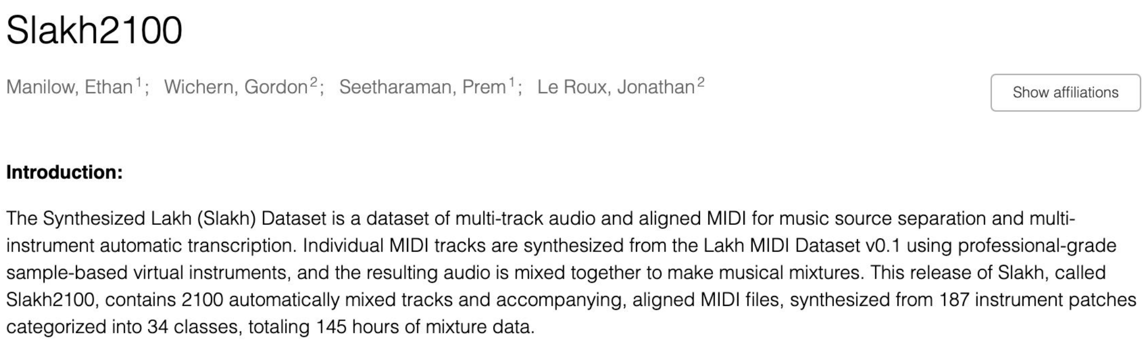
Speech recognition: >300 hrs data
Music sequence classification: 280 GB worth data

More datasets



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.

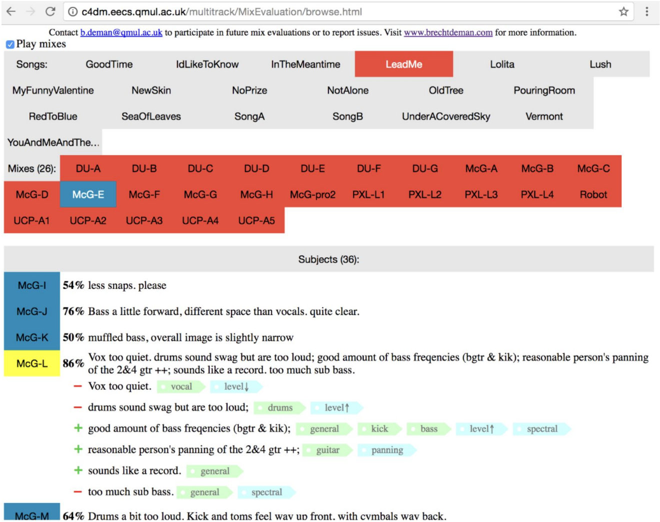


Slakh2100

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

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.



c4dm.eecs.qmul.ac.uk/multitrack/MixEvaluation/browse.html

Contact b.deman@qmul.ac.uk to participate in future mix evaluations or to report issues. Visit www.brechtdeman.com for more information.

☑ Play mixes

Songs: GoodTime IdLikeToKnow InTheMeantime LeadMe Lolita Lush MyFunnyValentine NewSkin NoPrize NotAlone OldTree PouringRoom RedToBlue SeaOfLeaves SongA SongB UnderACoveredSky Vermont YouAndMeAndThe...

Mixes (26): DU-A DU-B DU-C DU-D DU-E DU-F DU-G McG-A McG-B McG-C McG-D McG-E McG-F McG-G McG-H McG-pro2 PXL-L1 PXL-L2 PXL-L3 PXL-L4 Robot UCP-A1 UCP-A2 UCP-A3 UCP-A4 UCP-A5

Subjects (36):

McG-I **54%** less snaps. please

McG-J **76%** Bass a little forward, different space than vocals. quite clear.

McG-K **50%** muffled bass, overall image is slightly narrow

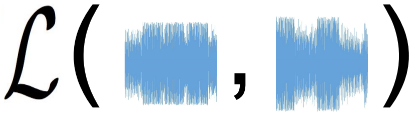
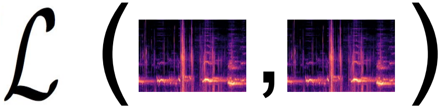
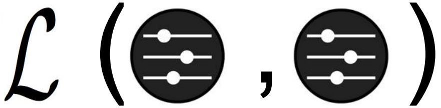
McG-L **86%** Vox too quiet. drums sound swag but are too loud; good amount of bass freqencies (bgtr & kik); reasonable person's panning of the 2&4 gtr ++; sounds like a record. too much sub bass.

- − Vox too quiet. vocal level↓
- − drums sound swag but are too loud; drums level↑
- + good amount of bass freqencies (bgtr & kik); general kick bass level↑ spectral
- + reasonable person's panning of the 2&4 gtr ++; guitar panning
- + sounds like a record. general
- − too much sub bass. general spectral

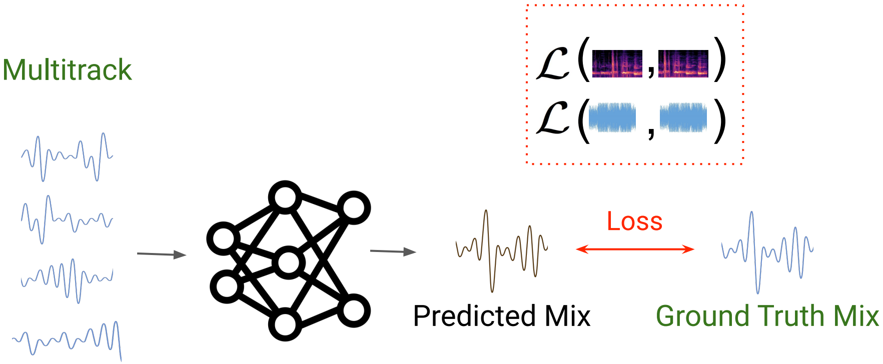
McG-M **64%** Drums a bit too loud. Kick and toms feel way up front. with cymbals way back.

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

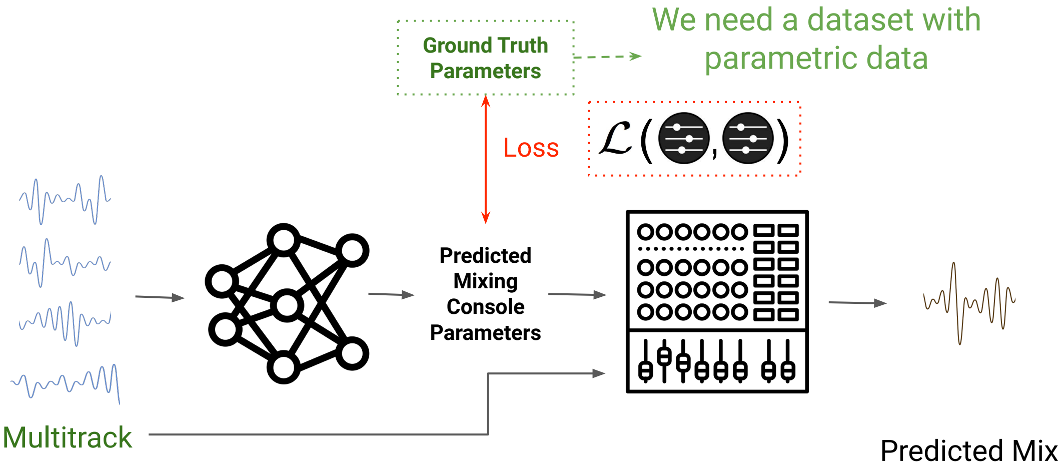
Model Types



Direct Transformation

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

Model Types

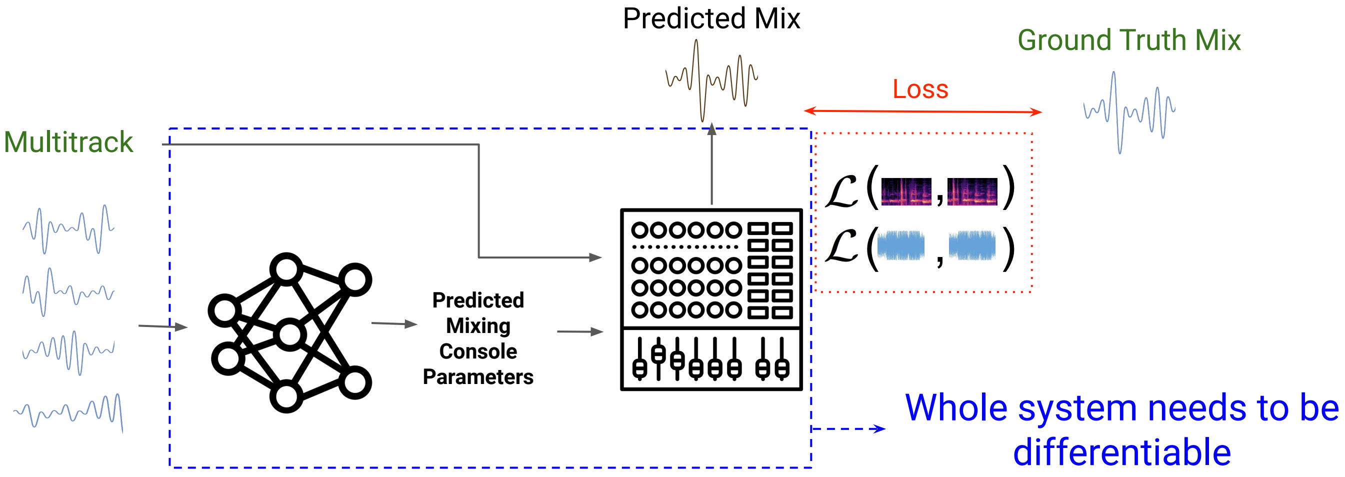


Ground Truth Parameters
We need a dataset with parametric data
Loss
$\mathcal{L}(\cdot,\cdot)$
Predicted Mixing Console Parameters
Multitrack
Predicted Mix

Parameter Estimation
(Parameter Loss)

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

Model Types



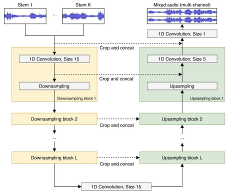
Whole system needs to be differentiable

Parameter Estimation
(Audio Loss)

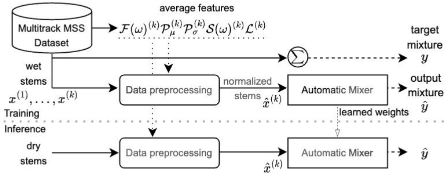
Black box system that allows interpretability and controllability (context not incorporated)

State of the Art

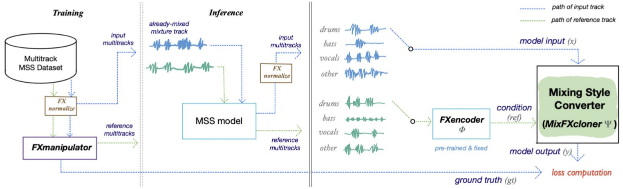
Direct Transformation



Wave-U-Net for drum mixing [a]

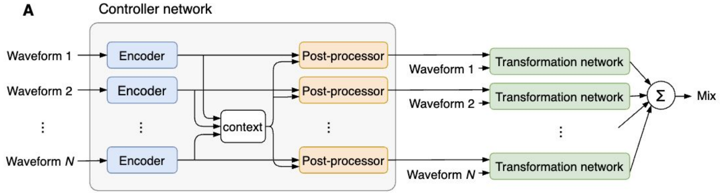


Mixing with out-of-domain data [c]



Mixing style transfer [d]

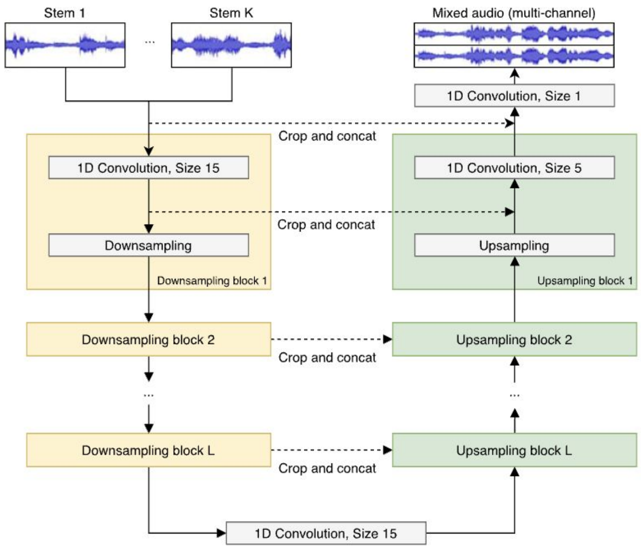
Parameter Estimation

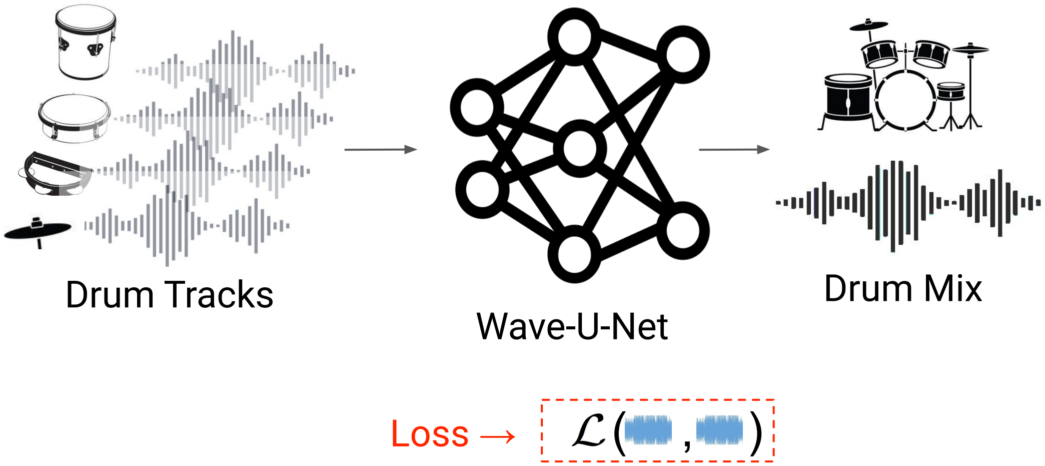


Mixing with neural mixing console [b]

[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)

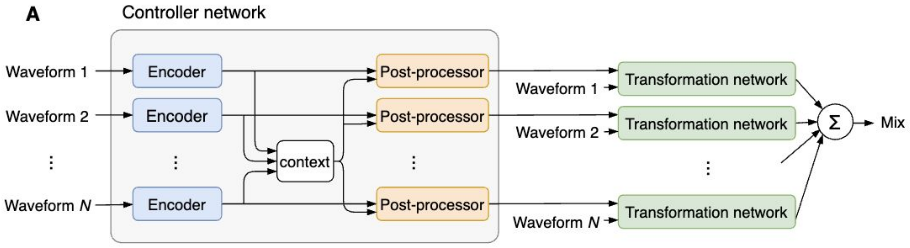
A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net



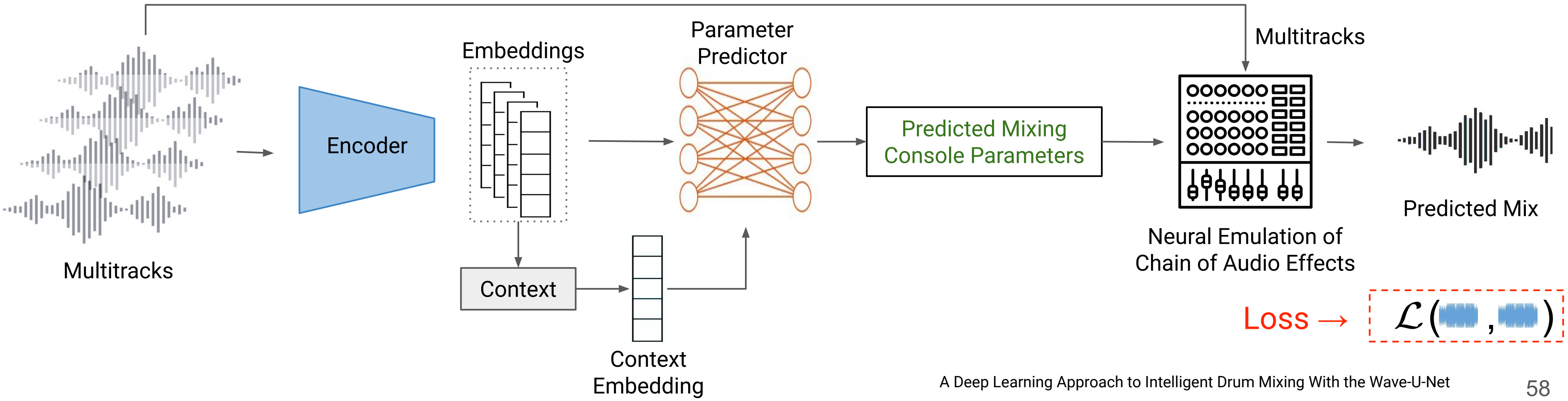


- Pros: directly learns the audio transformation
- Limitations: **Only drum mixing**, number of tracks is fixed

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

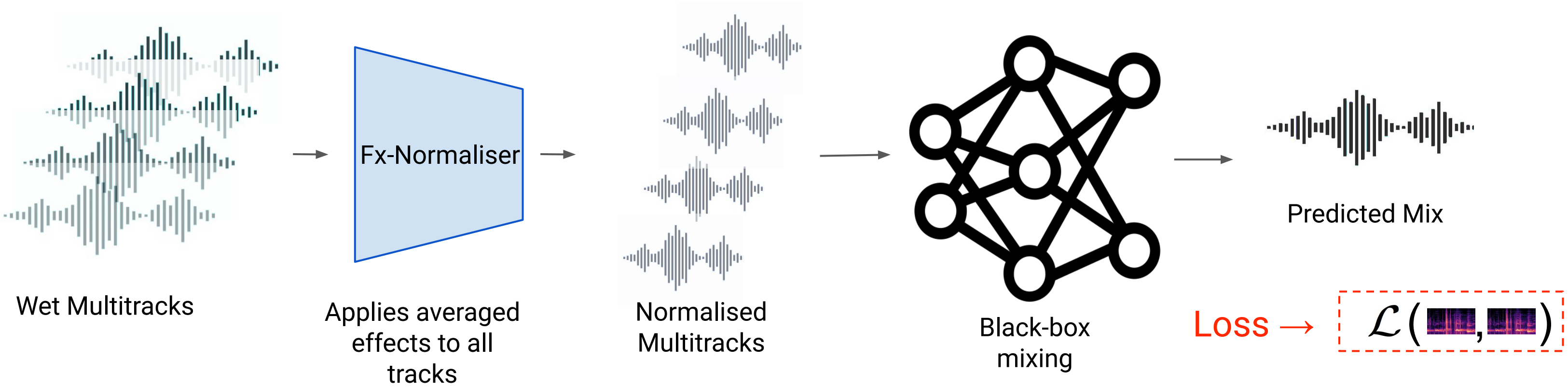


- Pros: Permutation invariant, works for any number of tracks, allows multitrack mixing
- Limitations: neural emulation of effects are difficult to train, **doesn't work well for all cases (Could be due to lack of enough data)**



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

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



Fx-Normaliser

Wet Multitracks

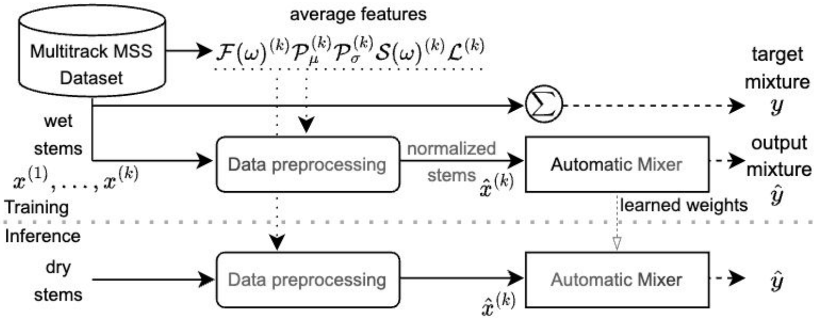
Applies averaged effects to all tracks

Normalised Multitracks

Black-box mixing

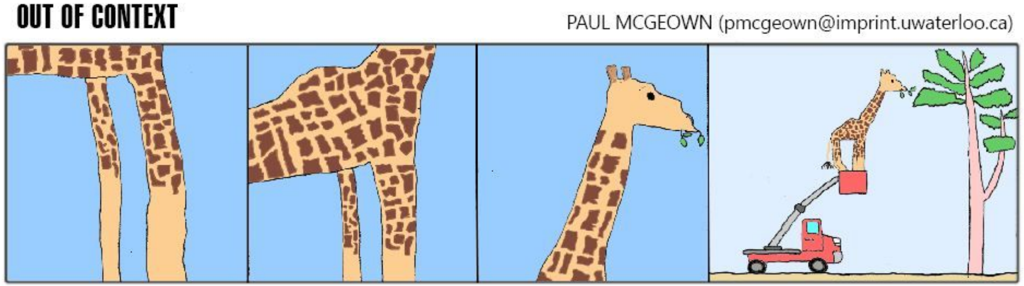
Predicted Mix

Loss → $\mathcal{L}(\ ,\ )$

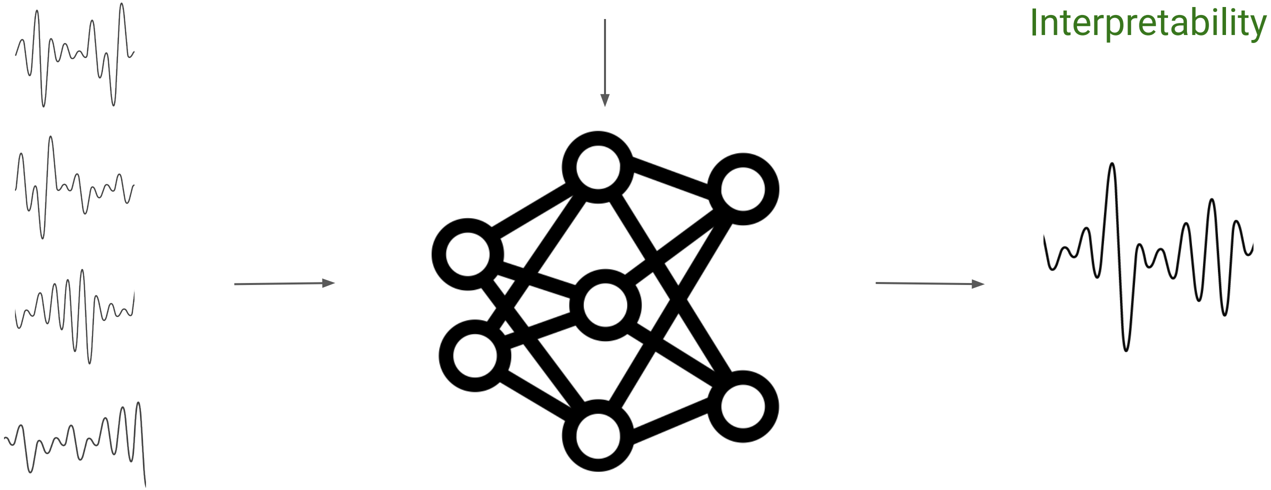


- Pros: uses of wet/processed stems to train, creates possibility for using extensive source separation datasets with wet stems
- Limitations: lacks interpretability and controllability, works for 4 stems

Limitations



What we want?



Context

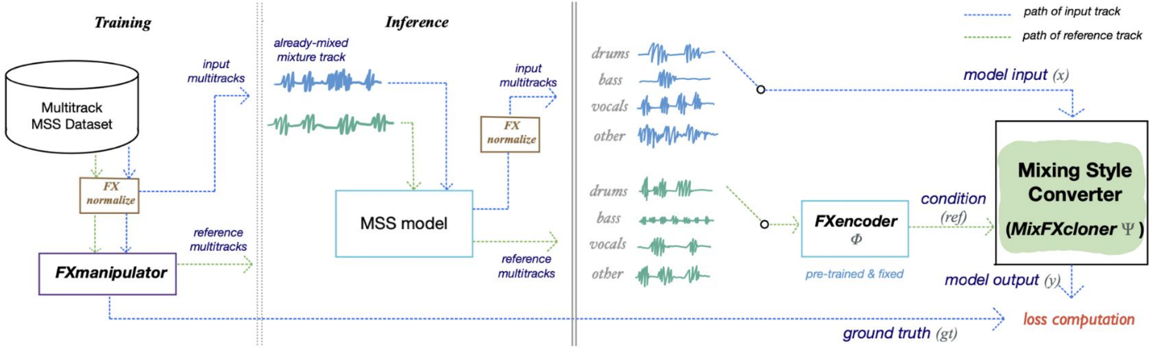
Controllability
Interpretability

Multitrack

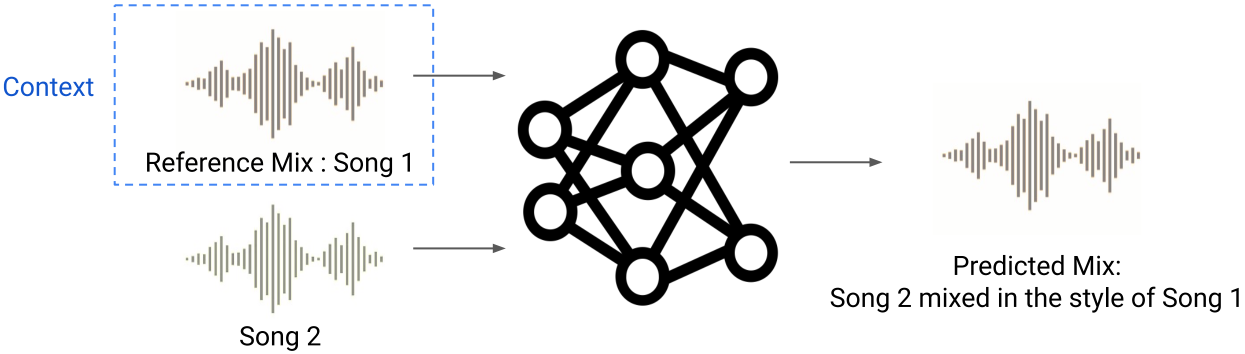
Neural
Network

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



Context

Reference Mix : Song 1

Song 2

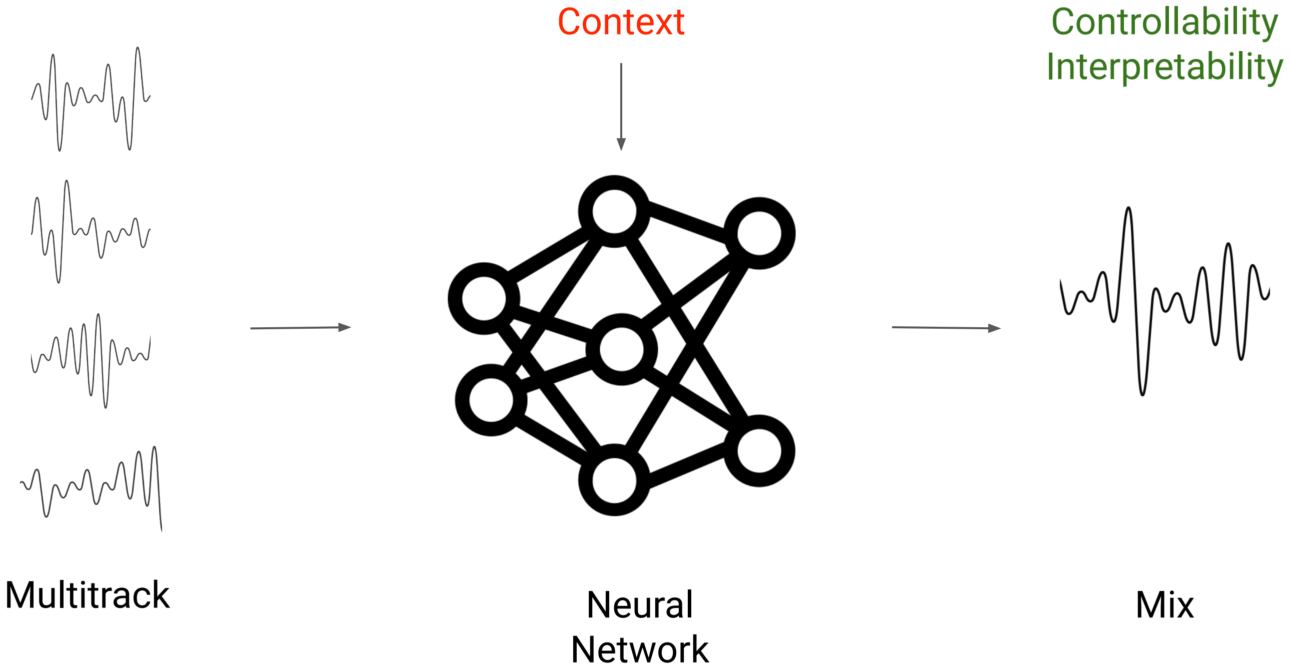
Predicted Mix:
Song 2 mixed in the style of Song 1

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?



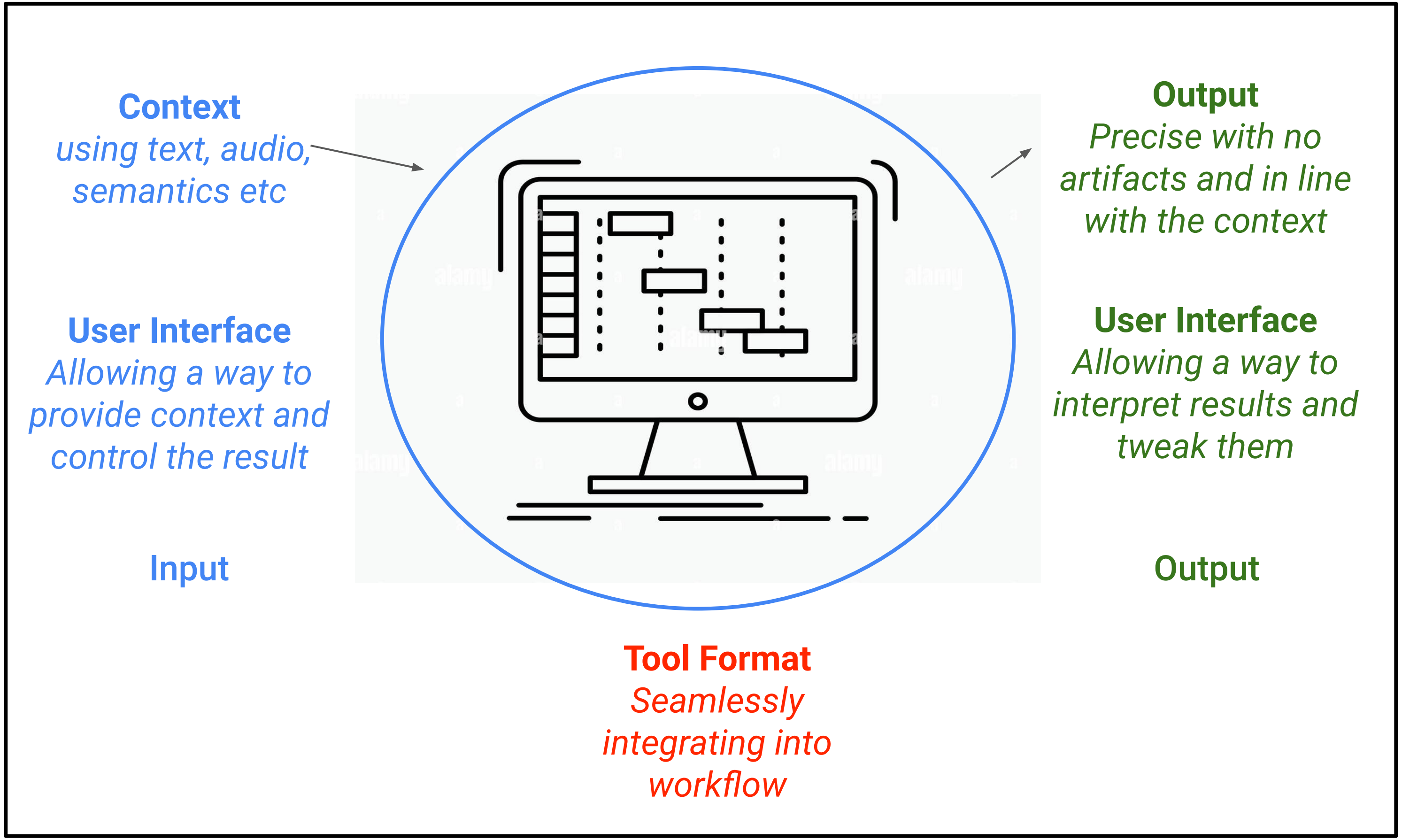
Context

Controllability
Interpretability

Multitrack

Neural
Network

Mix



Context
using text, audio, semantics etc

User Interface
Allowing a way to provide context and control the result

Input

Output
Precise with no artifacts and in line with the context

User Interface
Allowing a way to interpret results and tweak them

Output

Tool Format
Seamlessly integrating into workflow

Ideal design for an automatic mixing system

Part 3

Methods

Marco A. Martínez-Ramírez

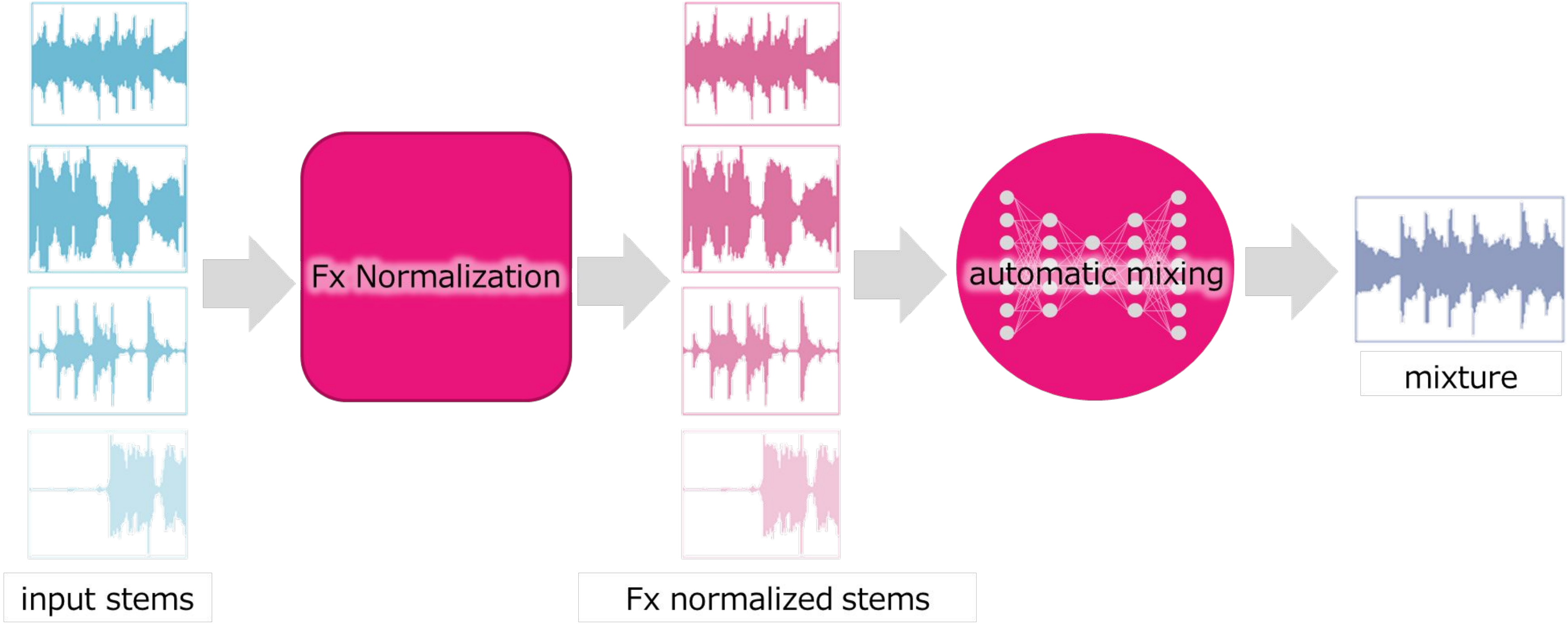
Junghyun (Tony) Koo

Christian J. Steinmetz

FX Normalization

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

Fx Normalization



Fx Normalization

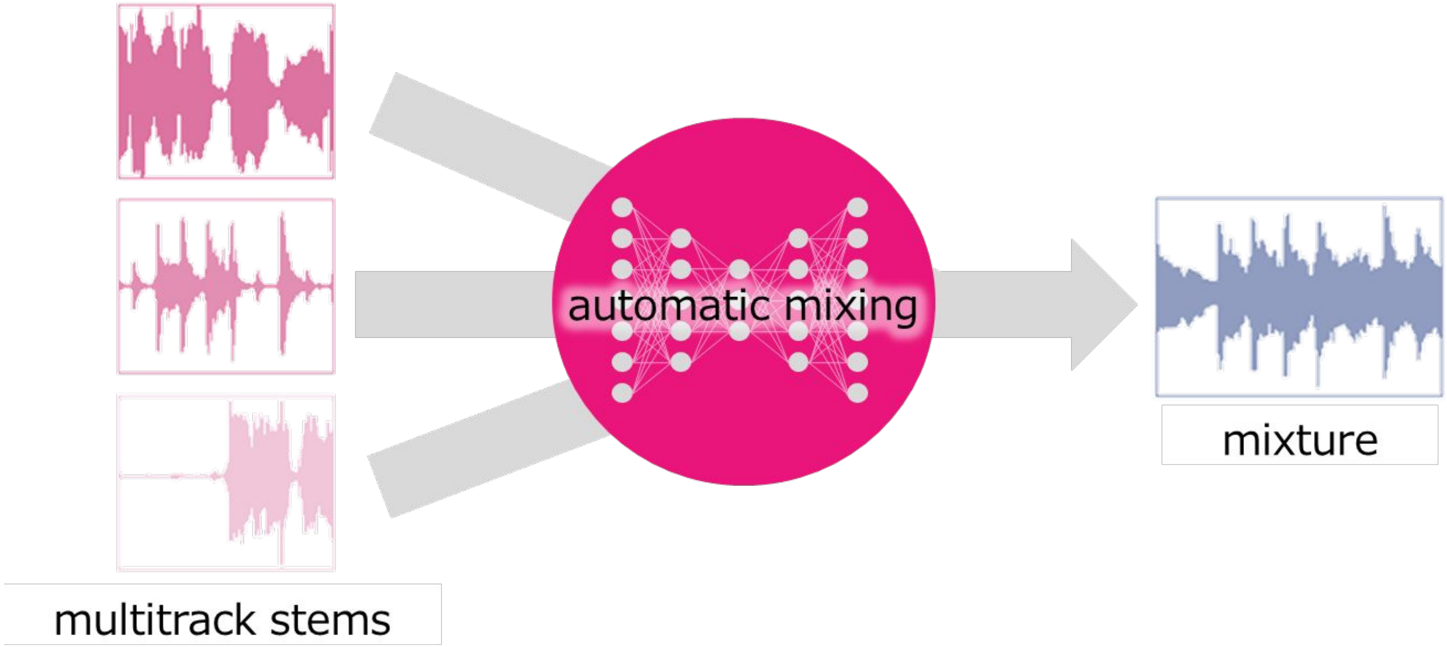
automatic mixing

mixture

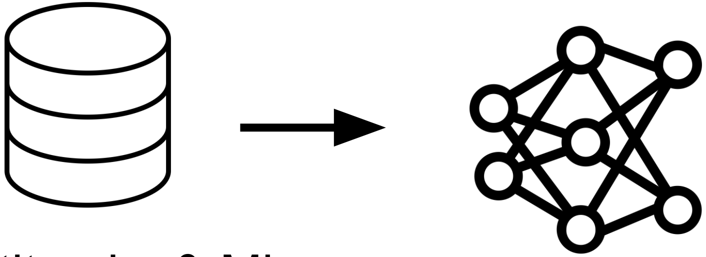
input stems

Fx normalized stems

Supervised Learning Approach



Challenging



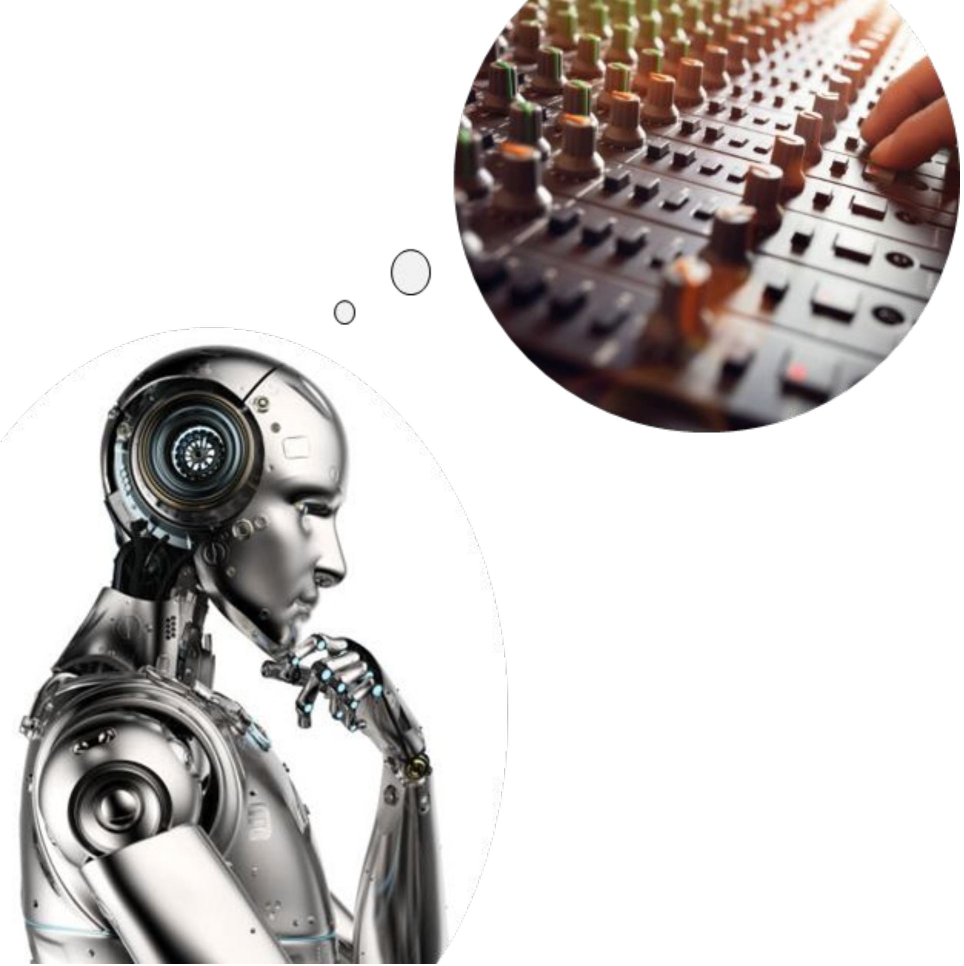
Dry multitracks & Mixes

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

Previous works

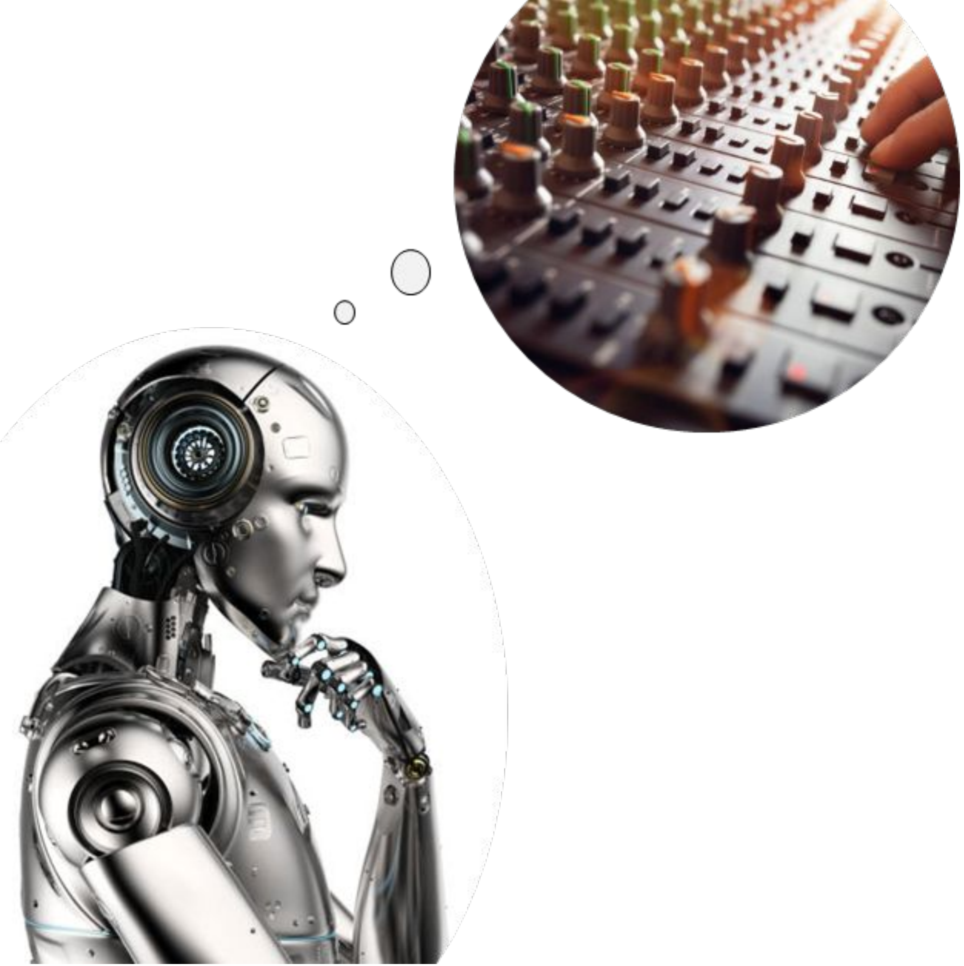
- Previous methods have not yet achieved the level of professional audio engineers mixes

- 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?*



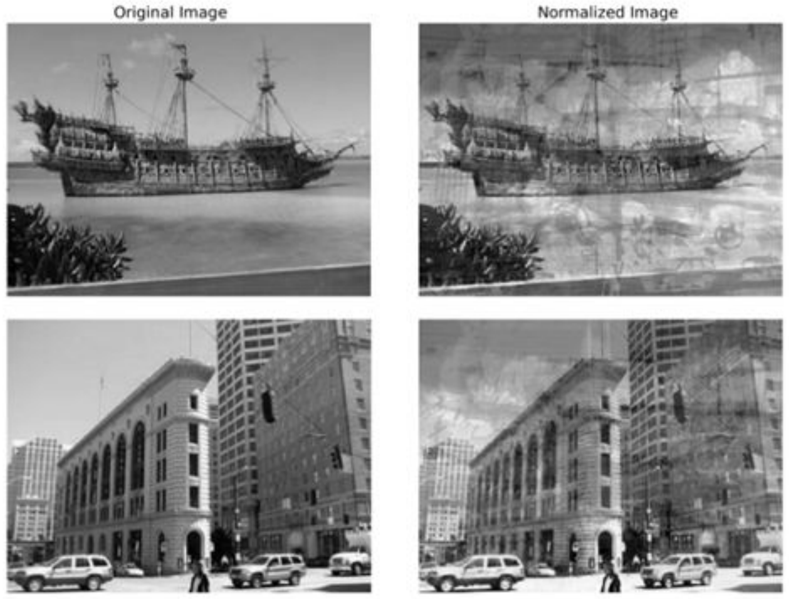
How ?

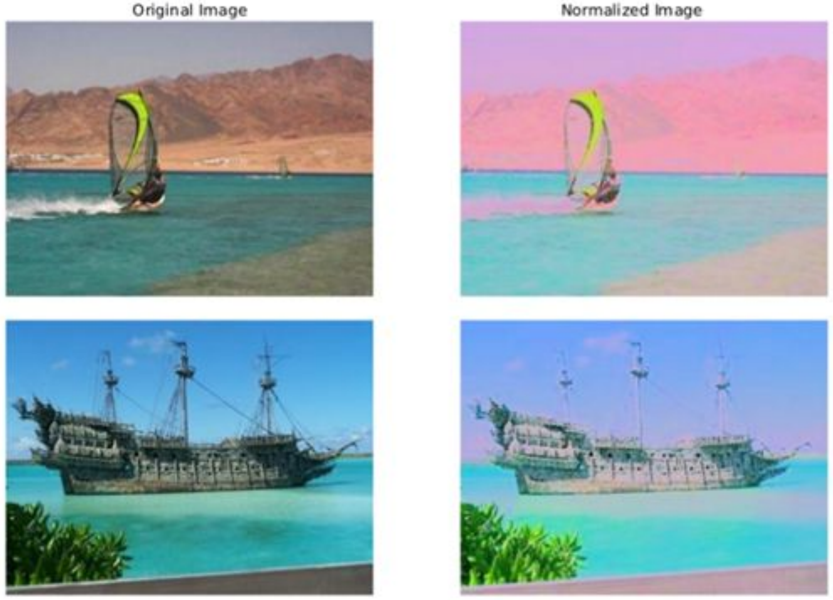
- ➢ ***Wet multitracks already contain the desired mixing effects****, which are what the networks need to learn*

🤔

Fx Normalization !

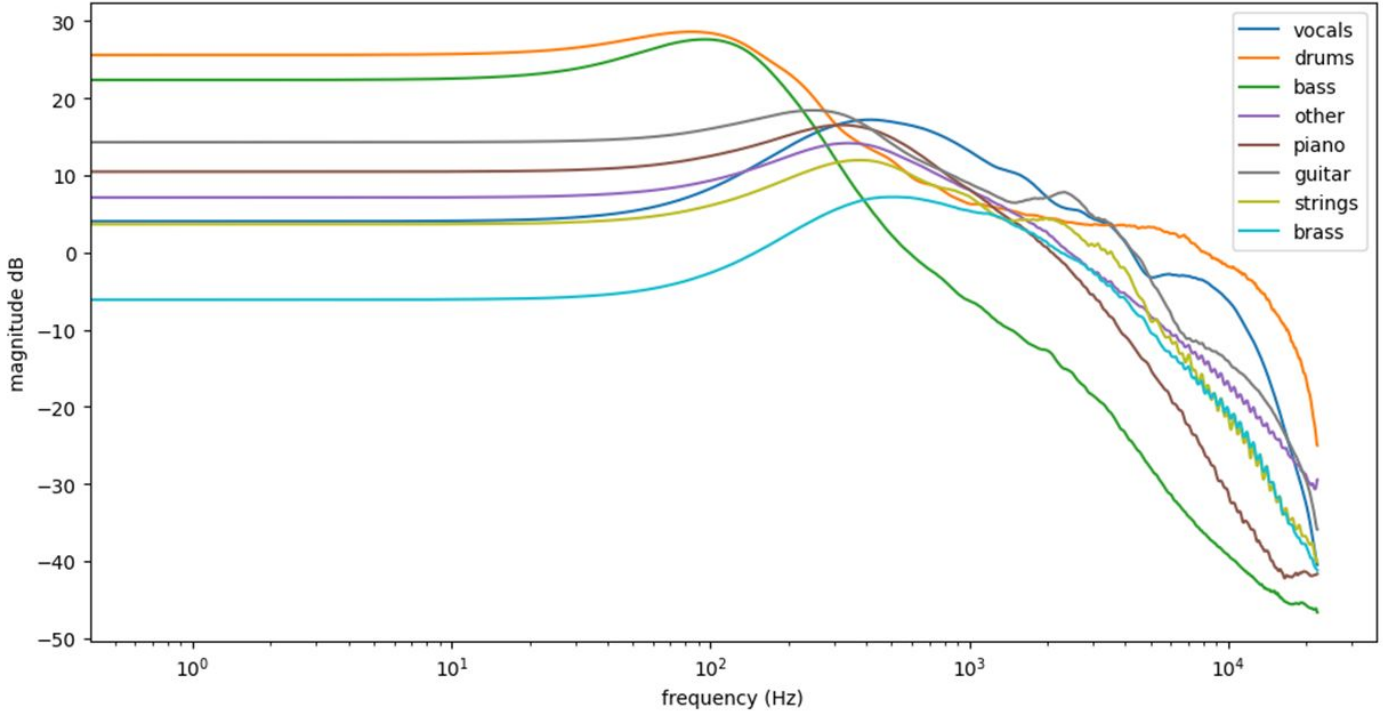
Data Normalization



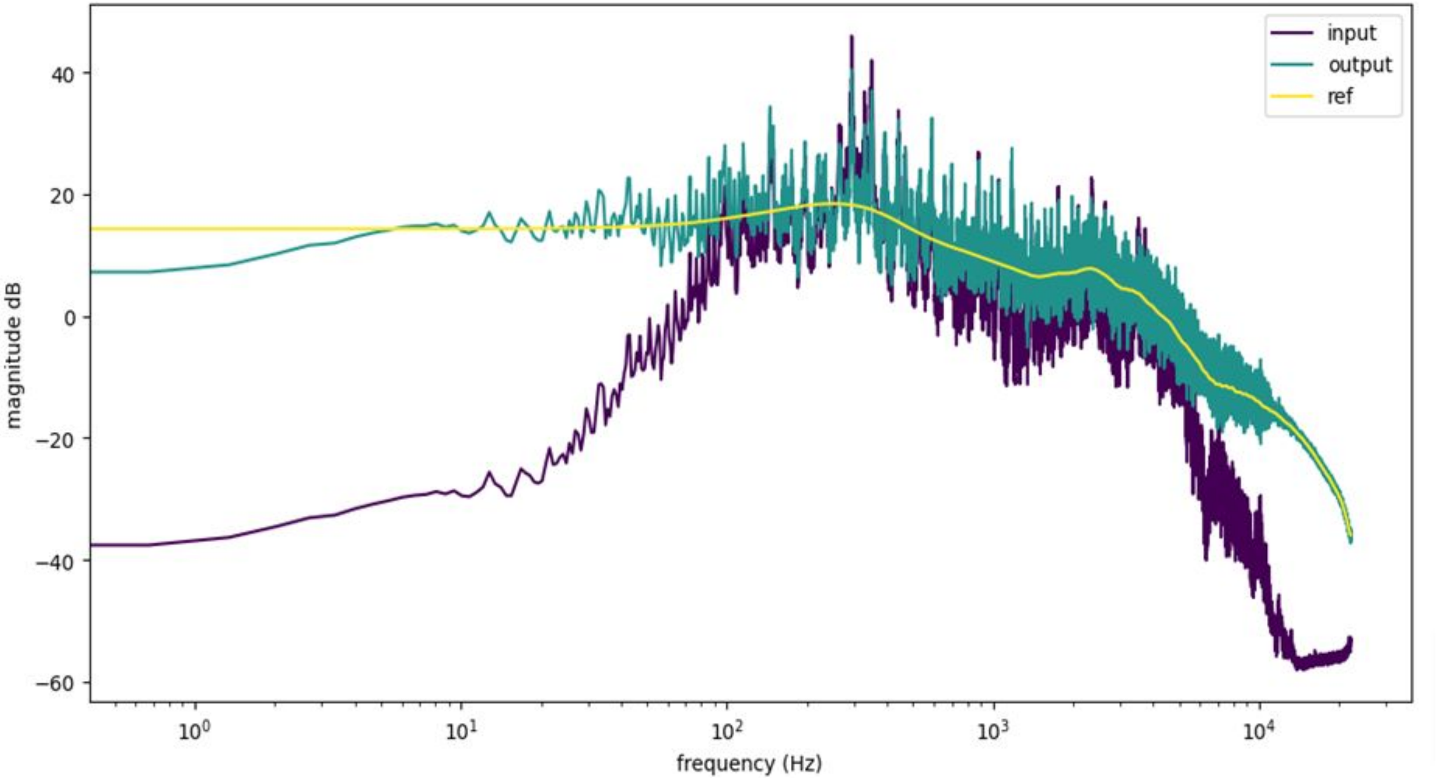


We apply the same to audio effects !

Fx Normalization–EQ average features

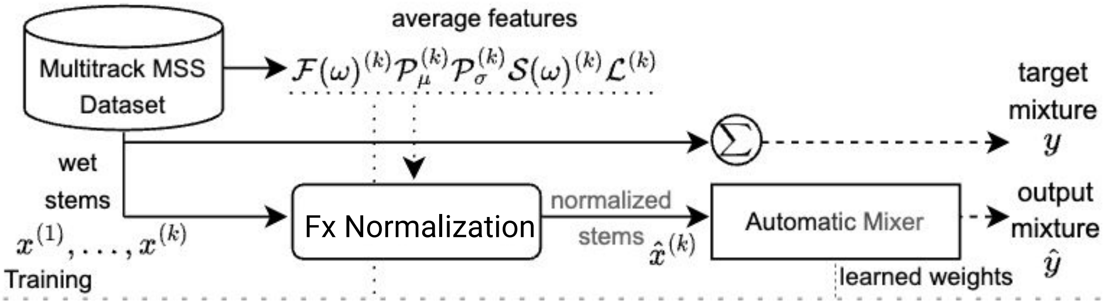


EQ Normalization



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

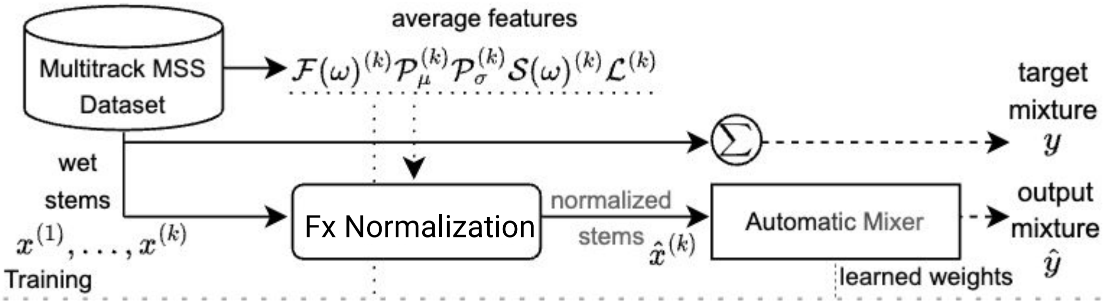
Method



Multitrack MSS Dataset
average features
$\mathcal{F}(\omega)^{(k)} \mathcal{P}_{\mu}^{(k)} \mathcal{P}_{\sigma}^{(k)} \mathcal{S}(\omega)^{(k)} \mathcal{L}^{(k)}$
$\sum$
target mixture y
wet stems
$x^{(1)}, \ldots, x^{(k)}$
Training
Fx Normalization
normalized stems $\hat{x}^{(k)}$
Automatic Mixer
output mixture $\hat{y}$
learned weights

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

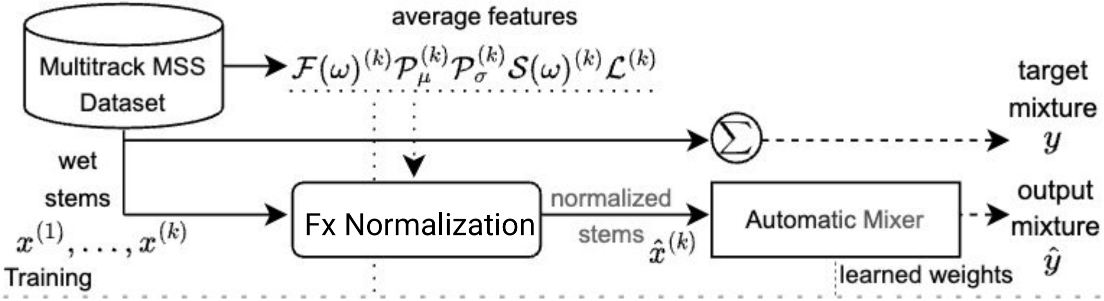
Method



Multitrack MSS Dataset
average features
$\mathcal{F}(\omega)^{(k)} \mathcal{P}_{\mu}^{(k)} \mathcal{P}_{\sigma}^{(k)} \mathcal{S}(\omega)^{(k)} \mathcal{L}^{(k)}$
target mixture y
$\sum$
wet stems
$x^{(1)}, \ldots, x^{(k)}$
Fx Normalization
normalized stems $\hat{x}^{(k)}$
Automatic Mixer
output mixture $\hat{y}$
Training
learned weights

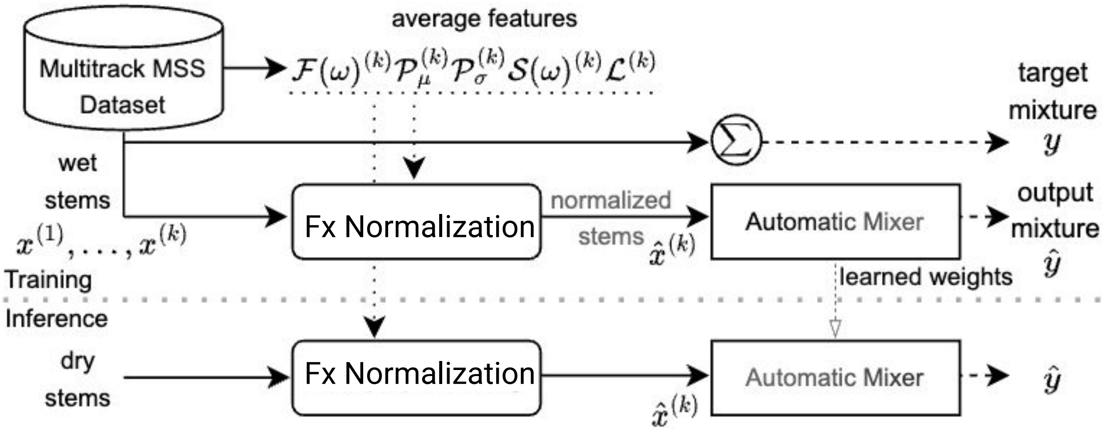
- Based on these features, we “effect-normalize” the wet stems and then train an automatic mixing network

Method



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

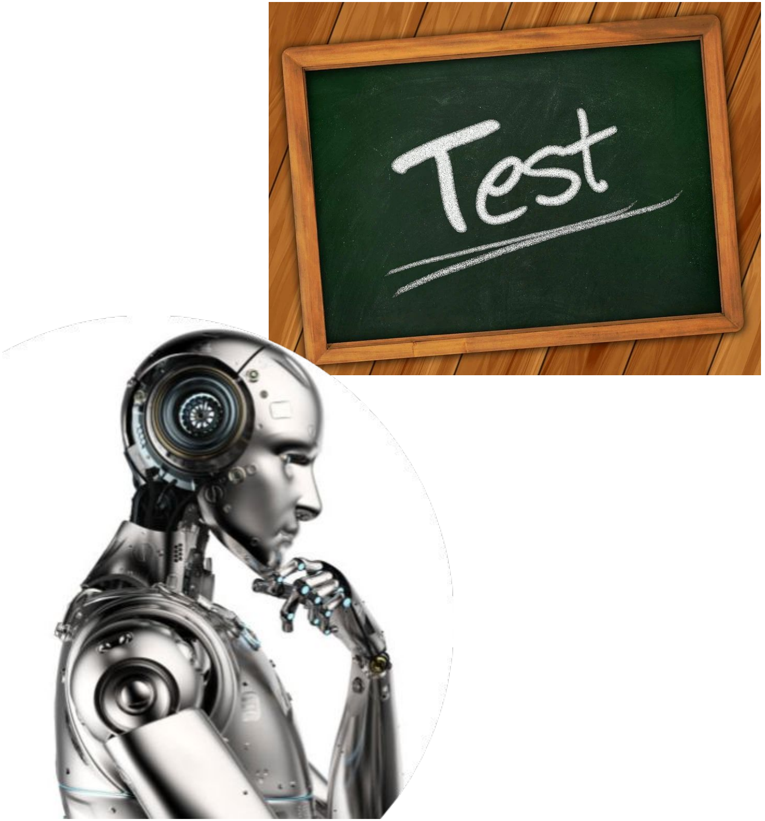
Method



- 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

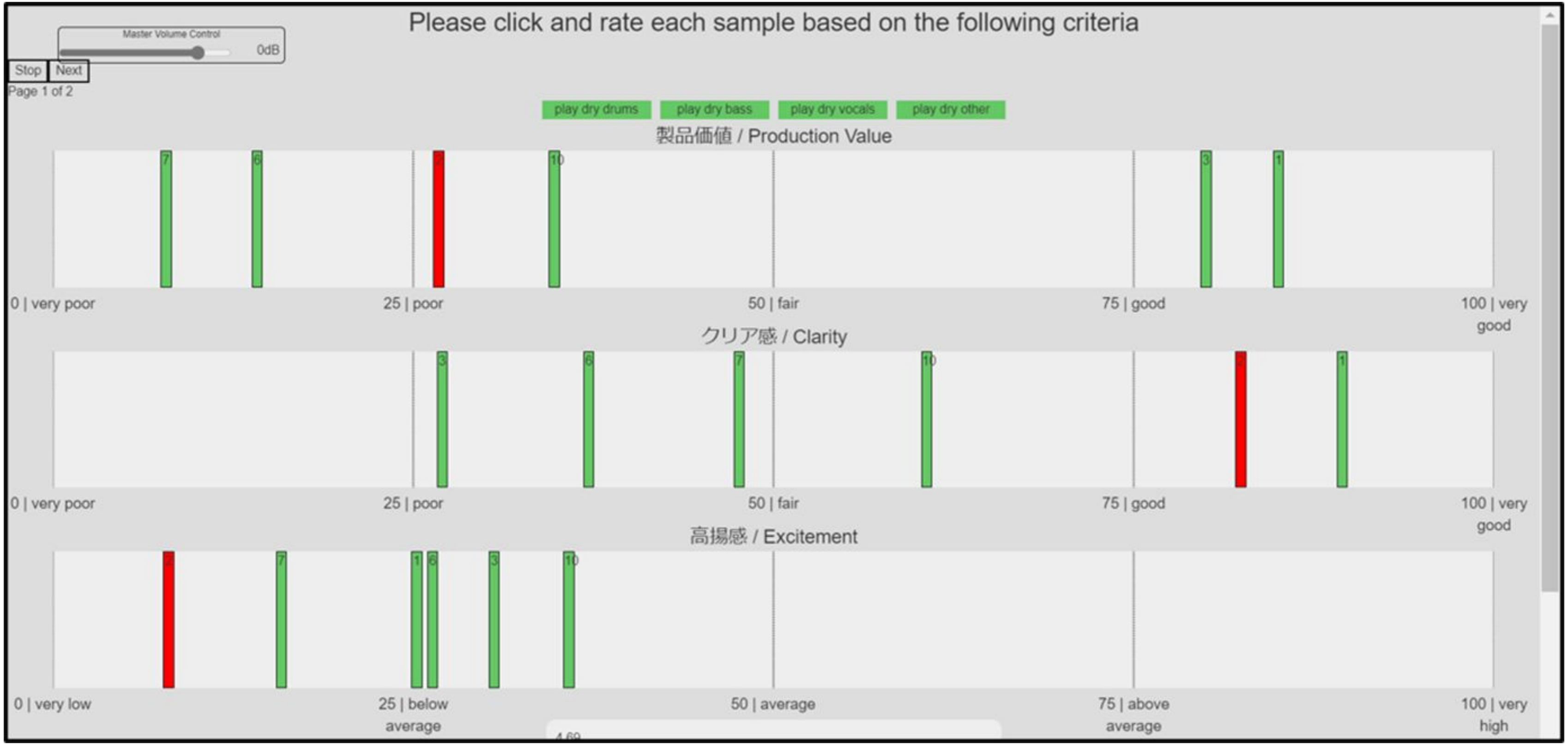
There is no standardized test type or platform

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



Please click and rate each sample based on the following criteria

Master Volume Control 0dB

Page 1 of 2

Criteria

Production Value

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

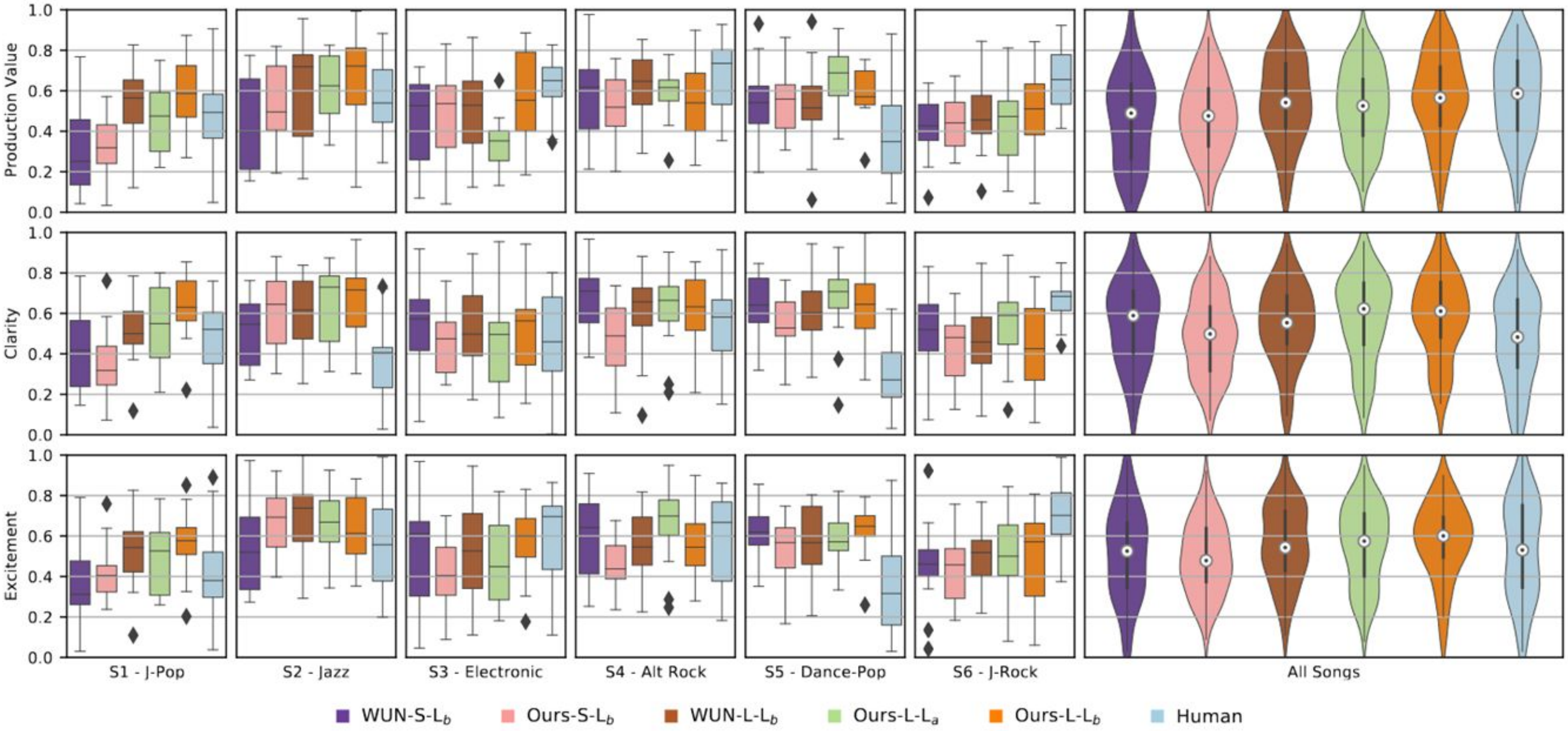
Clarity

- 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

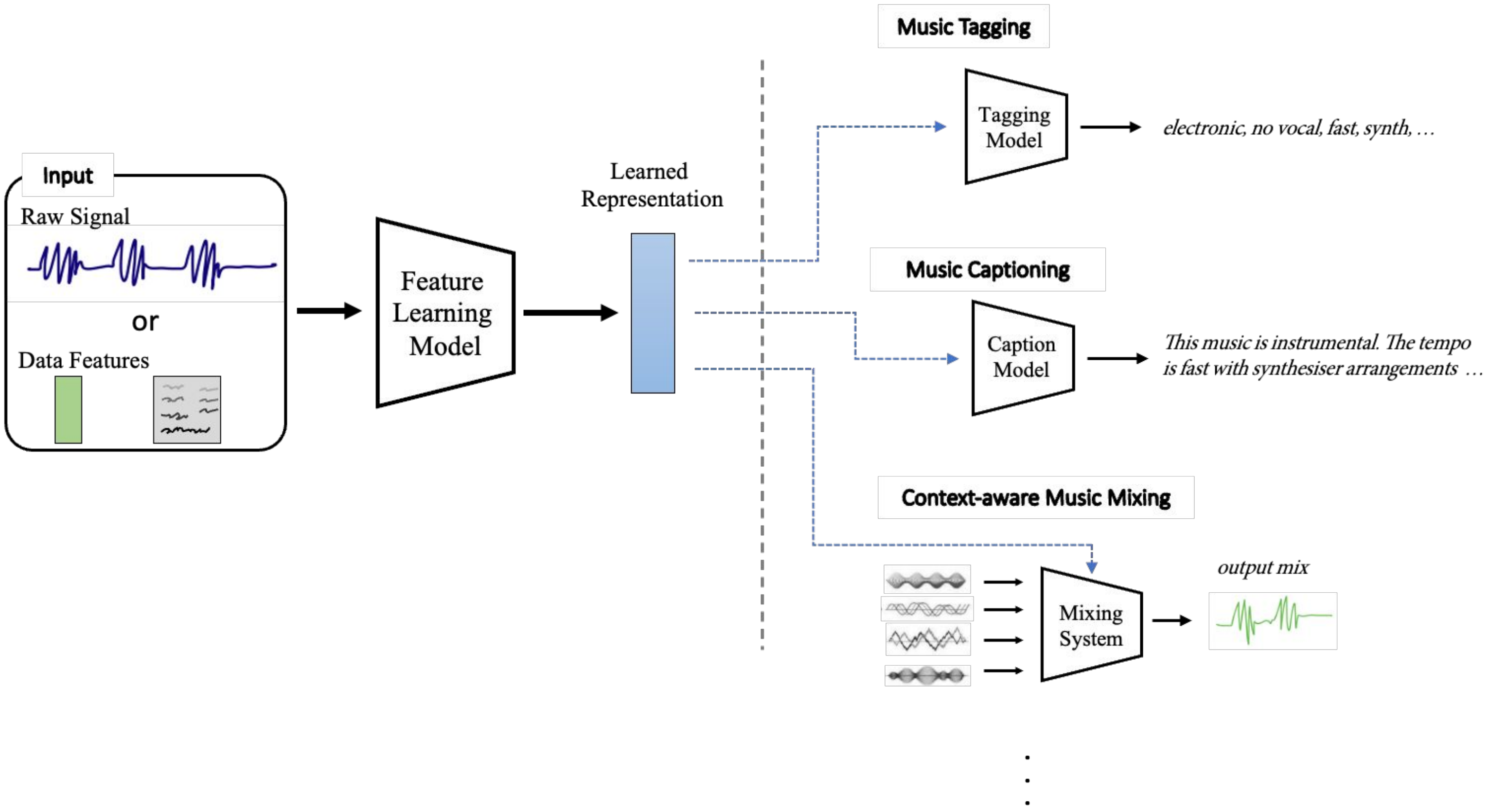
Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects ICASSP 23 Paper

Junghyun (Tony) Koo

What is Feature Learning?

Feature Learning

Task Transfer



Input

Raw Signal

or

Data Features

Feature Learning Model

Learned Representation

Music Tagging

Tagging Model

electronic, no vocal, fast, synth, …

Music Captioning

Caption Model

This music is instrumental. The tempo is fast with synthesiser arrangements …

Context-aware Music Mixing

Mixing System

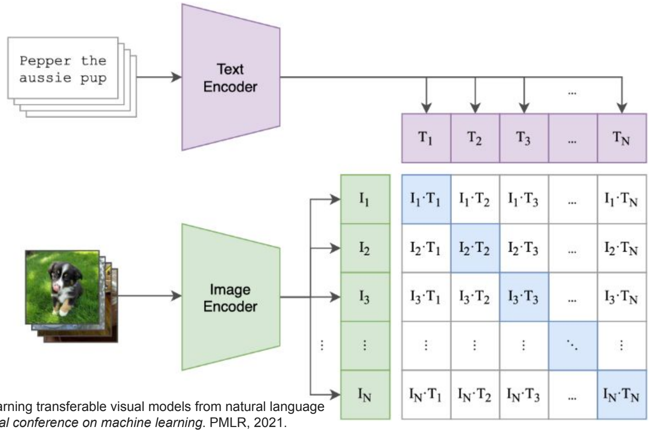
output mix

⋮

Contrastive Learning - Recent Applications

Contrastive Pre-training

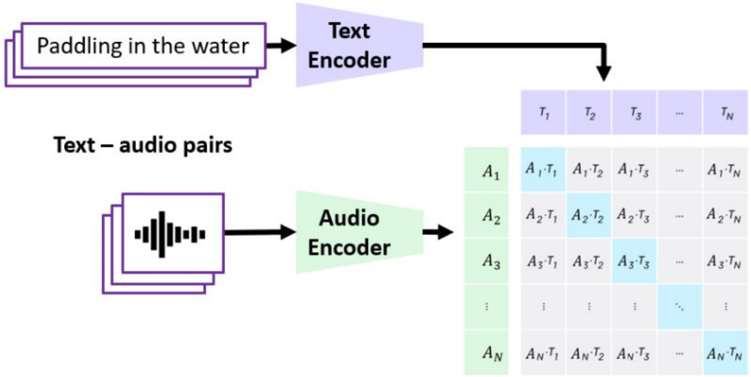
Image



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

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

Audio



	T_1	T_2	T_3	...	T_N
A_1	$A_1 \cdot T_1$	$A_1 \cdot T_2$	$A_1 \cdot T_3$	...	$A_1 \cdot T_N$
A_2	$A_2 \cdot T_1$	$A_2 \cdot T_2$	$A_2 \cdot T_3$	...	$A_2 \cdot T_N$
A_3	$A_3 \cdot T_1$	$A_3 \cdot T_2$	$A_3 \cdot T_3$	...	$A_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
A_N	$A_N \cdot T_1$	$A_N \cdot T_2$	$A_N \cdot T_3$	...	$A_N \cdot T_N$

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

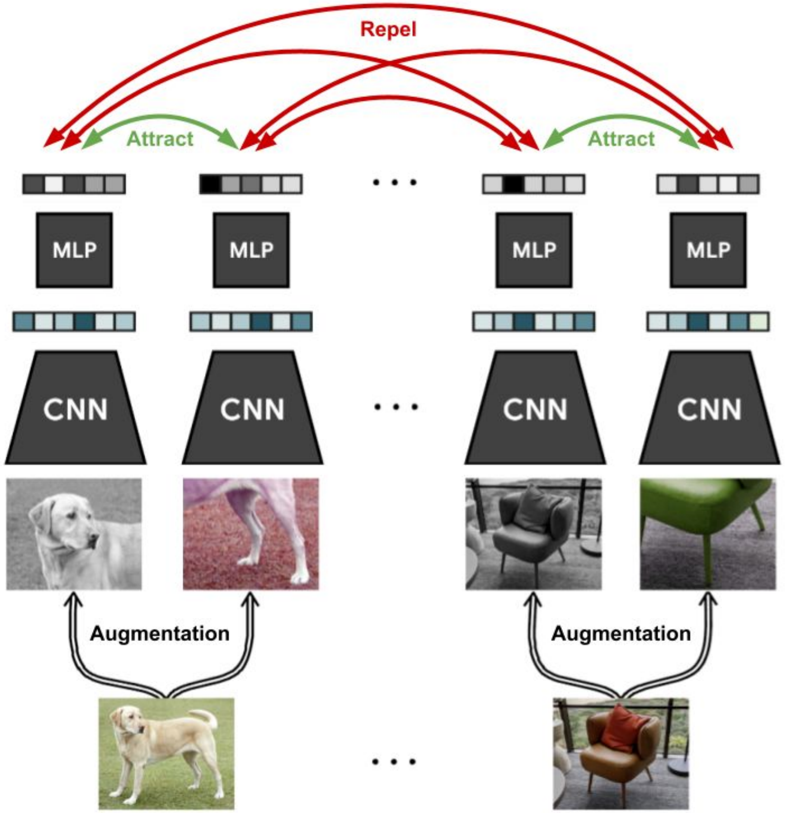
AudioGen

Meta MusicGen AI

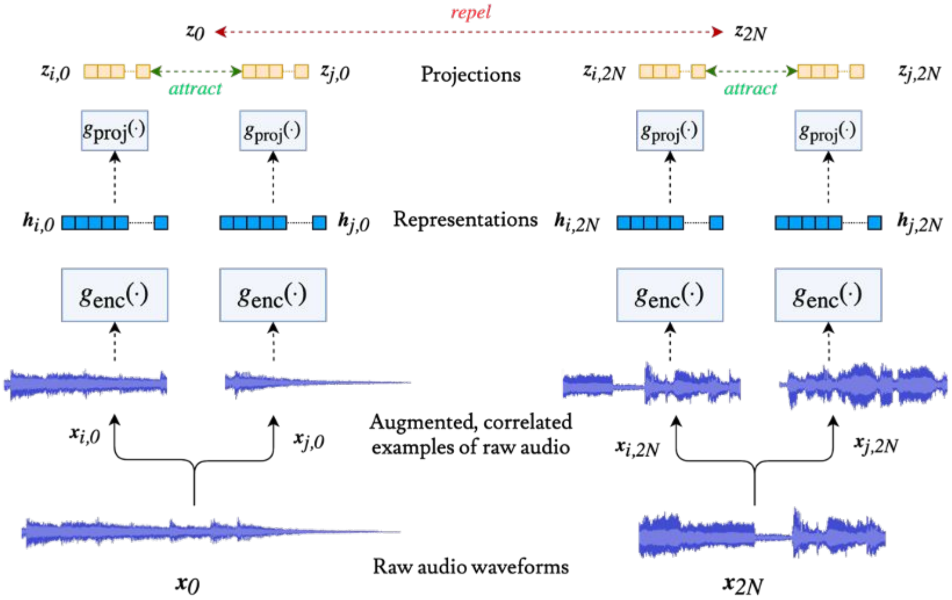
Mubert

Contrastive Learning - Training Method

SimCLR



CLMR



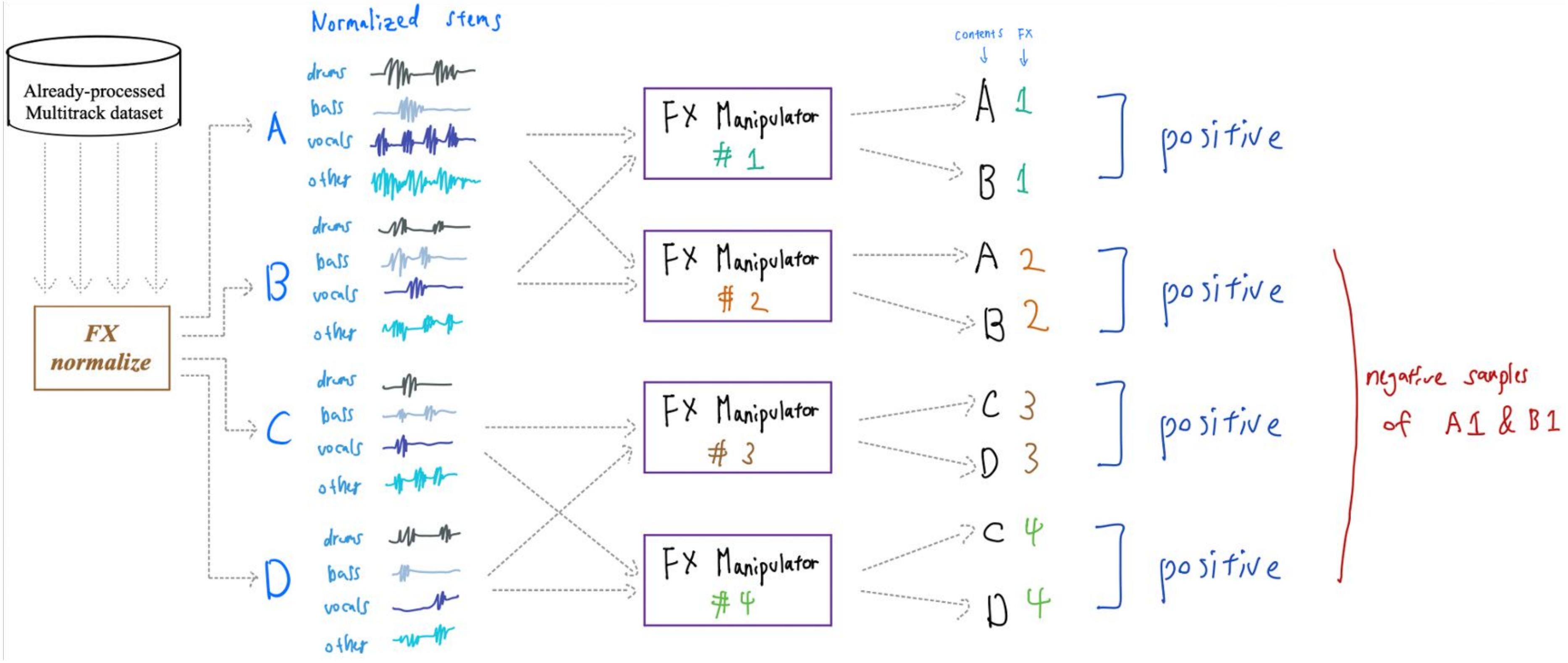
Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

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.

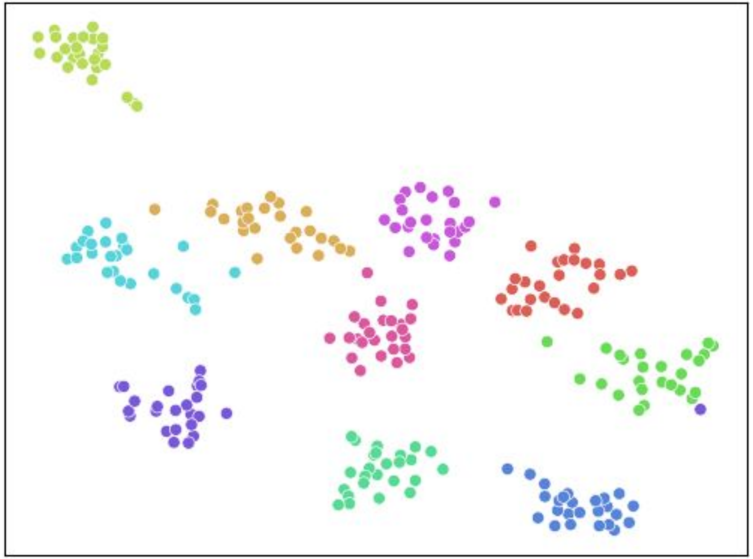
Training Procedure of the FXencoder



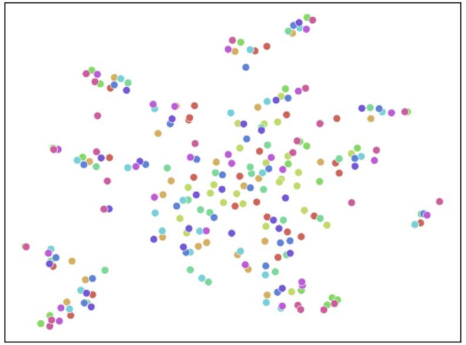
Koo, Junghyun, et al. "Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects." *ICASSP 2023*. IEEE, 2023.

Disentangled Representation

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



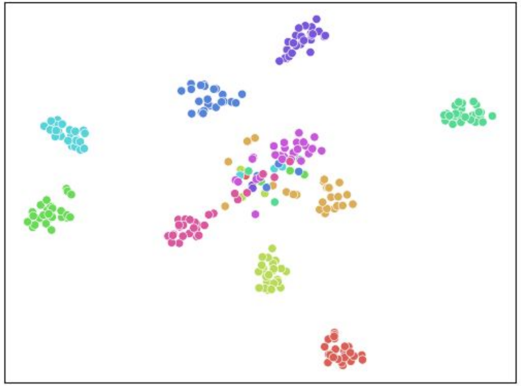
FXencoder



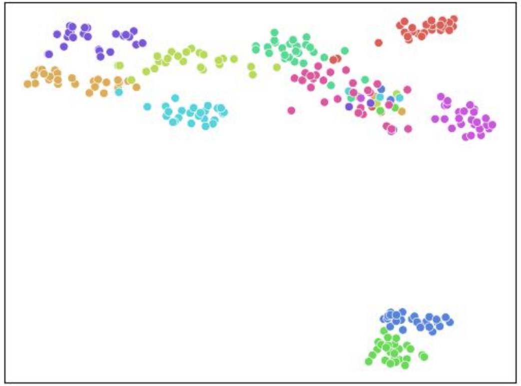
MEE
(model trained with standard approach)

Disentangled Representation - Individual Instrument

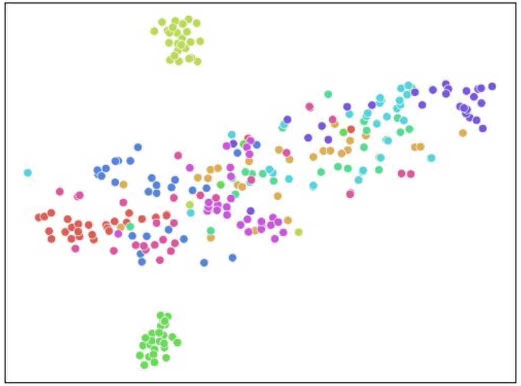
drums



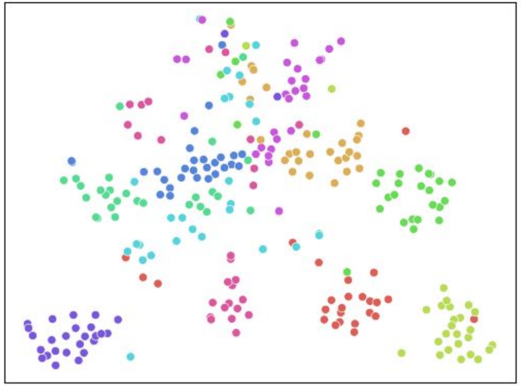
vocals



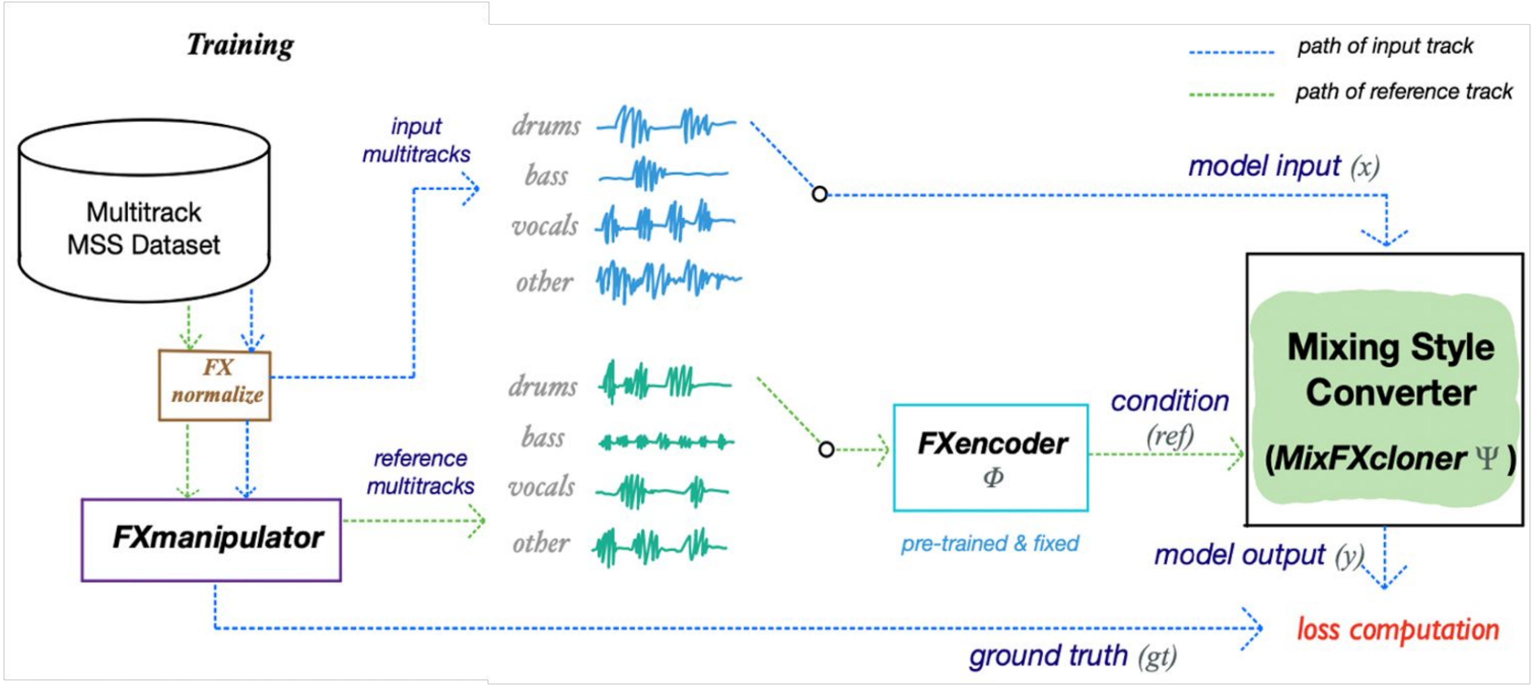
bass



other

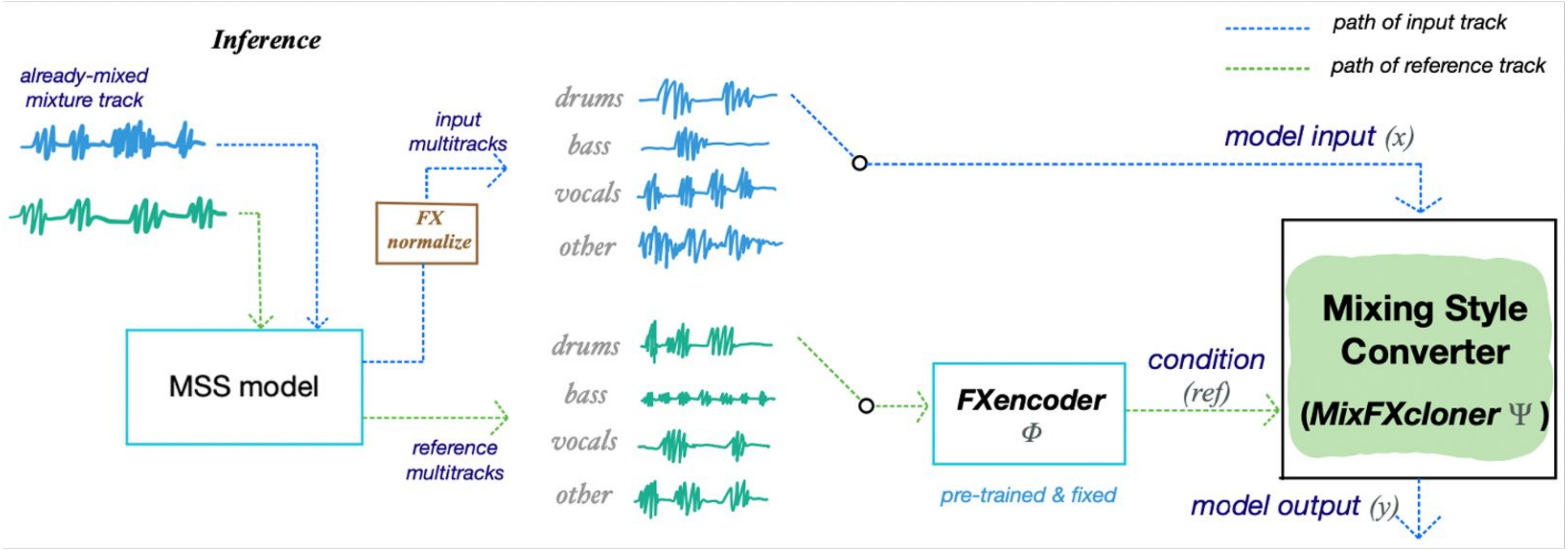


Music Mixing Style Transfer with FXencoder



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

Music Mixing Style Transfer with FXencoder



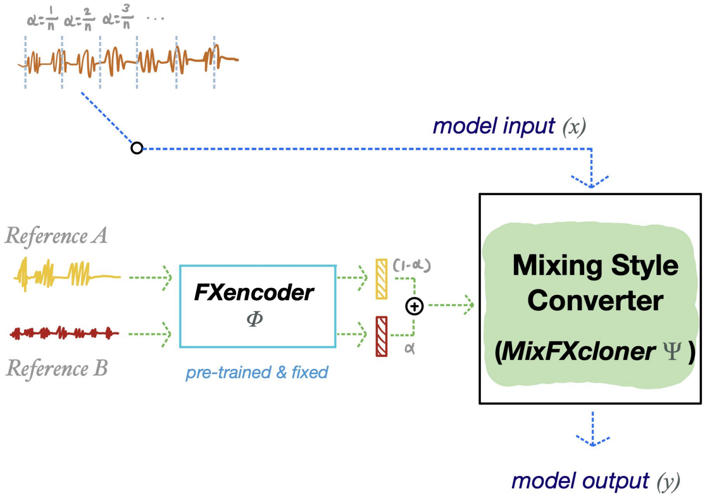
- During inference stage, we can transfer mixing style of mixture-wise inputs using a music source separation (MSS) model

Demo - Mixing Style Transfer

Input Mix: 🔊

	Reference A	Reference B
Target Style Mix	🔊	🔊
Individual Output	🔊	🔊

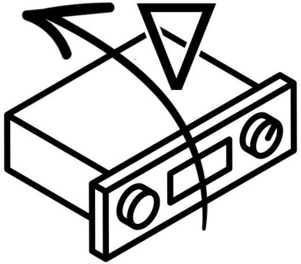
Interpolated Output 🔊



Try with your samples!

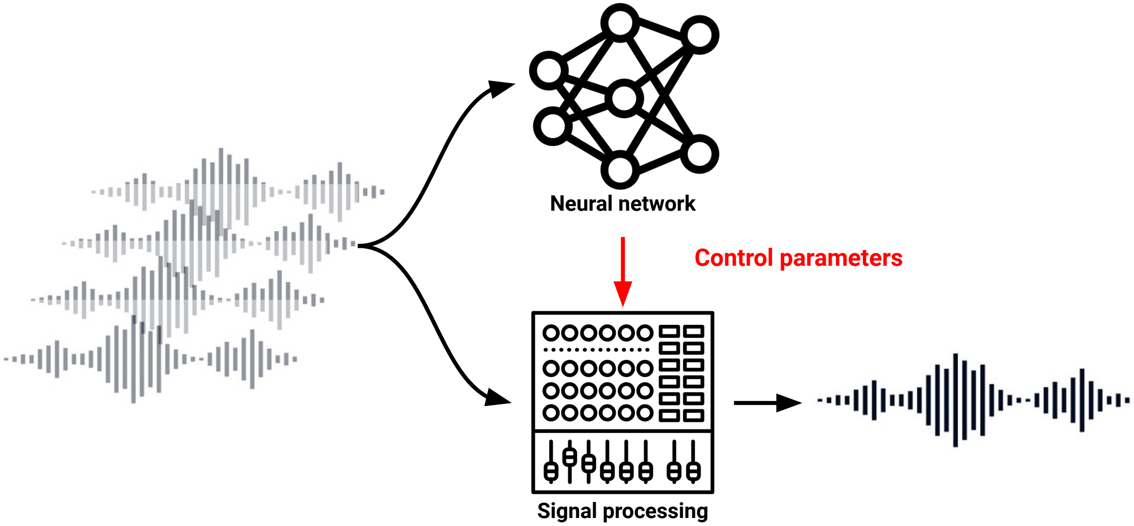
Differentiable signal processing

for automatic mixing



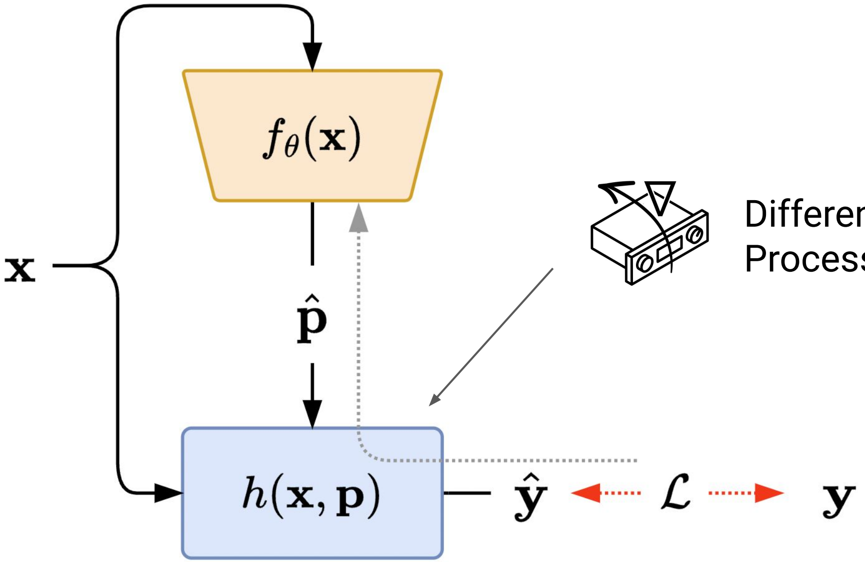
Christian Steinmetz

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



Differentiable Signal Processing

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

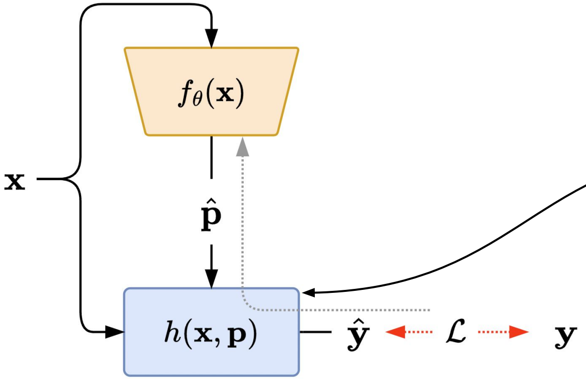
Techniques

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

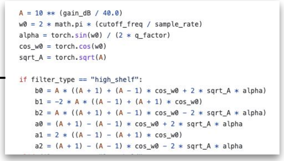
2. **Neural proxies and hybrids (NP)**
   Steinmetz et al. 2020, Steinmetz et al. 2022

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

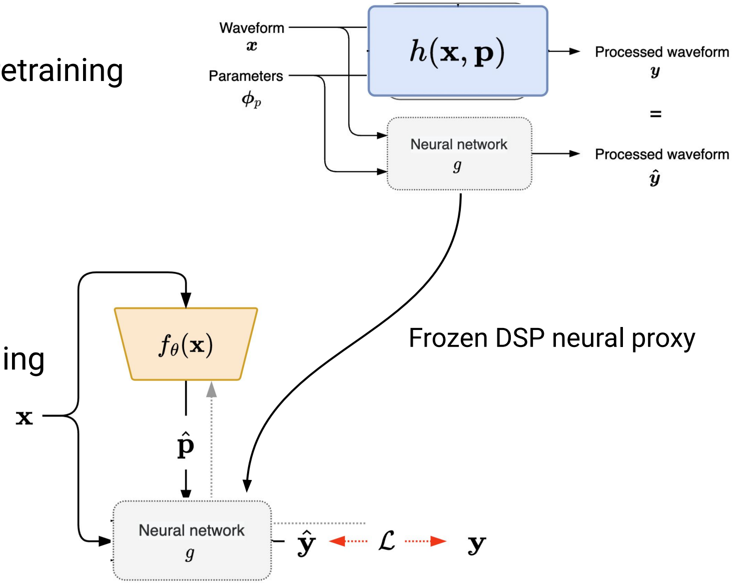
Engel, Jesse, et al. "DDSP: Differentiable digital signal processing." *ICLR* (2021).

Neural Proxy

(1) Pretraining

(2) Training

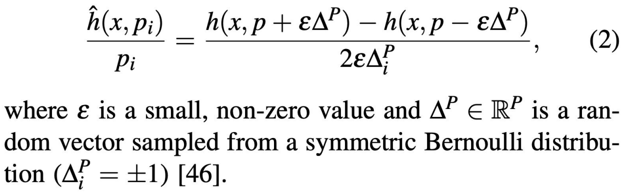
(3) Inference



Frozen DSP neural proxy

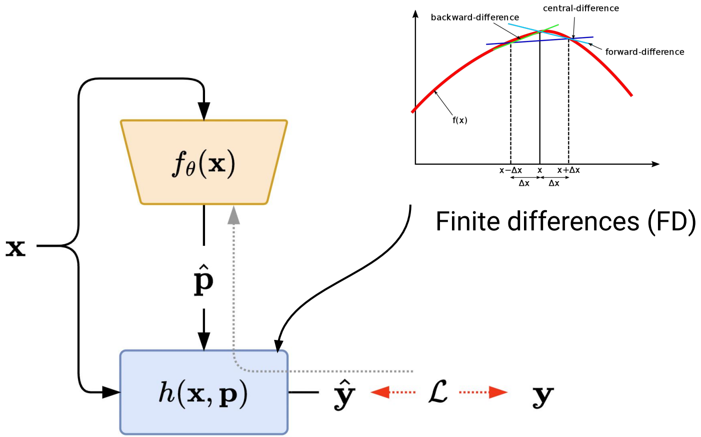
Steinmetz, Christian J., et al. "Automatic multitrack mixing with a differentiable mixing console of neural audio effects." ICASSP, 2021.

Gradient Approximation



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

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].

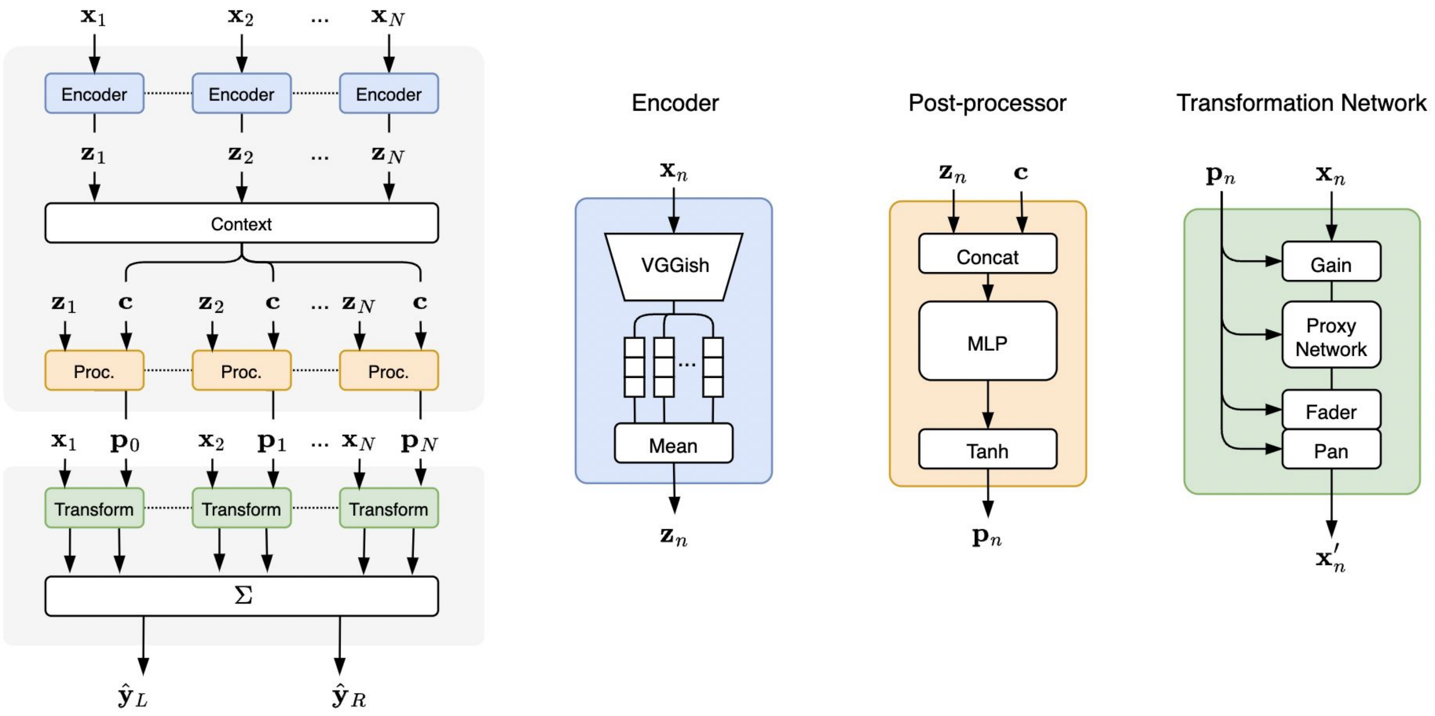


Finite differences (FD)

Simultaneous perturbation stochastic approximation (SPSA)

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

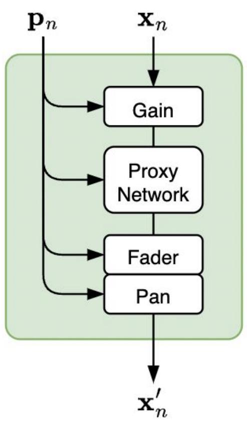
Creating a differentiable mixing console



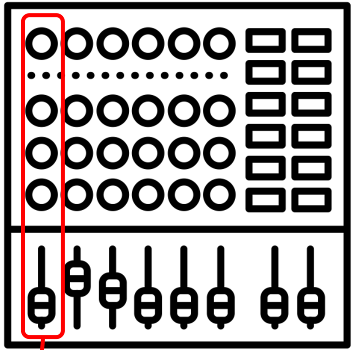
$\mathbf{x}_1$ $\mathbf{x}_2$ ... $\mathbf{x}_N$
Encoder
Encoder
Encoder
$\mathbf{z}_1$ $\mathbf{z}_2$ ... $\mathbf{z}_N$
Context
$\mathbf{z}_1$ $\mathbf{c}$ $\mathbf{z}_2$ $\mathbf{c}$ ... $\mathbf{z}_N$ $\mathbf{c}$
Proc.
Proc.
Proc.
$\mathbf{x}_1$ $\mathbf{p}_0$ $\mathbf{x}_2$ $\mathbf{p}_1$ ... $\mathbf{x}_N$ $\mathbf{p}_N$
Transform
Transform
Transform
Σ
$\hat{\mathbf{y}}_L$
$\hat{\mathbf{y}}_R$
Encoder
$\mathbf{x}_n$
VGGish
...
Mean
$\mathbf{z}_n$
Post-processor
$\mathbf{z}_n$ $\mathbf{c}$
Concat
MLP
Tanh
$\mathbf{p}_n$
Transformation Network
$\mathbf{p}_n$ $\mathbf{x}_n$
Gain
Proxy Network
Fader
Pan
$\mathbf{x}'_n$

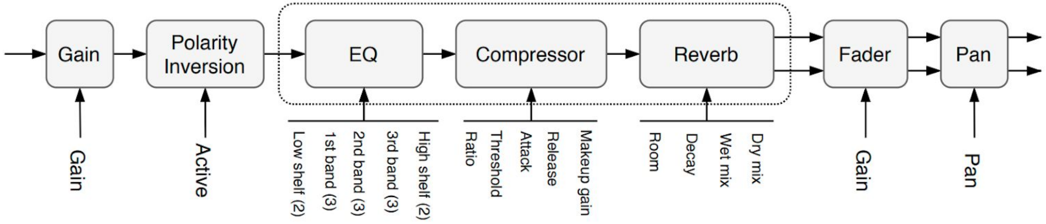
Steinmetz, Christian J., et al. "Automatic multitrack mixing with a differentiable mixing console of neural audio effects." ICASSP, 2021.

Creating a differentiable mixing console



Proxy network

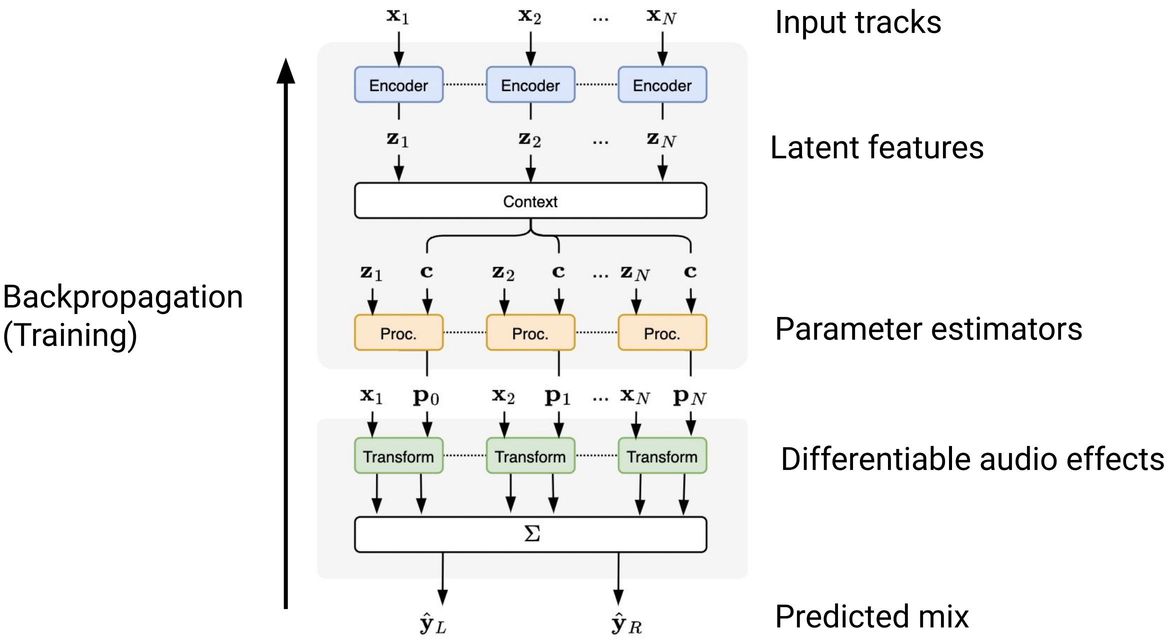




Differentiable channel strip

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

Creating a differentiable mixing console


$\mathbf{x}_1$ $\mathbf{x}_2$ ... $\mathbf{x}_N$
Encoder
Encoder
Encoder
$\mathbf{z}_1$ $\mathbf{z}_2$ ... $\mathbf{z}_N$
Context
$\mathbf{z}_1$ $\mathbf{c}$ $\mathbf{z}_2$ $\mathbf{c}$ ... $\mathbf{z}_N$ $\mathbf{c}$
Proc.
Proc.
Proc.
$\mathbf{x}_1$ $\mathbf{p}_0$ $\mathbf{x}_2$ $\mathbf{p}_1$ ... $\mathbf{x}_N$ $\mathbf{p}_N$
Transform
Transform
Transform
Σ
$\hat{\mathbf{y}}_L$ $\hat{\mathbf{y}}_R$

Input tracks

Latent features

Backpropagation
(Training)

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

roex®

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

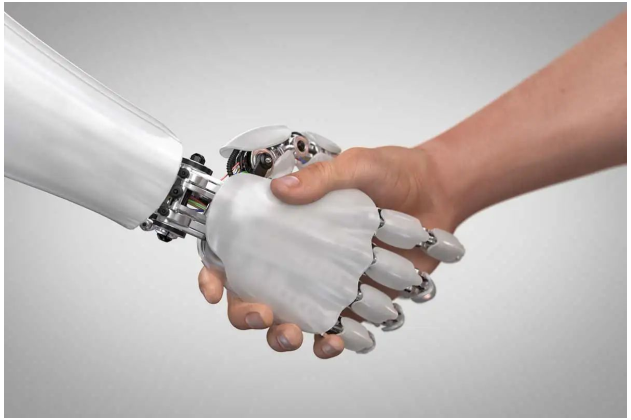
- What is a good mix? **Definition** and **target**
- **Complexity** and **variety** of genres
- **Balance** between user control and automation
- **Quality of input** audio is most likely not ideal

Current Market

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

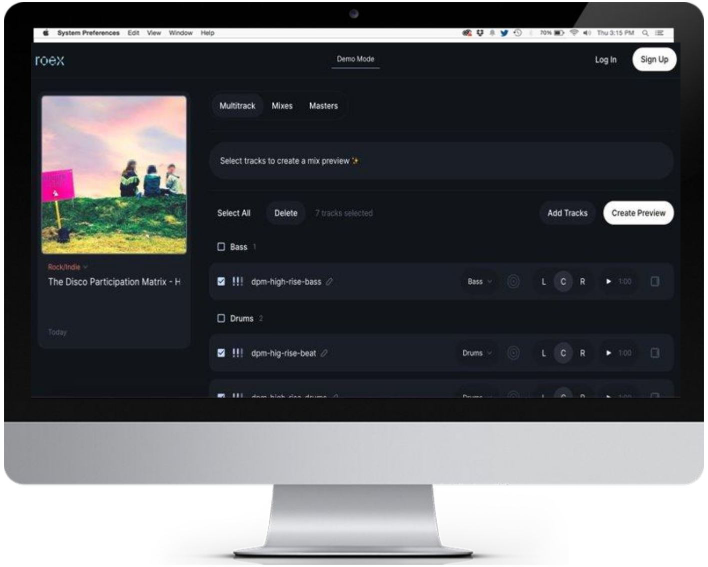
Our technology

- 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



Select tracks to create a mix preview ✨

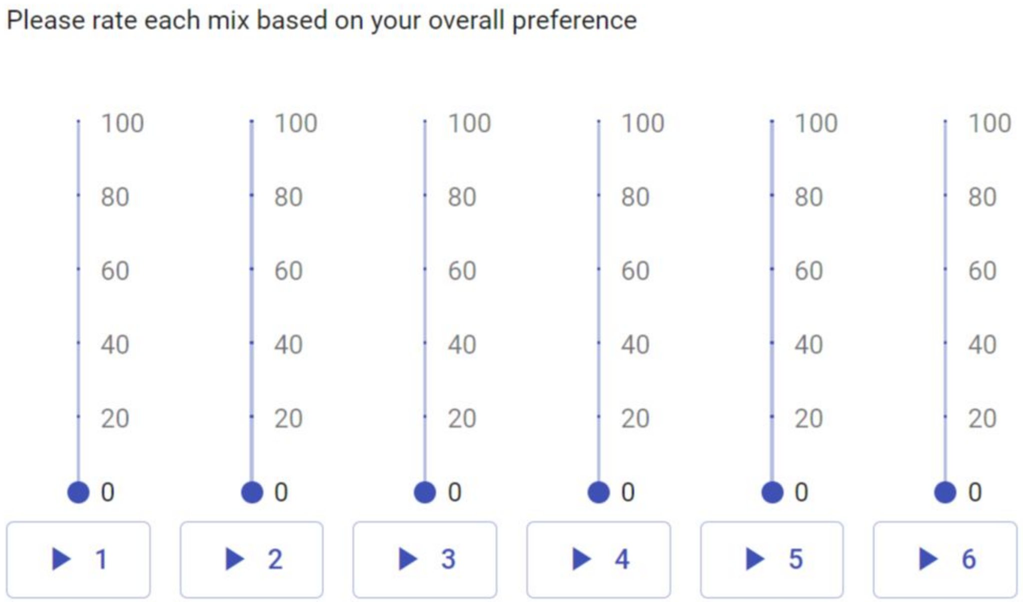
7 tracks selected

Roex Automix Demo

Provide some basic mix settings.

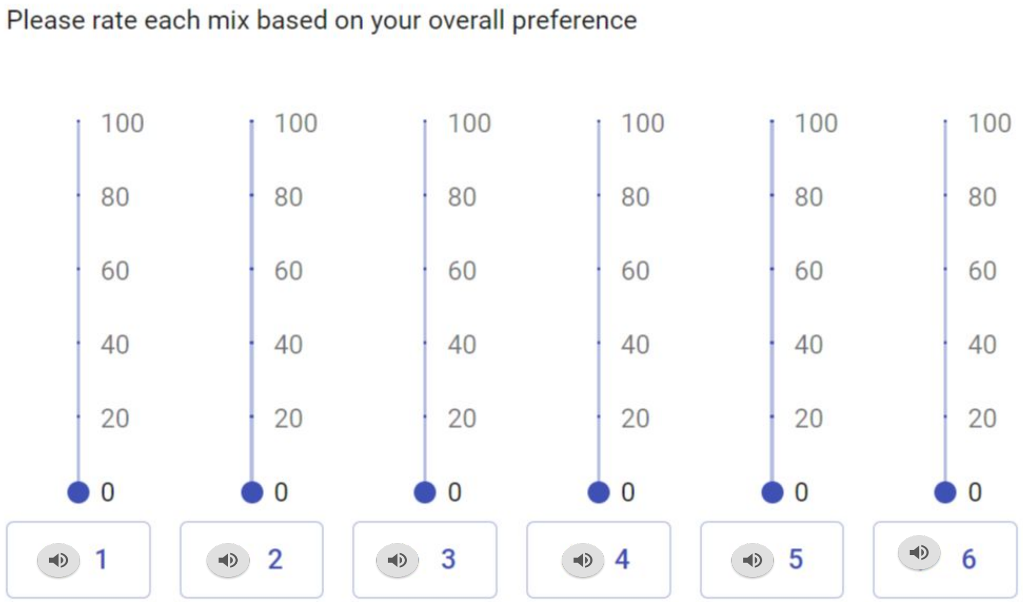
Demos

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

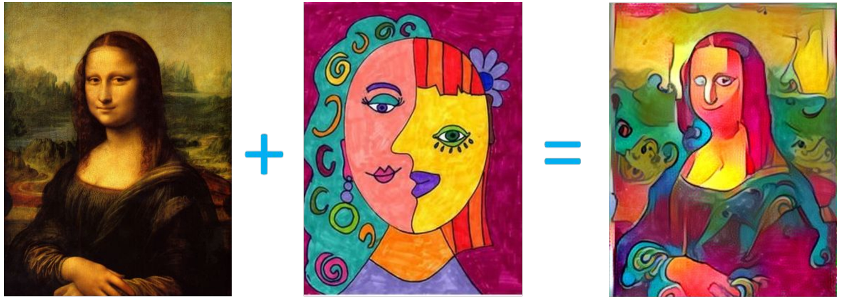
Generative AI

Functional art

Text prompt

Outpainting

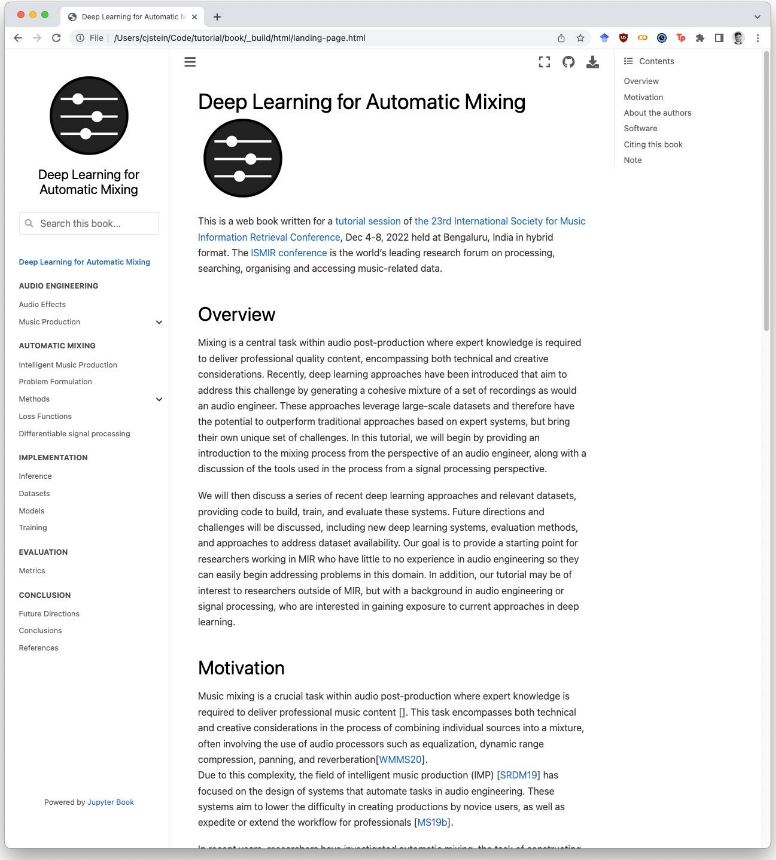
Style transfer



Resources

Book

https://dl4am.github.io/tutorial



Deep Learning for Automatic Mixing

This is a web book written for a tutorial session of the 23rd International Society for Music Information Retrieval Conference, Dec 4-8, 2022 held at Bengaluru, India in hybrid format. The ISMIR conference is the world's leading research forum on processing, searching, organising and accessing music-related data.

Overview

Mixing is a central task within audio post-production where expert knowledge is required to deliver professional quality content, encompassing both technical and creative considerations. Recently, deep learning approaches have been introduced that aim to address this challenge by generating a cohesive mixture of a set of recordings as would an audio engineer. These approaches leverage large-scale datasets and therefore have the potential to outperform traditional approaches based on expert systems, but bring their own unique set of challenges. In this tutorial, we will begin by providing an introduction to the mixing process from the perspective of an audio engineer, along with a discussion of the tools used in the process from a signal processing perspective.

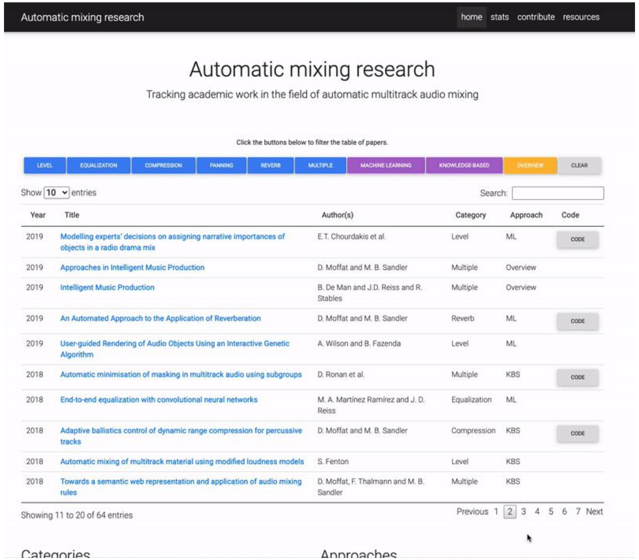
We will then discuss a series of recent deep learning approaches and relevant datasets, providing code to build, train, and evaluate these systems. Future directions and challenges will be discussed, including new deep learning systems, evaluation methods, and approaches to address dataset availability. Our goal is to provide a starting point for researchers working in MIR who have little to no experience in audio engineering so they can easily begin addressing problems in this domain. In addition, our tutorial may be of interest to researchers outside of MIR, but with a background in audio engineering or signal processing, who are interested in gaining exposure to current approaches in deep learning.

Motivation

Music mixing is a crucial task within audio post-production where expert knowledge is required to deliver professional music content []. This task encompasses both technical and creative considerations in the process of combining individual sources into a mixture, often involving the use of audio processors such as equalization, dynamic range compression, panning, and reverberation[WMMS20].
Due to this complexity, the field of intelligent music production (IMP) [SRDM19] has focused on the design of systems that automate tasks in audio engineering. These systems aim to lower the difficulty in creating productions by novice users, as well as expedite or extend the workflow for professionals [MS19b].

In recent years, researchers have investigated automatic mixing, the task of constructing

More works on automatic mixing research



Automatic mixing research

Tracking academic work in the field of automatic multitrack audio mixing

Click the buttons below to filter the table of papers.

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

Categories

Approaches

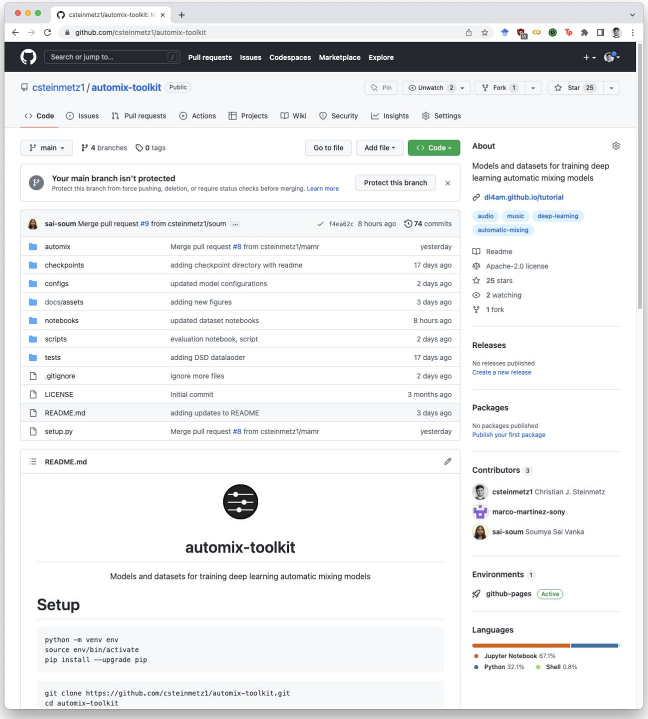
Searchable/filterable table of relevant papers and stats

https://csteinmetz1.github.io/AutomaticMixingPapers

automix-toolkit

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

★ Star   Star it on GitHub



Your main branch isn't protected
Protect this branch from force pushing, deletion, or require status checks before merging. Learn more

sai-soum Merge pull request #9 from csteinmetz1/soum ... ✓ f4ea62c 8 hours ago 74 commits

automix	Merge pull request #8 from csteinmetz1/mamr	yesterday
checkpoints	adding checkpoint directory with readme	17 days ago
configs	updated model configurations	2 days ago
docs/assets	adding new figures	3 days ago
notebooks	updated dataset notebooks	8 hours ago
scripts	evaluation notebook, script	2 days ago
tests	adding DSD datalaoder	17 days ago
.gitignore	ignore more files	2 days ago
LICENSE	Initial commit	3 months ago
README.md	adding updates to README	3 days ago
setup.py	Merge pull request #8 from csteinmetz1/mamr	yesterday

README.md

automix-toolkit

Models and datasets for training deep learning automatic mixing models

Setup

```
python -m venv env
source env/bin/activate
pip install --upgrade pip
```

```
git clone https://github.com/csteinmetz1/automix-toolkit.git
cd automix-toolkit
```

About

Models and datasets for training deep learning automatic mixing models

dl4am.github.io/tutorial

audio music deep-learning automatic-mixing

Readme
Apache-2.0 license
25 stars
2 watching
1 fork

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 3

csteinmetz1 Christian J. Steinmetz
marco-martinez-sony
sai-soum Soumya Sai Vanka

Environments 1

github-pages Active

Languages

Jupyter Notebook 67.1%
Python 32.1%
Shell 0.8%

Thank You

Questions?