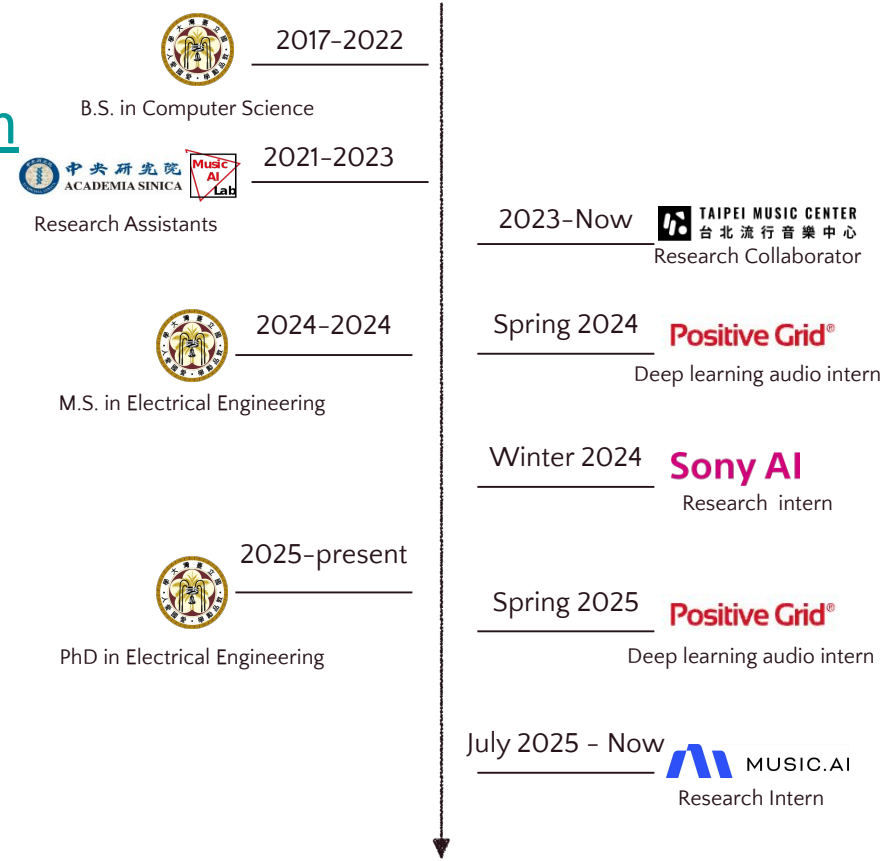
AI Mixing

Yen-Tung Yeh
National Taiwan University

About Me

- Yen-Tung (Arthur) Yeh
- Email: ytsrt66589@gmail.com
- Research Topics:
  - Audio Effects
  - Guitar Tone
  - Mixing/Mastering

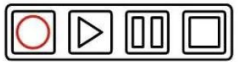


Introduction

What makes Spotify music sound professional while your iPhone recordings don't?

Audio Production

STAGES OF AUDIO ENGINEERING

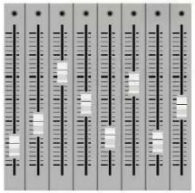


TRACKING

RECORDING OF DIFFERENT INSTRUMENTS, SOUNDS AND VOCALS

EDITING

EDITING AND CORRECTING THE RECORDINGS



MIXING

ART OF PROCESS OF COMBINING ALL THE INSTRUMENTS AND MIXING THE SOUNDS

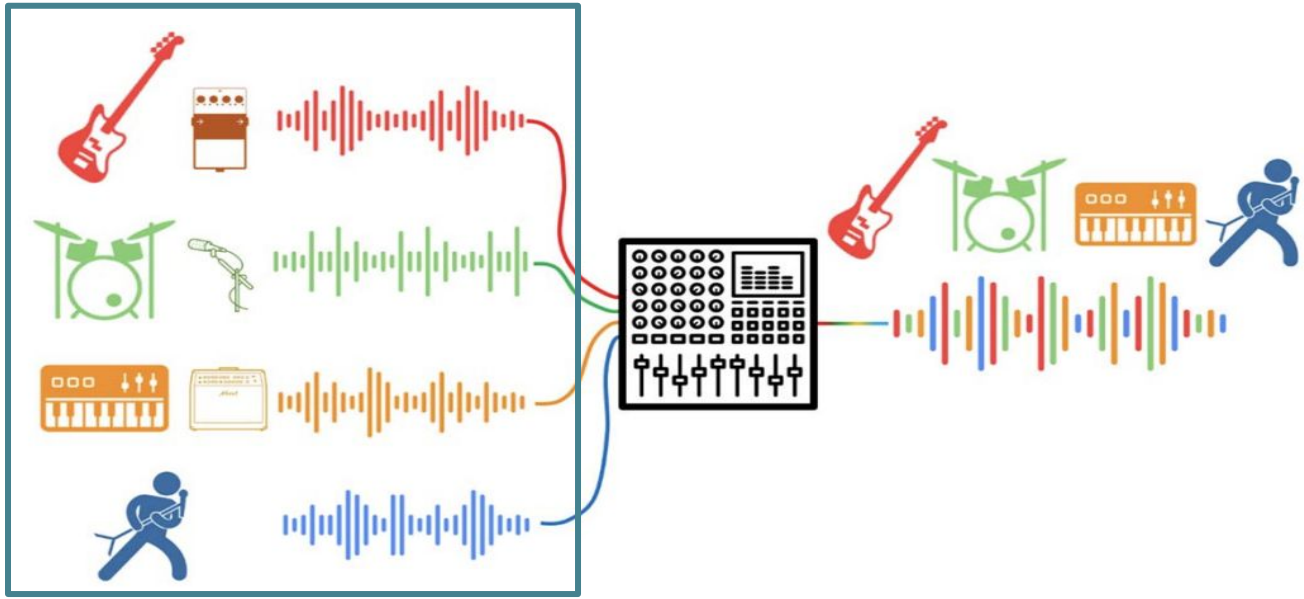
MASTERING

PROCESS OF MAKING ALL THE SONGS SOUND COHERENT

Mixing

- **Audio mixing** is the process of blending multitrack recordings by using chains of audio effects (Gain, Panning, EQ, Reverb, Compression, etc.)

Multitrack Recordings



Audio Effects (AFx)

- Audio effects are processing tools that intentionally change how recorded sound behaves
- Input audio -> output audio

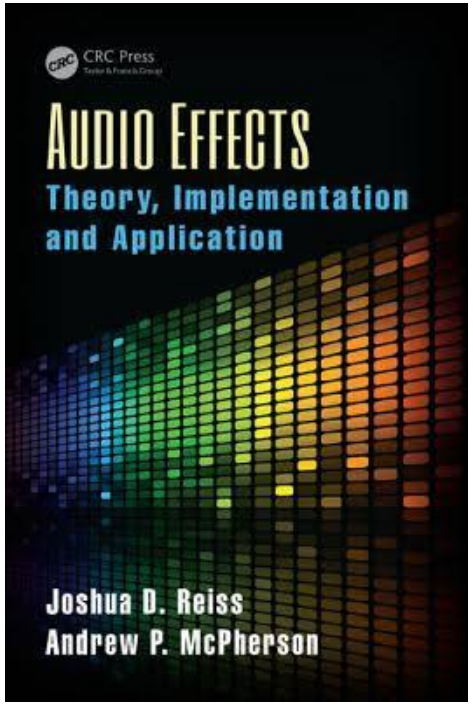
Audio Effects (AFx)

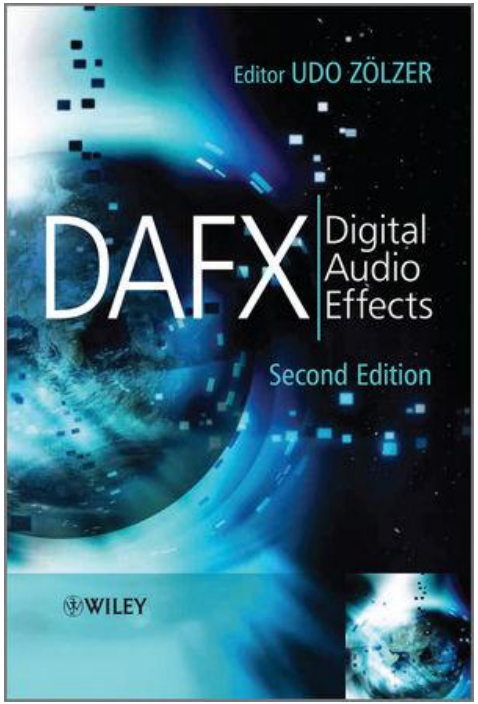
Mixing & Audio Effects

- Highly correlated topics. Mixing problem is often too difficult to analyze but audio effects problem is much clear for research. Most of good researchers in mixing field are also an expert at audio effects topic

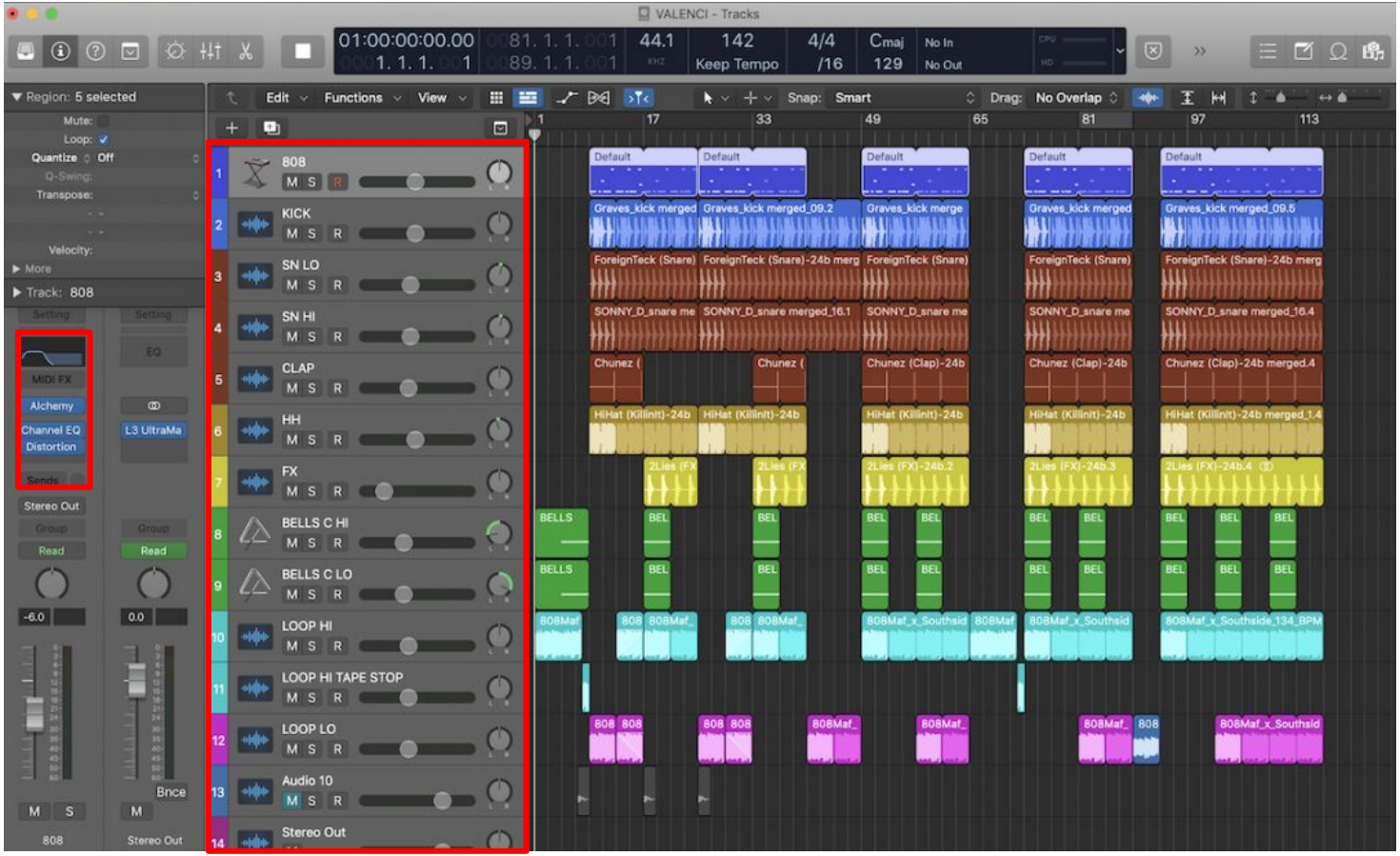
Audio Effects Area is a Big Research Field

- Conferences: International Conference on Digital Audio Effects (DAFx)
- Books:
  - link1
  - link2





Mixing in DAW



Mixing in Industry

AUDIO REPAIR

RX 10

Industry-standard audio repair tool used on movies and TV shows to restore damaged, noisy audio to pristine condition.

Shop now

Learn More

MIXING

Neutron 4

Mix smarter and faster while staying in your flow. Eight professional plug-ins combine to create your modern and intelligent mixing experience.

Shop Now

MASTERING

Ozone 11

Harness the power of Ozone 11, the industry-standard mastering suite. Featuring new processing like Clarity, Stem Focus, and Transient/Sustain for professional sound with ease and precision.

VOCAL PRODUCTION

Nectar 4

Get your vocals to sit in the mix with the most sophisticated set of tools for vocal production.

Shop Now

Learn More

Mixing is Difficult

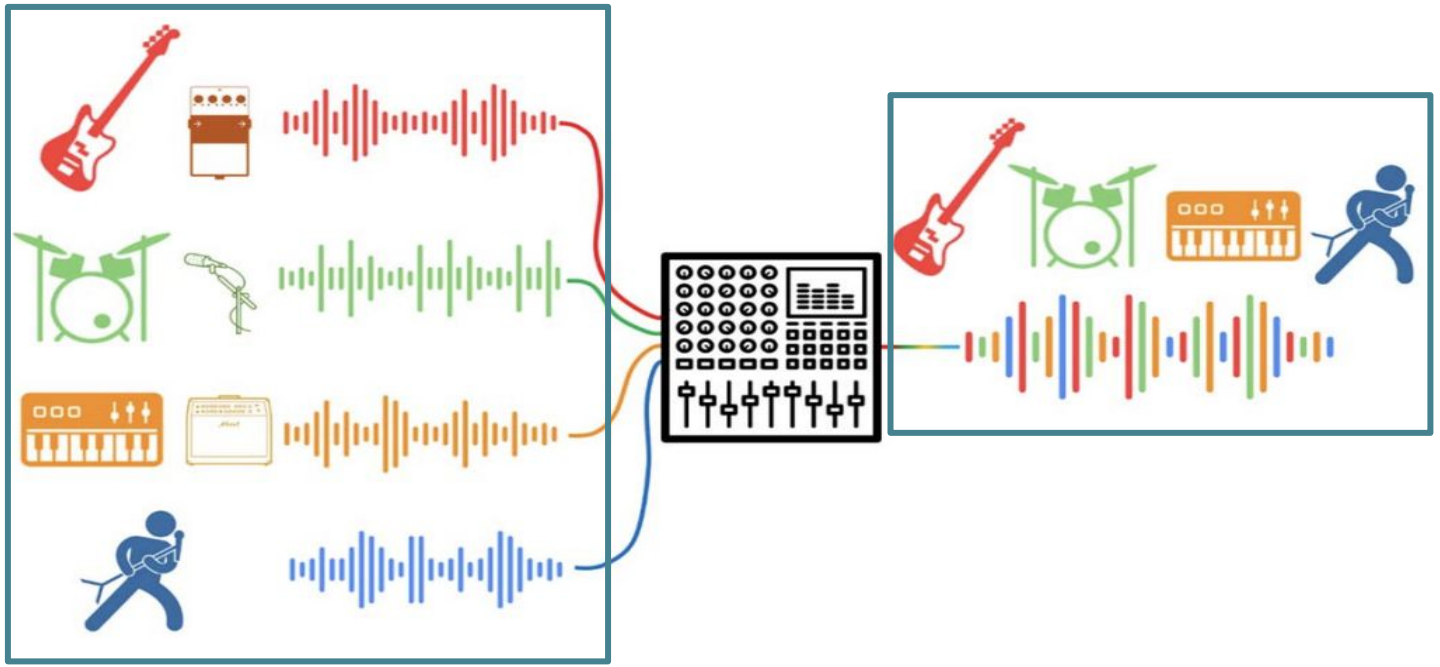
- High demand of mixing technique
  - Requires good listening skills
  - Requires creativity
  - Requires understanding of music (instruments, genre)
  - Requires skills of handling large number of input tracks
- Time consuming

Can we let machine intelligently help us mix audio?

Background

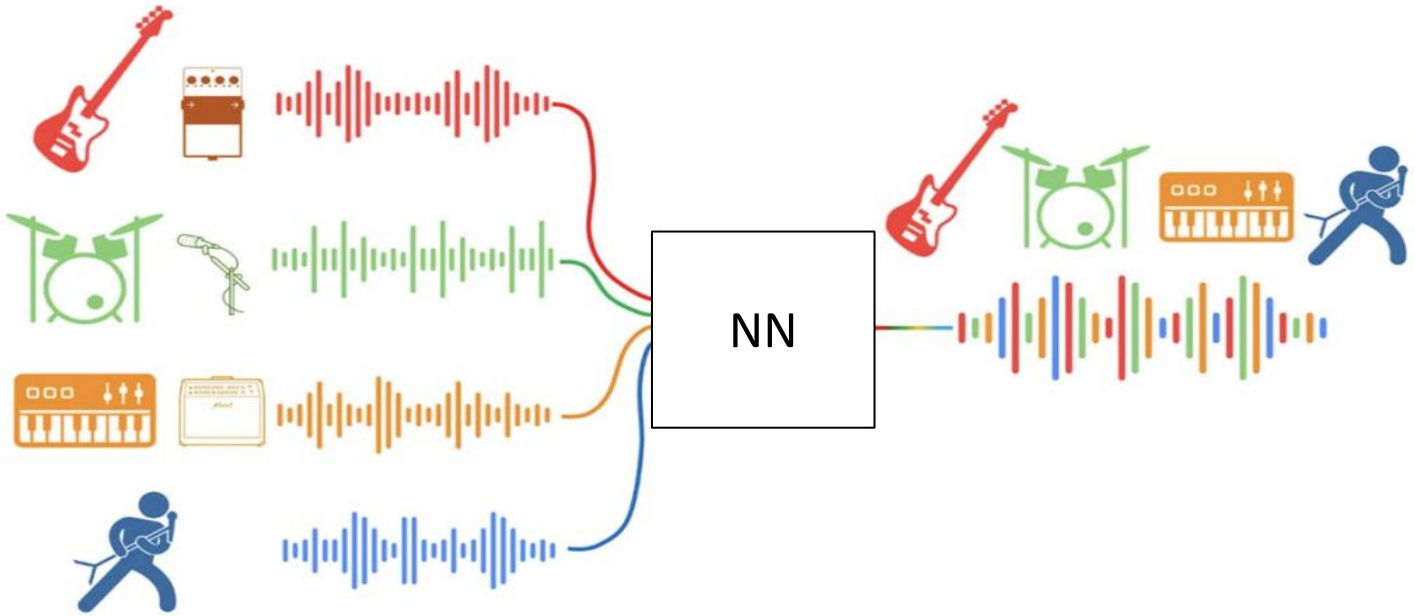
High-Level

- Audio2Audio Transformation



Problem Formulation of Mixing

- Create a final good mixture by given multitrack audio

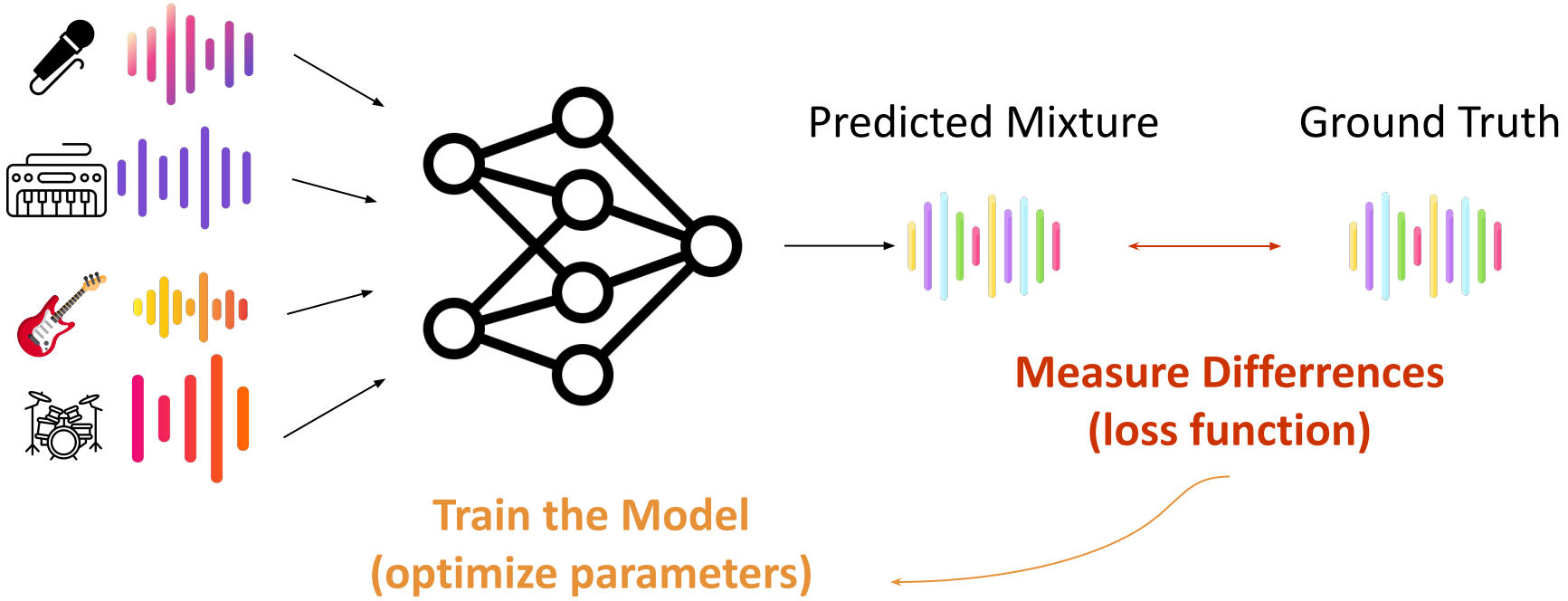


In Short

- Mixing is about learning an **audio transformation**
  - Input: Multitrack audio
  - Output: Mixture (audio)

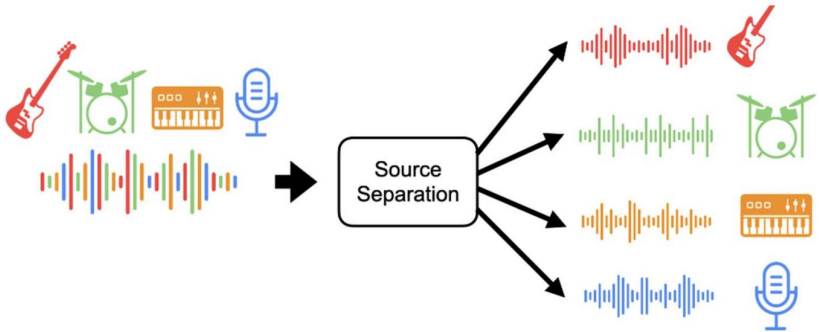
Automatic Mixing

Deep Learning Approach

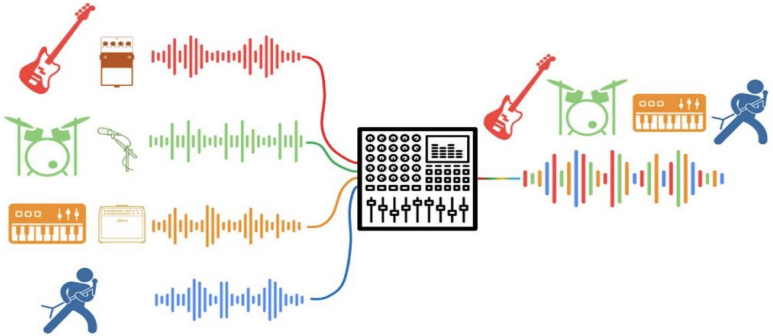


Rethinking: Mixing & Source Separation

- Source separation



- Mixing



Mixing = Inverse Source Separation

- Source separation:
  - Input mixture, output multitrack audio
- Mixing
  - Input multitrack audio, output mixture

Multitrack Audio as Input

- Sol: Concatenate the multitrack audio along channel
  - Single Stereo Input: [Ch, Length]
  - N Stereo Input: [Ch * N, Length]

MixWaveUNet

Input (Stack of Multitracks)

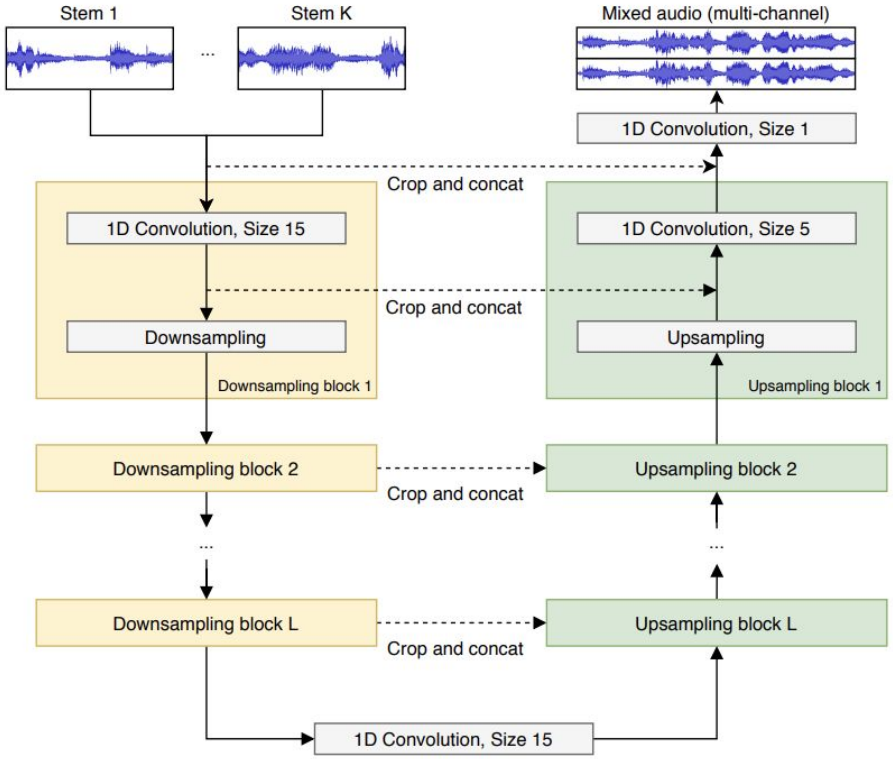


Fig. 1. Block diagram of the adapted *Wave-U-Net* for automatically mixing K stems using L layers

Output (Mixture)

Martinez Ramirez, M., Daniel Stoller, and David Moffat. "A deep learning approach to intelligent drum mixing with the wave-u-net." Audio Engineering Society, 2021.

MixWaveUNet

- Directly learns the audio transformation
- High-quality audio output

However, this model suffers from several limitations

Limitation: Stack of Multitrack Input

- Stack of multitrack is problematic in two senses:
  - Fixed number of inputs
  - Fixed order of inputs
- The fixed scenario is because the model requires the static setting. For example: model expect input channel as 8. Then the input audio channel should always be 8.

Fixed Number of Inputs

- If the model trained on 4 input tracks, then the input should always be 4 tracks.
- What if you only want to mix 3 tracks?
  - Add a zero value track as the dummy
- What if you want to mix 8 tracks?
  - :(

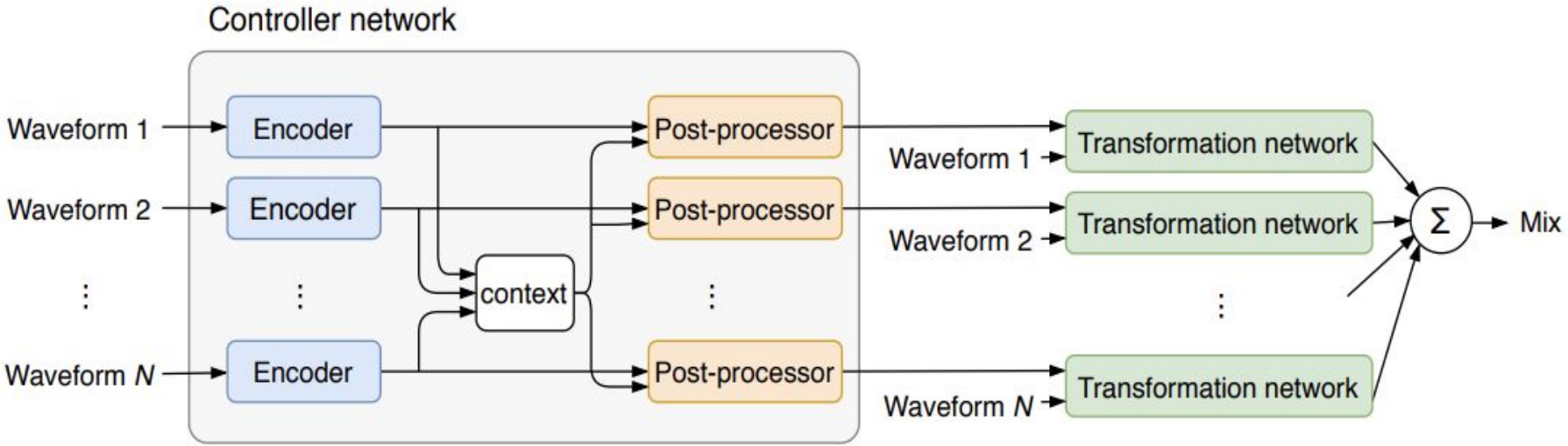
Fixed order of inputs

- If the model trained with the order
  - First channel is drums
  - Second channel is vocals
  - Third channel is bass
  - Fourth channel is guitar
- When inference, you should **follow this exact order**
- What if randomly permuting the order
  - The mixing behavior will be unexpected, leading to **worse result**

Controllability & Interpretability

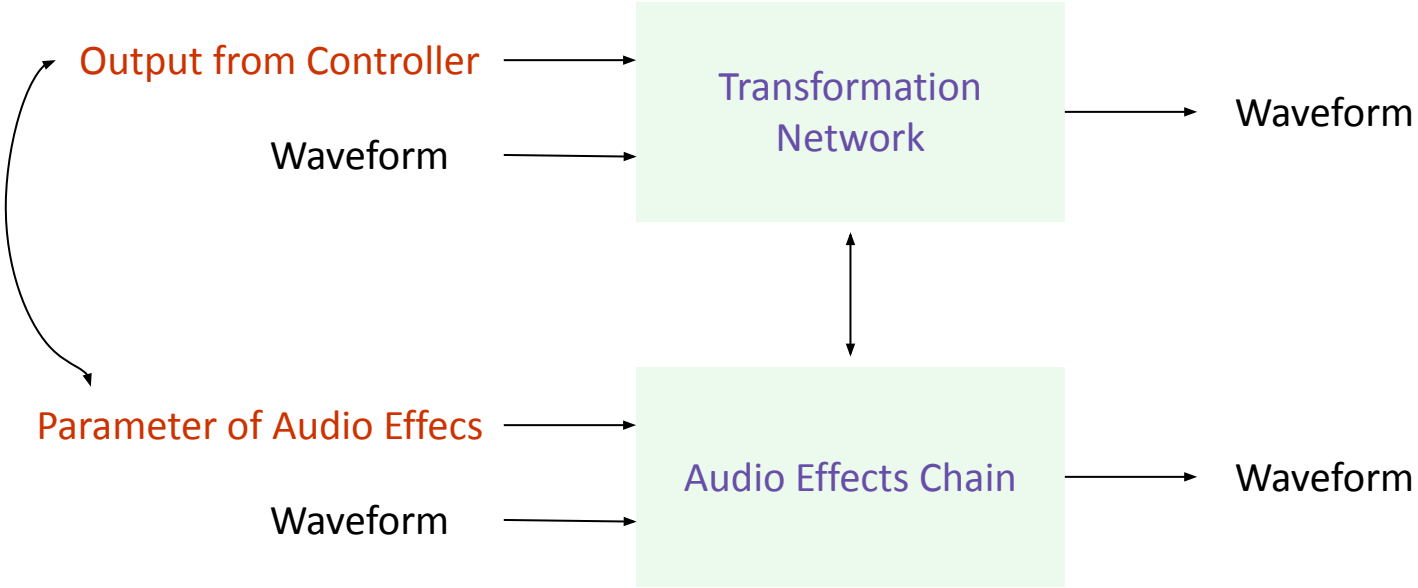
- In many real-world cases, users may want to understand **how the mixing is done** (interpretability)
- Users may want to **control the mixing model** (controllability)
- MixWaveUNet is a pure black-box model which is not controllabale and lack of interpretability

DMC = Controller + Transformation network



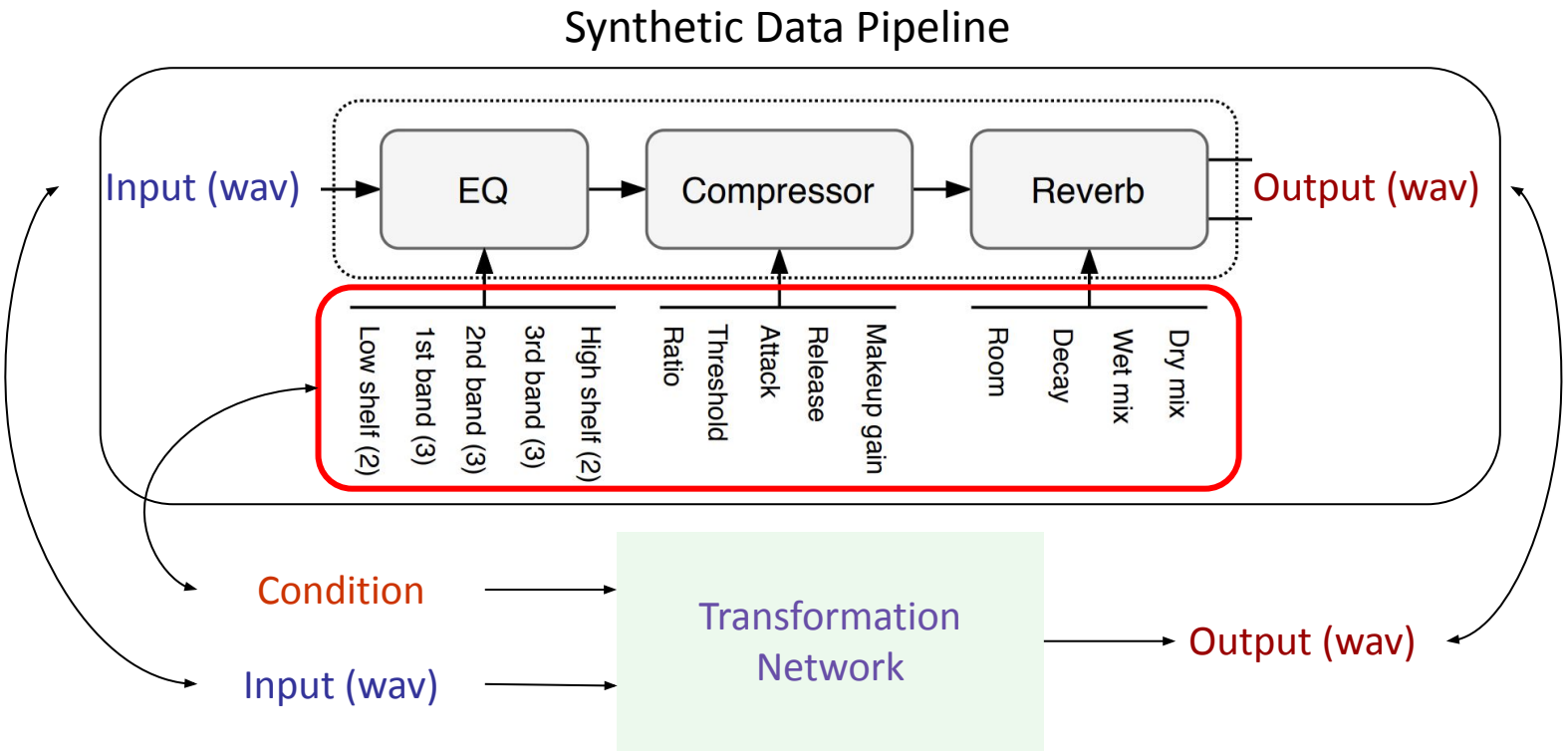
Steinmetz, C. J., Pons, J., Pascual, S., & Serra, J. (2021, June). Automatic multitrack mixing with a differentiable mixing console of neural audio effects. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 71-75). IEEE.

Transformation Network



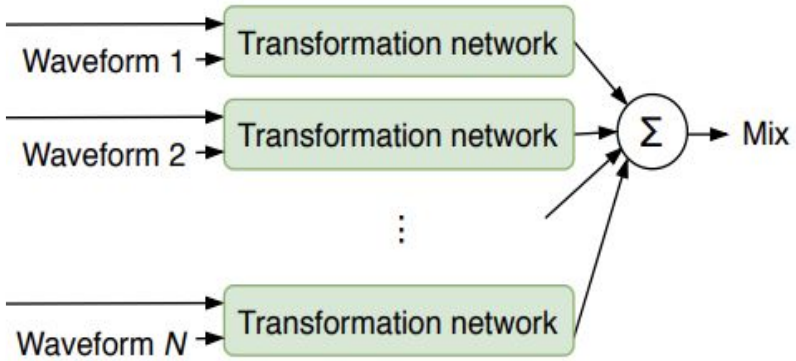
How do we guarantee the Transformation Network will have the same behavior as audio effects chain?

Audio Effects Emulation (Pre-trained)

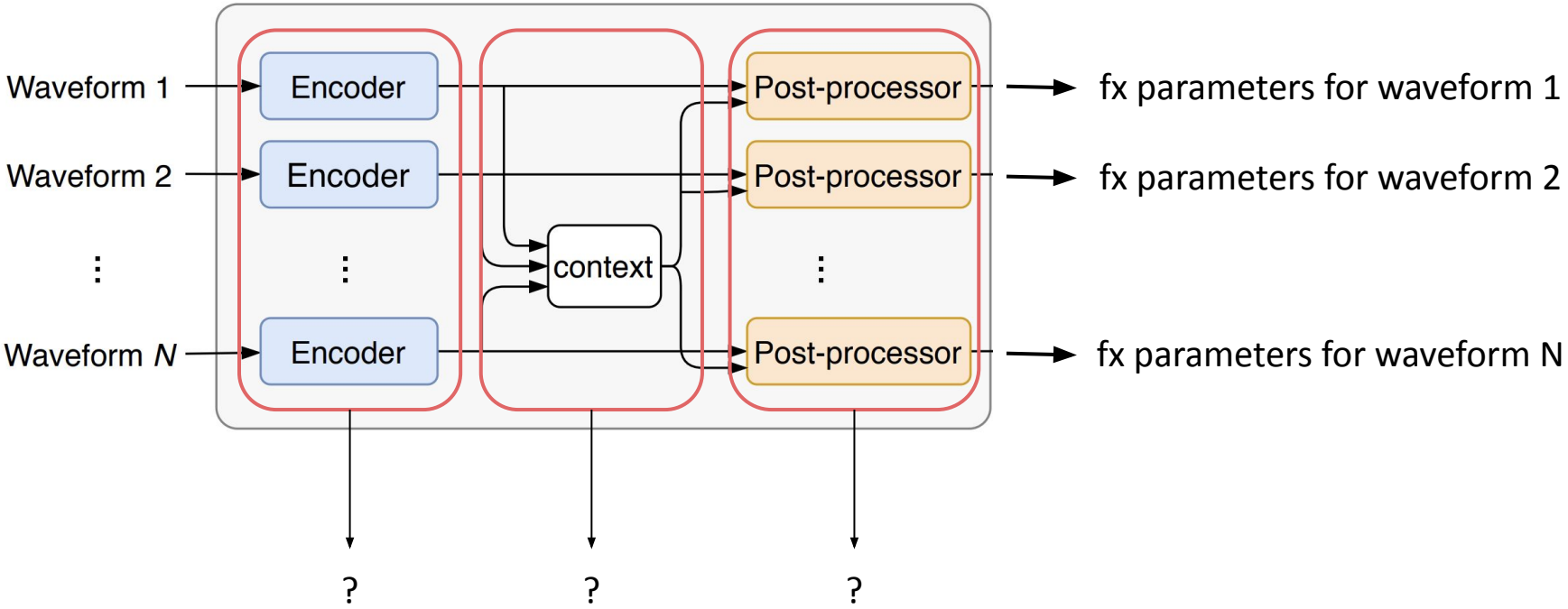


Transformation Network in DMC

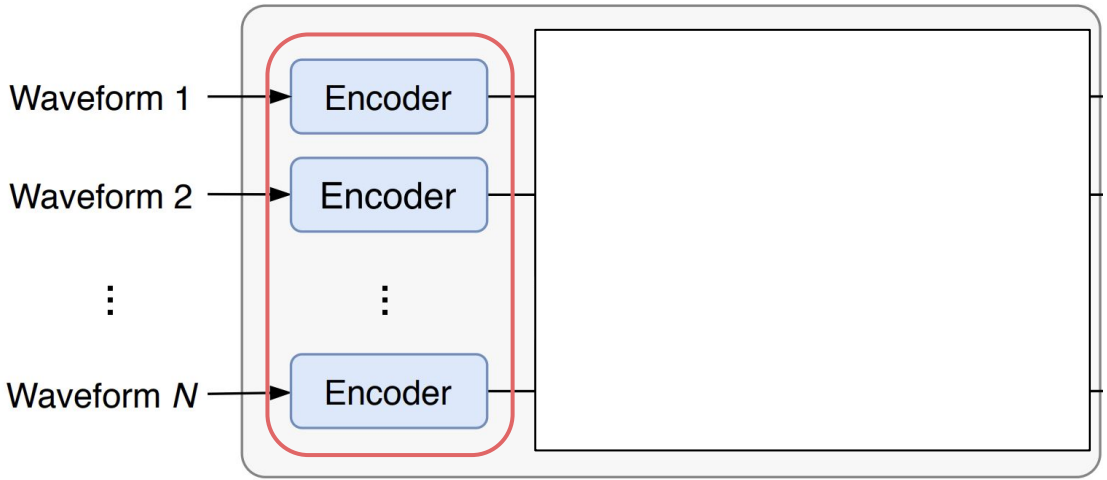
- Purpose: applying audio effects to the input waveform
- After processing all inputs, get the summation of all stems as the final mixture
- Analogy to the real-world mixing console



Controller

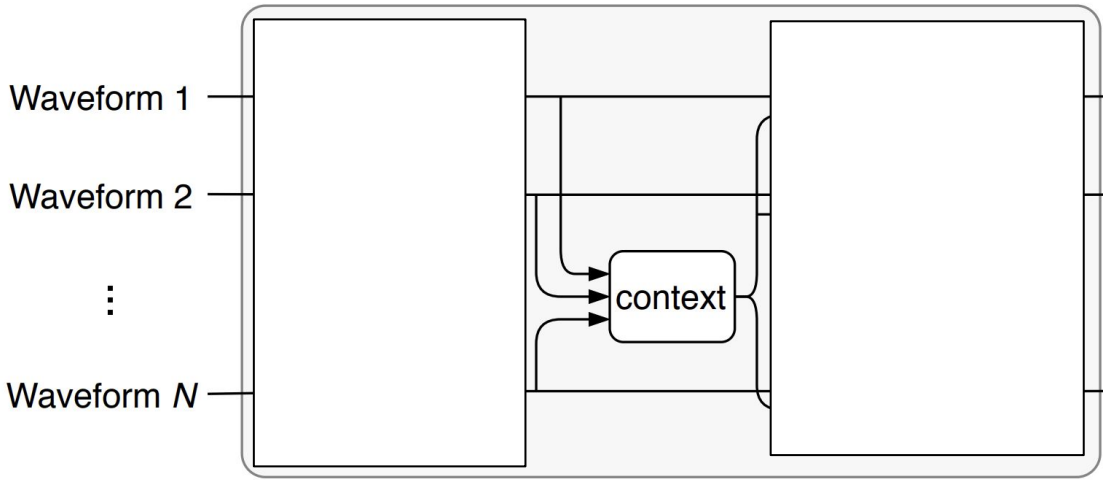

Waveform 1
Waveform 2
Waveform N
Encoder
Encoder
Encoder
context
Post-processor
Post-processor
Post-processor
fx parameters for waveform 1
fx parameters for waveform 2
fx parameters for waveform N
?
?
?

Controller



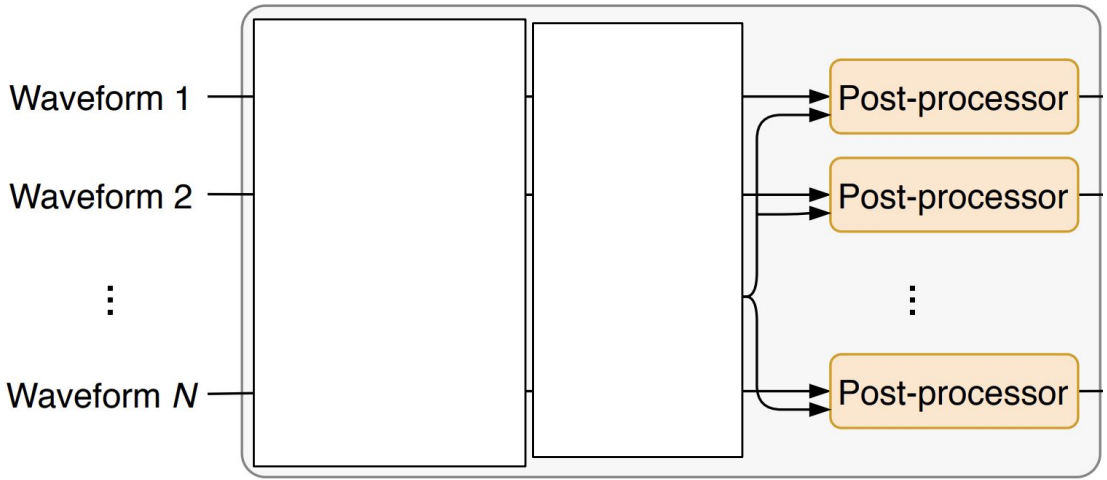
Purpose: To get the semantic embedding of each audio input

Controller



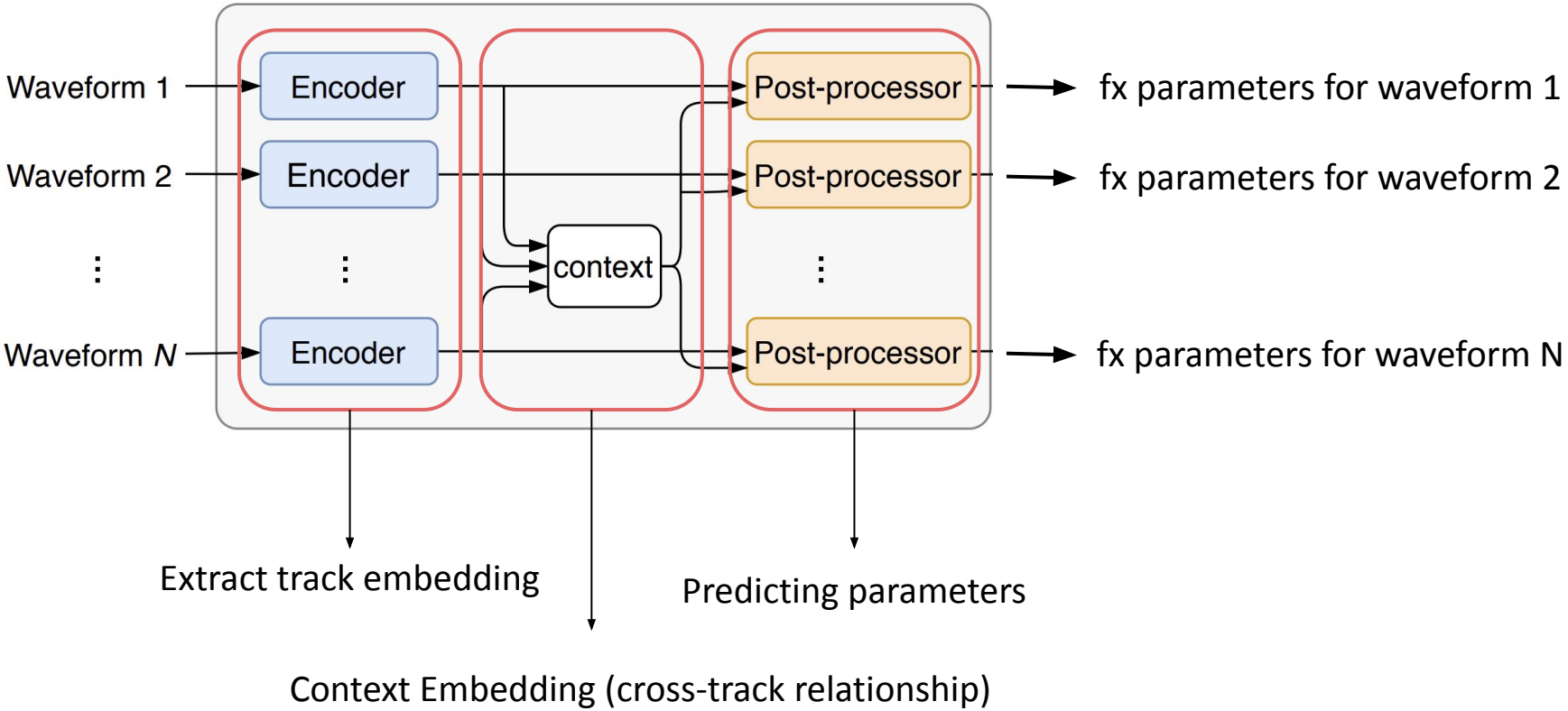
Purpose: To include the cross-track relationship. Context is an embedding, creating from the mean embedding of multitrack input

Controller

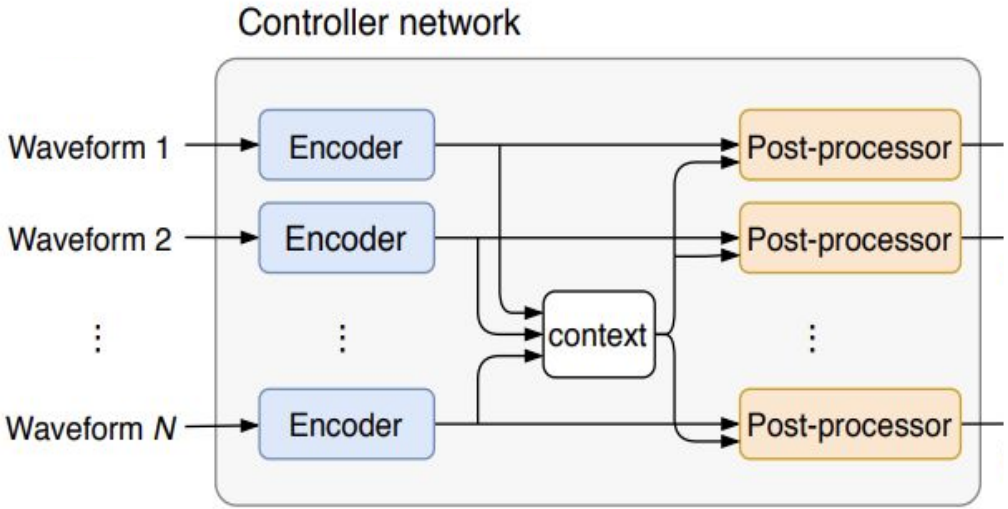


Purpose: Given **track embedding** and **context embedding**, predicting the audio effects parameters of each waveform

Controller

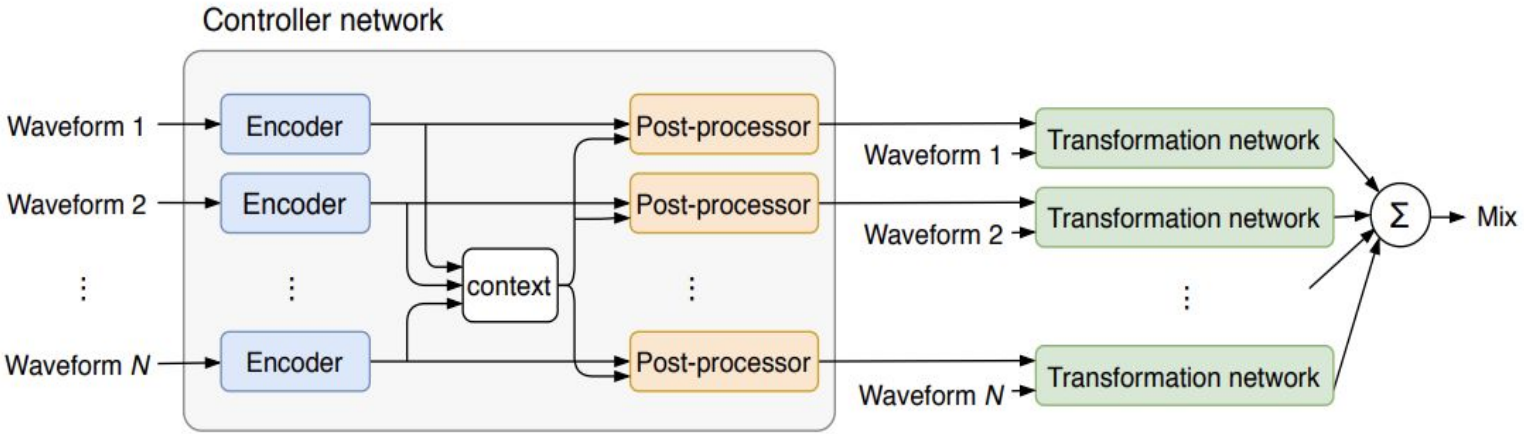

Waveform 1
Waveform 2
Waveform N
Encoder
Encoder
Encoder
context
Post-processor
Post-processor
Post-processor
fx parameters for waveform 1
fx parameters for waveform 2
fx parameters for waveform N
Extract track embedding
Predicting parameters
Context Embedding (cross-track relationship)

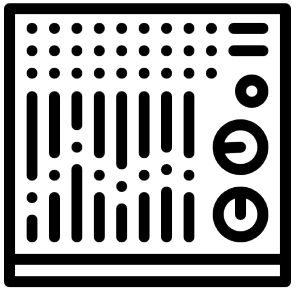
Controller in DMC



- Purpose: predicting audio effects parameters
- Analogy to the mixing engineer

DMC = Controller + Transformation network





DMC

- Handling arbitrary input tracks with arbitrary order
  - Controller
- Controllable and interpretable
  - Due to the audio effects parameters and transformation network

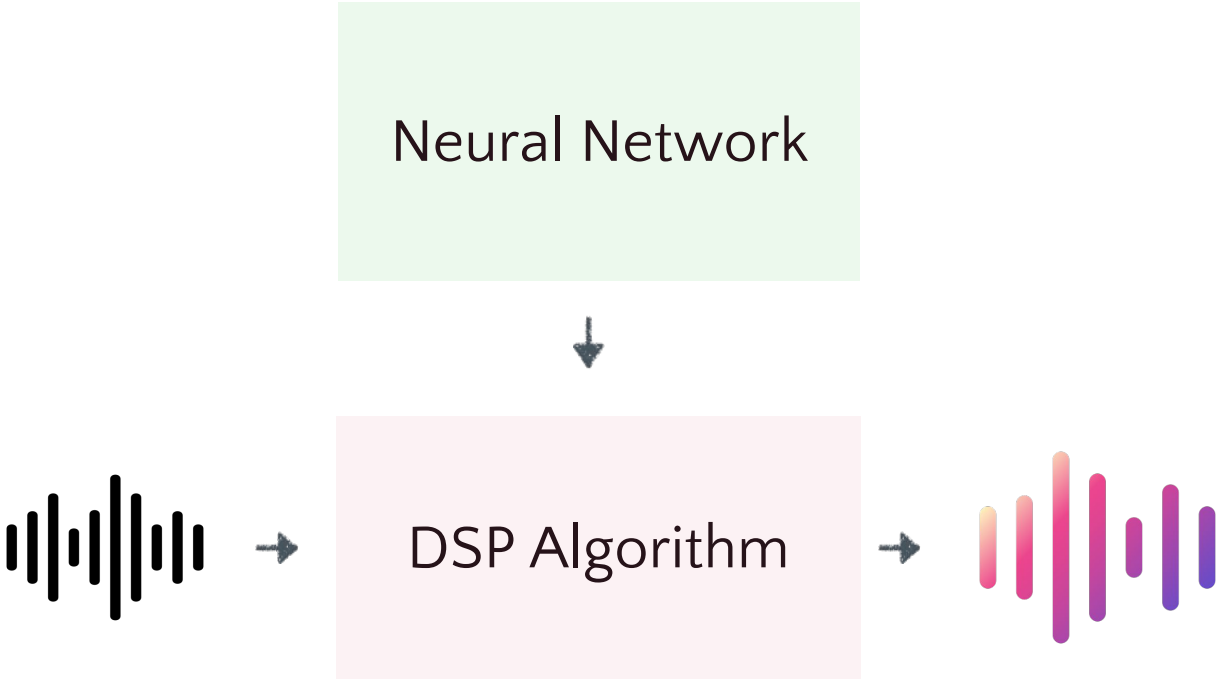
Insights from DMC

- Using NN to control audio effects processors
  - ex. Using Controller to control Transformation Network
- This brings up the important concept about how we can leverage the well-developed signal processing approach to deep learning framework

Deep Learning = Throwing away DSP tools?

Differentiable Signal Processing

- Using Neural Network to control DSP algorithms



Differentiable and Efficiency

- To make sure the backpropagation, the DSP algorithm should be differentiable
- To make sure the efficient usage of GPU, we need to optimize the algorithm for faster processing

Why Leveraging DSP?

- DSP algorithm makes sure the high-quality output audio
  - Those DSP algorithms are well-developed for over decades
- It aligns with the logic of current workflow and audio plugin
  - Most of underlying logic of audio effects are based on DSP

However…..

- MixWaveUnet & DMC have not yet achieved the level of professional audio engineers mixes
- Hypothesis:
  - Bottleneck of performance can be resolved with a large dataset

Therefore, we need to introduce the current dataset for mixing first.

Mixing Dataset

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

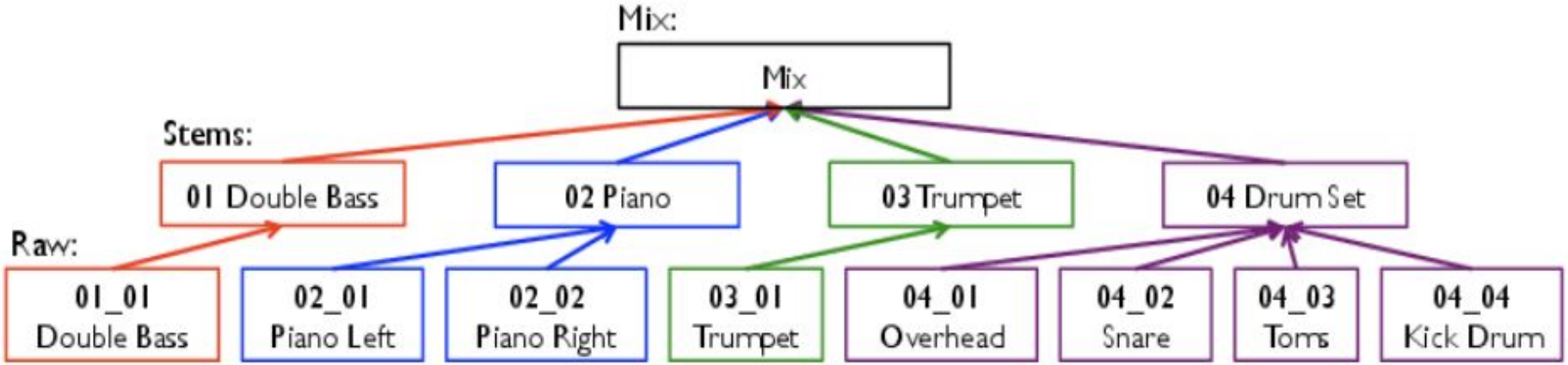
Dataset	Size(Hrs)	no. of Songs	no. of Instrument Category	no. of tracks	Type	Usage Permissions	Other info	Remarks
MedleyDB	7.2	122	82	1-26	Multitrack, Wav	Open	44.1KHz, 16 bit, stereo	-
ENST Drums	1.25	-	1	8	Drums, Wav/AVI	Limited	44.1KHz, 16 bit, stereo	Drums only dataset
Cambridge Multitrack	>3	>50	>5	5-70	Multitrack, Wav	open	44.1KHz, 16/24 bit, Stereo	Not time alligned, recordings for all the songs are not uniform

Very limited data for music mixing !

Source Separation Dataset?

- Source Separation dataset assume the final mixture is just the summation of each stem
  - Which means no hope to learn mixing !
  - The underlying of multitrack input to mixture is just pure summation
- BTW, a great example of understanding your data is extremely important for deep learning project

Dry Track / Wet Track (Stem) / Mixture

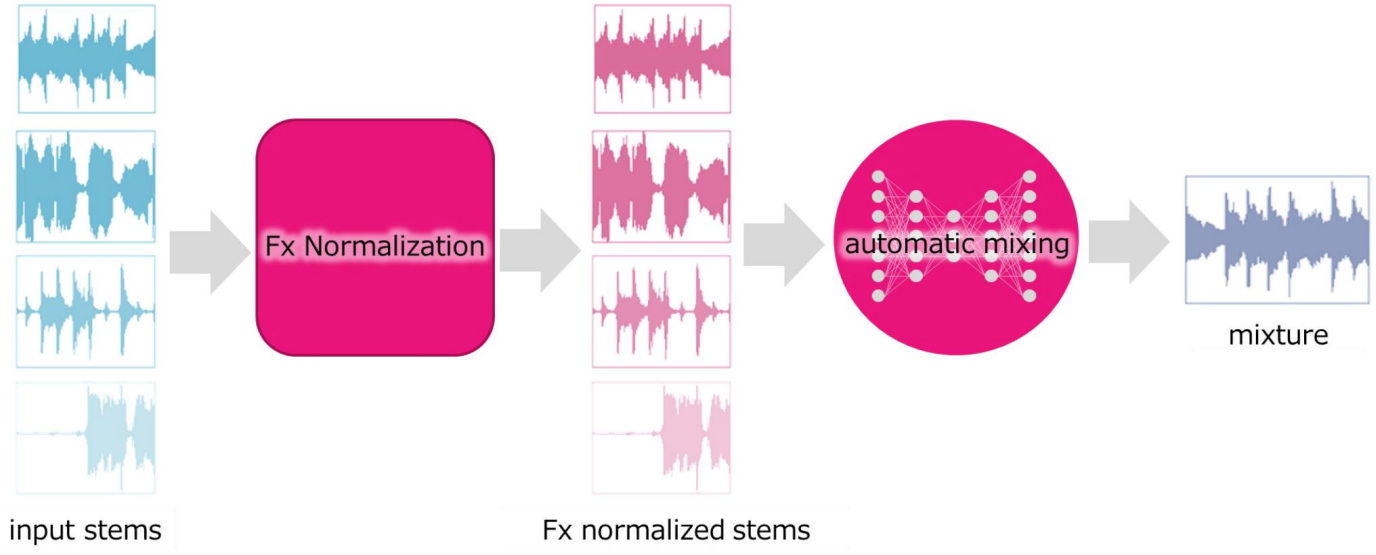


Addressing Data Bottleneck

- Challenge: Collecting dry multitrack input with corresponding mixes is difficult
  - Time-consuming
  - Copyright issue
  - Expensive
- Can we use wet multitrack music data and repurpose it to train models for mixing?

Audio Effects (Fx) Normalization

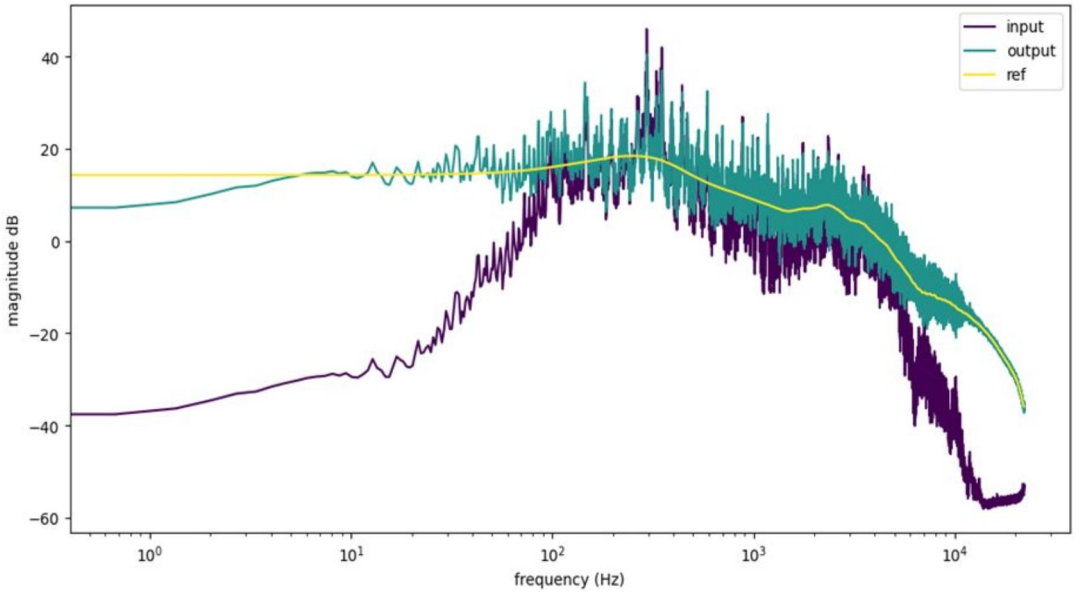
- A normalization technique that design for audio effects
- Motivation: to get the unprocessed (normalized) multitrack



Martínez-Ramírez, Marco A., et al. "Automatic music mixing with deep learning and out-of-domain data." in Proc. of ISMIR 2022

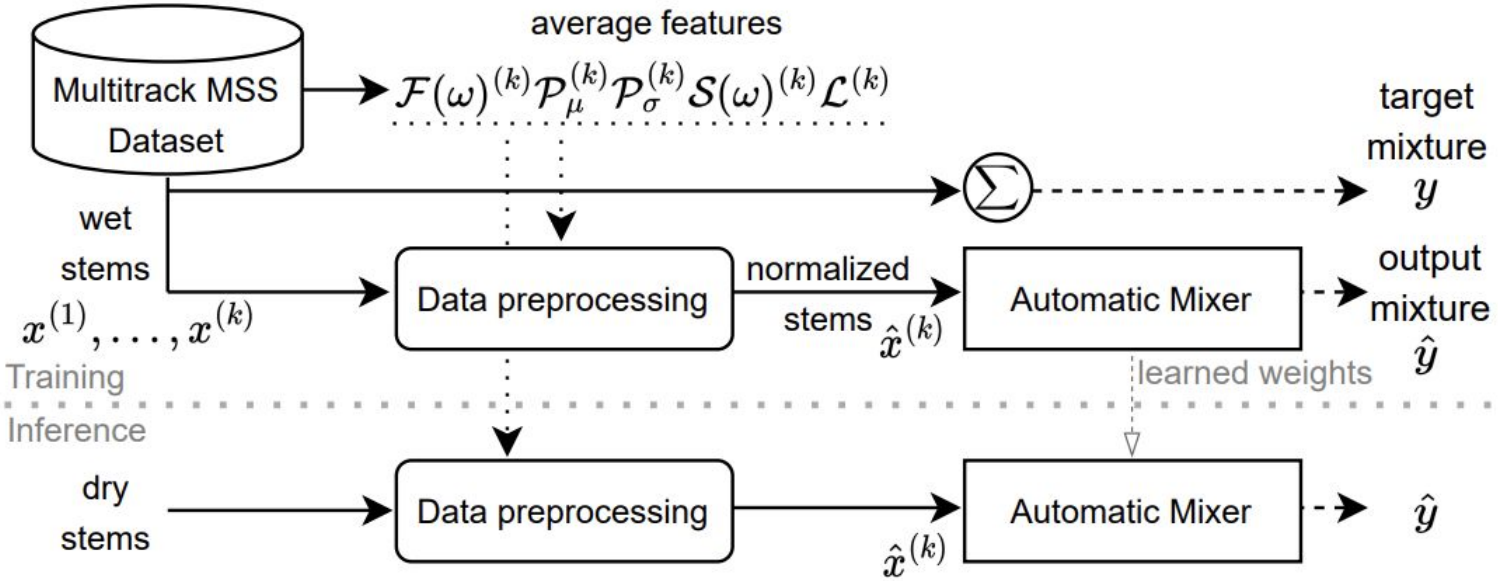
Core Idea

- Collecting the average audio effects features for each type of stems
- Map each audio to the average audio effects features
- Ex. EQ Normalization



Learning Objective of Model

- Model learns how to “denormalize” multitrack input and blend it to the final mixture

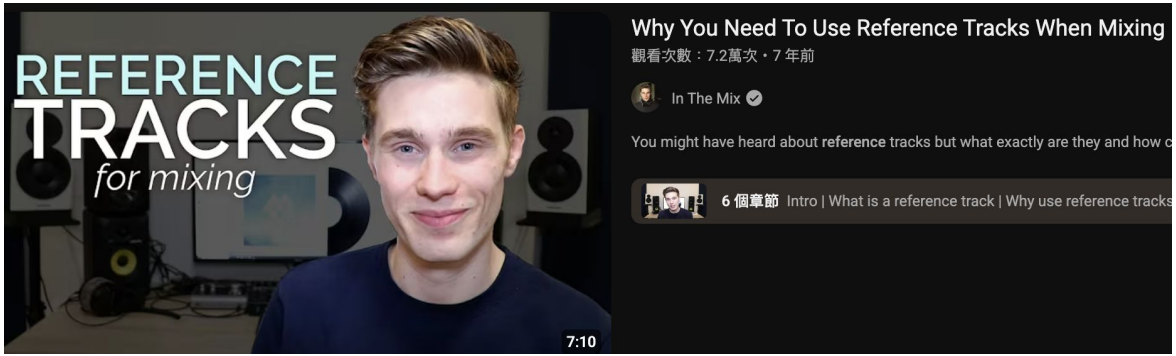


Context-Aware Mixing & Effects Representation for Mixing

Rethinking Mixing

- Previous models focus on creating good mixes
- However, what are good mixes? and which criteria does the mixing process model follows?
  - Subjective concept, highly depends on individuals
- This brings up an important concept:

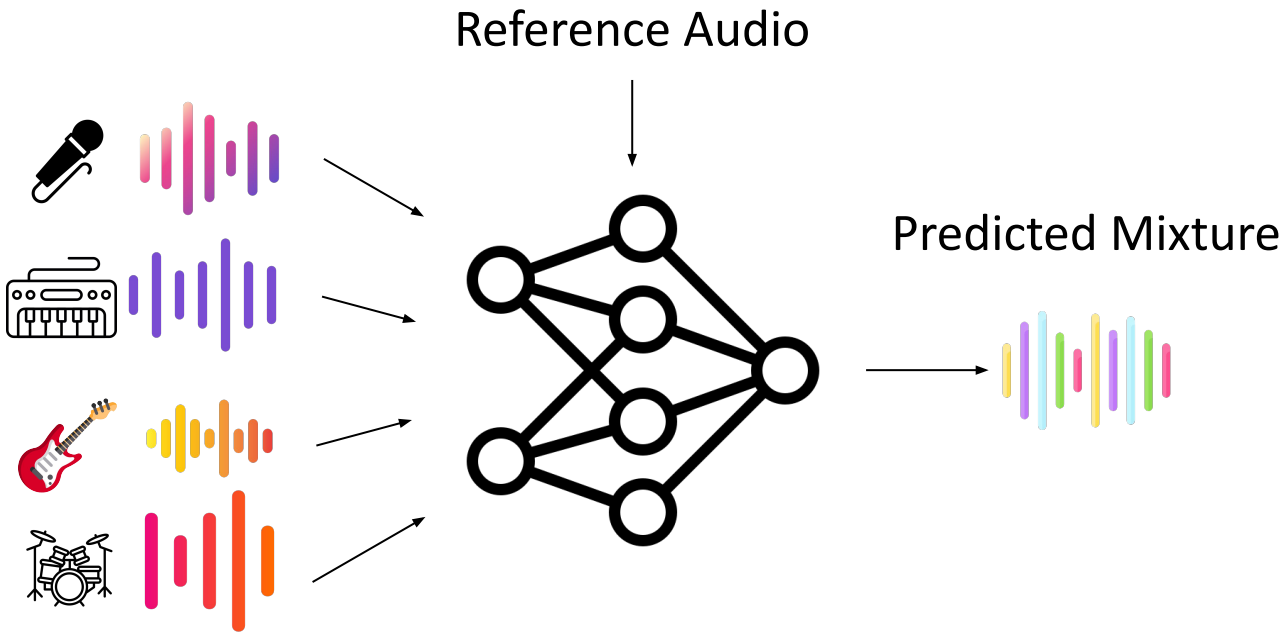
Reference Audio



Reference Audio

- Reference Audio is commonly used by real-world professional mixing engineer
- The role of the reference audio: knowing the mixing way done in this mixture
- One of the interpretation: knowing how audio effects are applied to the multitrack input

Problem Formulation of Mixing Style Transfer



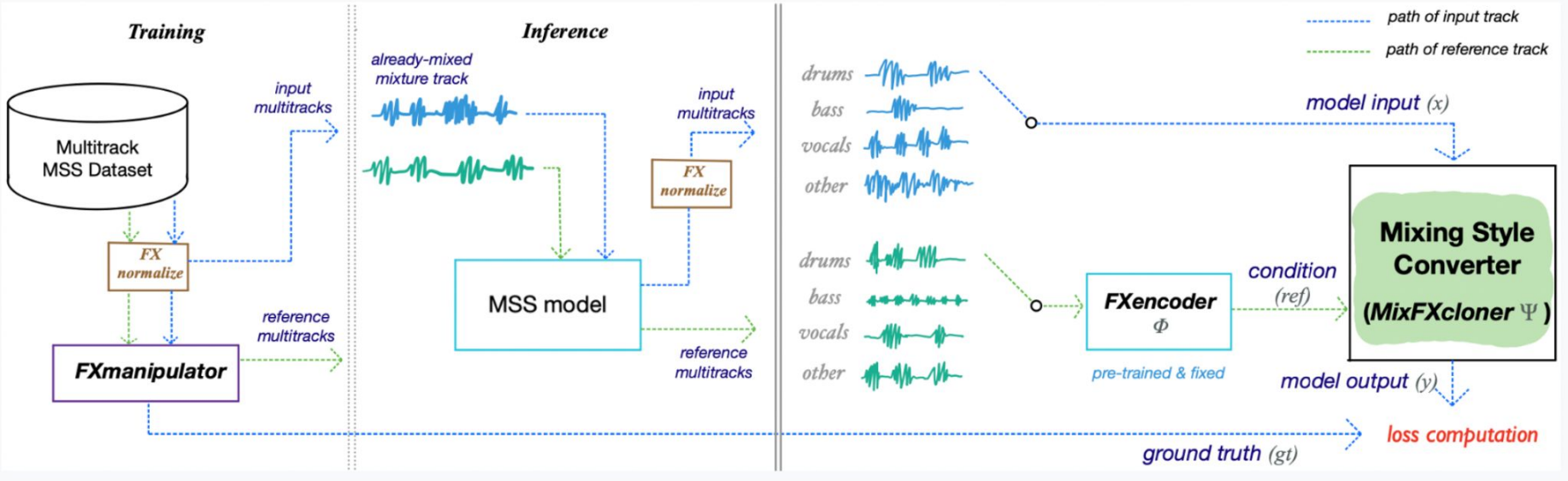
Given multitrack input, the model should mix the input according to the reference audio

Mixing Style Representation

- Trained jointly with mixing style transfer system
  - Diff-MST [1]
- Pre-trained by specific design
  - Fx-Encoder [2]
  - Fx-Encoder++ [3]

[1] Vanka, Soumya Sai, et al. "Diff-mst: Differentiable mixing style transfer." in Proc. of ISMIR 2024.
[2] Yeh, Yen-Tung, et al. "Fx-Encoder++: Extracting Instrument-Wise Audio Effects Representations from Mixtures." in Proc. of ISMIR 2025.
[3] Koo, J., Martínez-Ramírez, M. A., Liao, W. H., Uhlich, S., Lee, K., & Mitsufuji, Y. (2023, June). Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

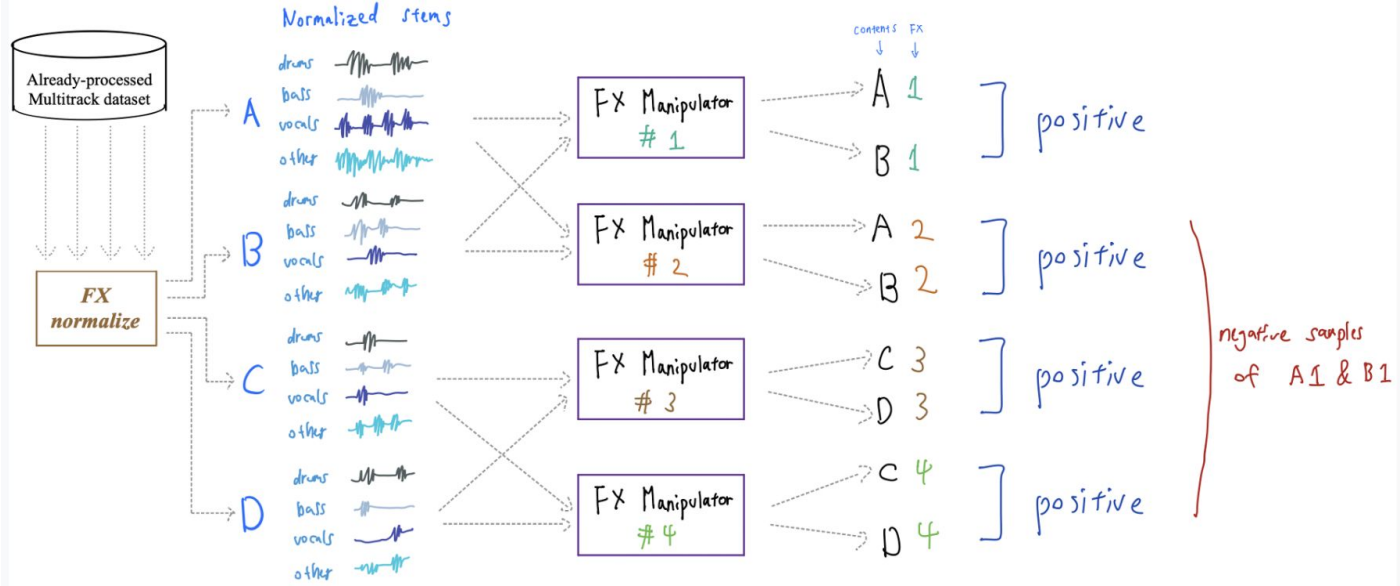
The First Mixing Style Transfer System



Koo, J., Martínez-Ramírez, M. A., Liao, W. H., Uhlich, S., Lee, K., & Mitsufuji, Y. (2023, June). Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

FXencoder

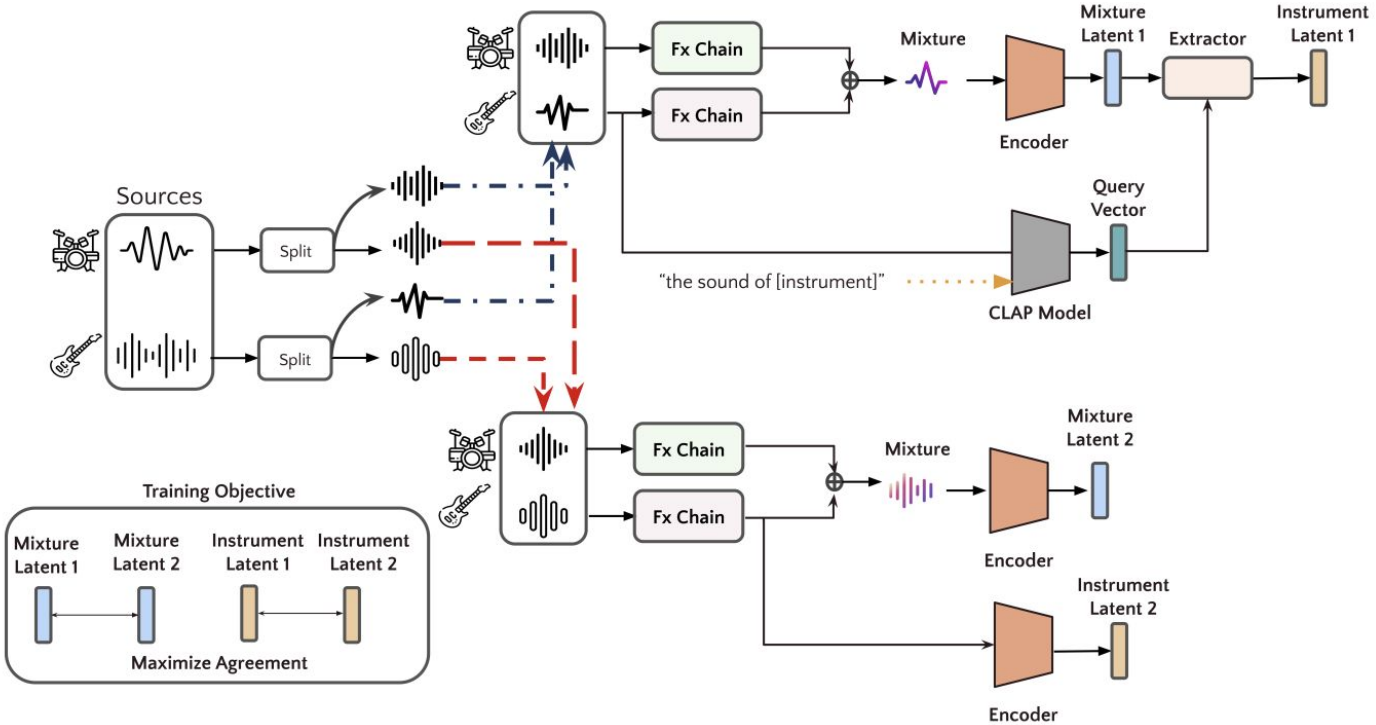
- Using contrastive learning for audio effects representation learning



Koo, J., Martínez-Ramírez, M. A., Liao, W. H., Uhlich, S., Lee, K., & Mitsufuji, Y. (2023, June). Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Fx-Encoder++

- Improved version by two-level contrastive objectives



Yeh, Yen-Tung, et al. "Fx-Encoder++: Extracting Instrument-Wise Audio Effects Representations from Mixtures." in Proc. of ISMIR 2025.

Contrastive Learning

- Core idea of contrastive learning
  - Maximize agreement between positive pairs and minimize agreement between negative pairs
- A common approach for representation learning
  - Multi-modal model: CLIP, CLAP

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.
Wu, Yusong, et al. "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

Contrastive Learning for Audio Contents

- Usually, the audio representation is trained for focusing on **contents**
- This brings up the difficult usage for us because we only want the audio effects information
- To be noticed: this doens’t mean the general audio representation didn’t include the audio effects information. It is about all information are entangled and hard to control [1]

[1] Hawley, Scott H., and Christian J. Steinmetz. "Leveraging neural representations for audio manipulation." arXiv preprint arXiv:2304.04394 (2023).

Contrastive Learning for Audio Effects

- Redesigning the contrastive objective to focus on audio effects
- Positive Sample: the audio content that share the same audio effects processing
  - The embedding should be **invariance** to the audio content
  - It means no matter what the audio content is, if it shared the same audio effects processing, then the embedding should be clustered
- This is the core idea of FXEncoder and Fx-Encoder++

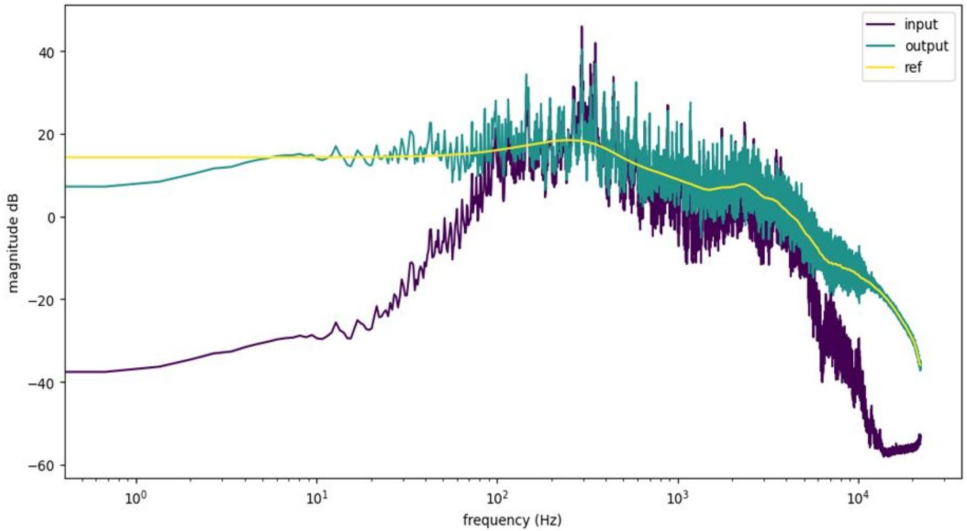
[1] Hawley, Scott H., and Christian J. Steinmetz. "Leveraging neural representations for audio manipulation." arXiv preprint arXiv:2304.04394 (2023).

Positive Sample for Audio Effects is Ambiguous

- Recall the contrastive objective of positive sample:
  - the audio content that share the same audio effects processing
- This brings up several ambiguous scenarios:
  - the source is often already processed by some audio effects
  - how to determine whether the input source is already processed
  - some audio effects are inherently from the source
- Ambiguous and unclear definition for creating positive sample

Back to Audio Effects Normalization

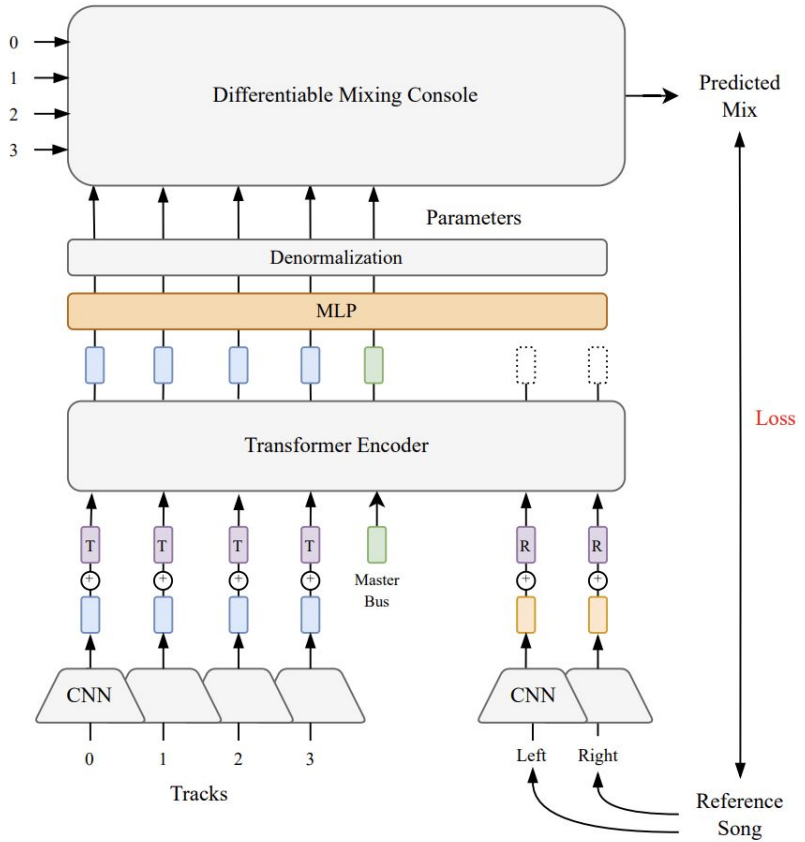
- Recall audio effects normalization: Map each audio to the average audio effects features



Audio effects normalization gives us the proper definition to design positive pair for audio effects

Diff-MST

- Using different segments from the same song (assuming it shares the same mixing style)



Vanka, Soumya Sai, et al. "Diff-mst: Differentiable mixing style transfer." in Proc. of ISMIR 2024.

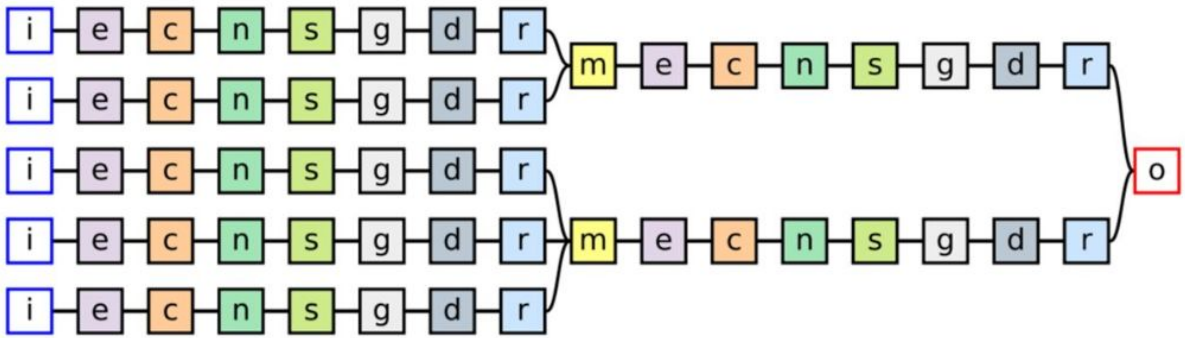
Reverse Engineering of Mixing

Reverse Engineering

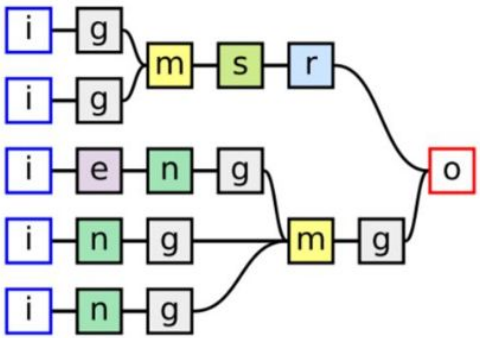
- Given multitrack input & final mixture, can we understand the intermediate audio effects processors?
- Different from previous methods, reverse engineering aims to get how exactly this mixture is mixed by the given input
- Previous methods aim to train a model that can automatically help us mix

Search for Music Mixing Graphs

- The idea is about we can interpret the mixing process as the graph-like structure
- After using the full complex effects graphs to fit specific song, we can **prune** it to get the more efficient mixing graph



(a) Full mixing console (before pruning)



(b) Pruned graph

Lee, Sungho, et al. "Searching for music mixing graphs: A pruning approach." in Proc. of DAFx (2024).

Evaluation

Evaluation is Difficult for Mixing

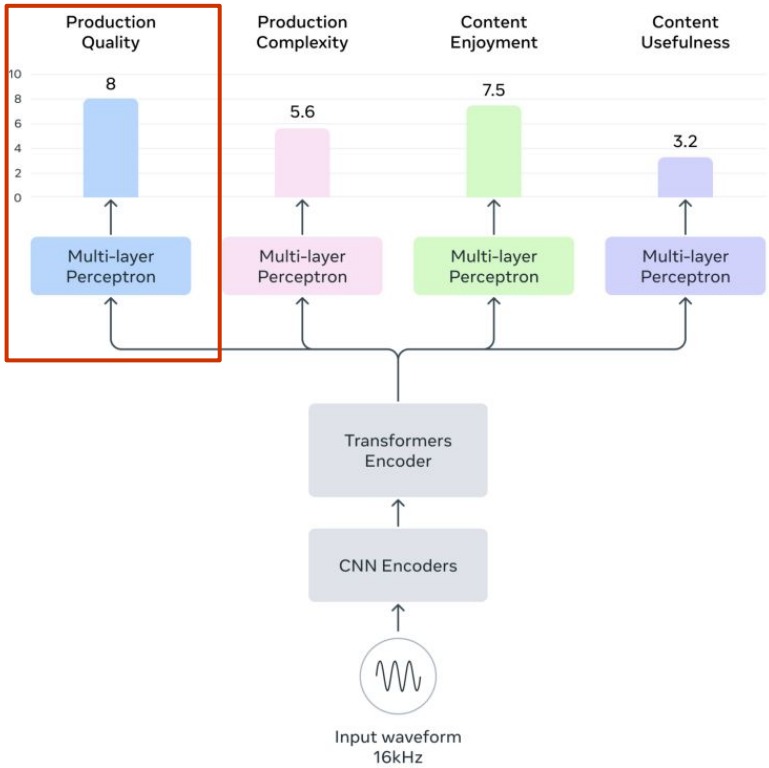
- No single metric can evaluate the “mixing quality”
- Comparing with ground truth gives us certain sense but not good enough
  - This is because even the same multitrack audio, there are several good mixes can be generated

Objective Evaluation

- Objective evaluation of music production tasks remains an open field of research
  - This is because most of criteria don’t actually capture “production quality”
- Current way is using
  - audio features related to audio effects
  - audio objective evaluation metric in other task

Meta Audiobox Aesthetics

- Can’t evaluate production quality properly, this is because
  - The input is **mono**, **16kHz** audio



Subjective Evaluation

- Subjective evaluation is the most reliable approach to evaluate quality of mixing
  - Time-consumiing
  - Expensive
- However, as I know, when developing the music production product, subjective evaluation is still the most promising one

References

Researchers

- Marco A. Martínez-Ramírez
  - Research Scientists at Sony AI
- Christian J. Steinmetz
  - Research Scientists at Suno
- Junghyun (Tony) Koo
  - Research Scientists at Sony AI
- Soumya Sai Vanka
  - Ph.D. student at Queen Mary University
- Sungho Lee
  - Ph.D. student at Seoul University
- Chin-Yun Yu
  - Ph.D. student at Queen Mary University
- Yen-Tung Yeh
  - Ph.D. student at National Taiwan University

Tutorials & Toolkit

- automix-toolkit
  - Helping people to facilitate the AI mixing
- deep learning for automixing
  - The book for AI mixing tutorials
- graphfx
  - The library for music mixing graph processing

If you are interested in doing anything related to Mixing, feel free to contact me