Use **wet multitrack music data and repurpose it** to train supervised deep learning models that perform **automatic music mixing**

We use **data preprocessing** that calculates **average features** related to **audio effects** on a music source separation dataset. Based on these features, we **"effect-normalize"** the wet stems and then train an automatic mixing network.

At inference, the same preprocessing is applied to dry data.

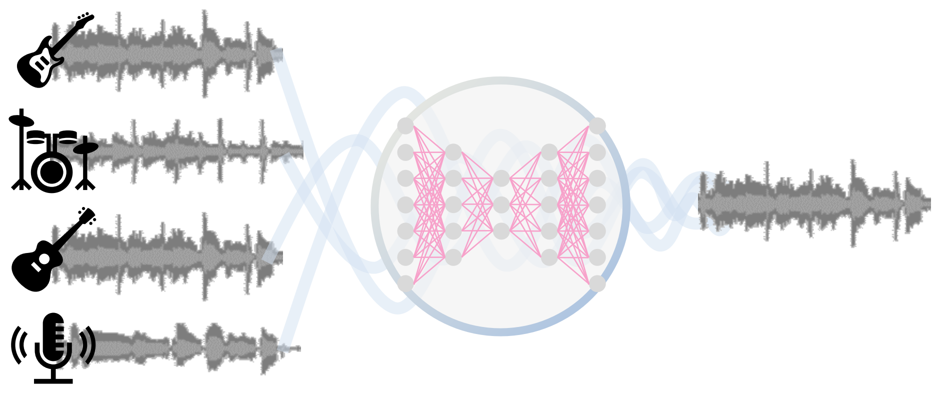
paper, mixes and code

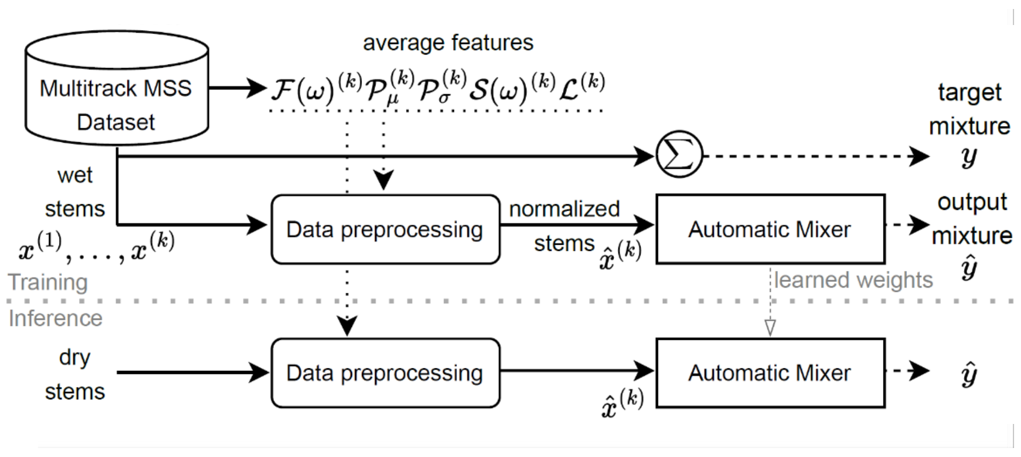
SONY
R&D Center

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

Marco A Martínez-Ramírez[1], Wei-Hsiang Liao[1], Giorgio Fabbro[2], Stefan Uhlich[2], Chihiro Nagashima[1], Yuki Mitsufuji[1]

[1]Sony Group Corporation, Tokyo, Japan
[2]Sony Europe B.V., Stuttgart, Germany





Methodology

Our method consists of a data preprocessing which computes average features related to audio effects on an out-of-domain dataset, e.g. a music source separation (MSS) dataset (4 stems; vocals, drums, bass, other).

Instead of removing audio effects, we normalize each stem based on the audio features related to each class of audio effects.

During training, we expect the models to learn how to undo or denormalize the input stems and thus approximate the original mix. At inference, the same data preprocessing is applied to dry data, thus yielding a fully automatic mixing system.

We propose normalization schemes for **loudness**, based on the average LUFS level; **equalization**, based on the average frequency magnitude spectrum; **panning**, based on the average spectral-panning position; **compression**, based on the average onset peak level; and **reverberation**, based on a stochastic data augmentation.

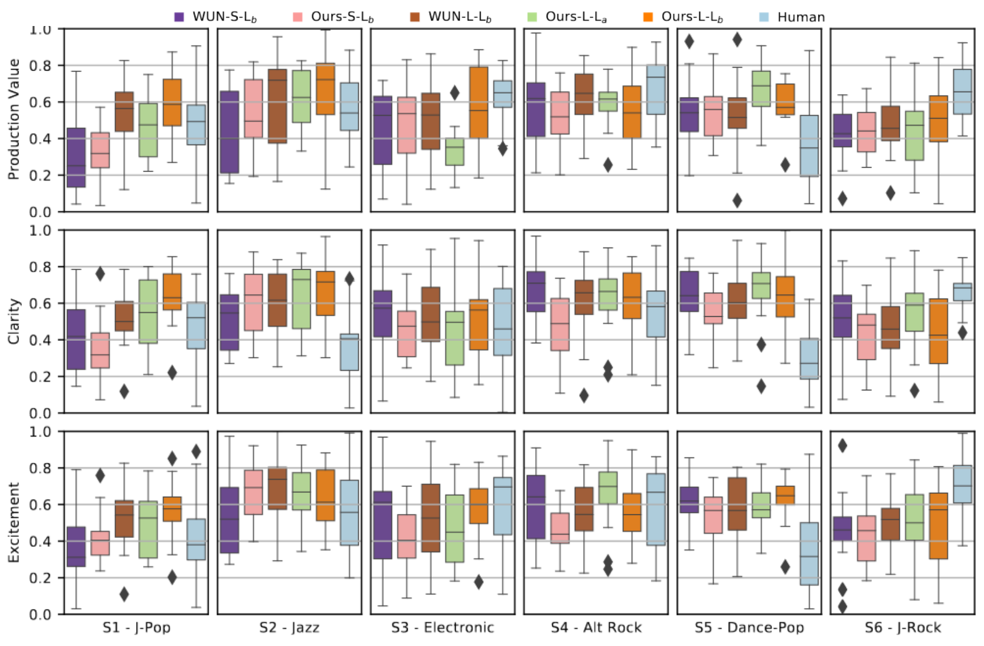
Contributions

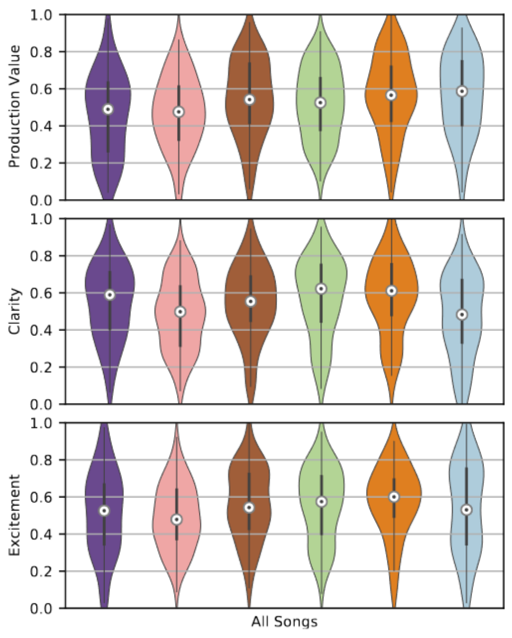
1. Novel data preprocessing step that allows training with out-of-domain data
2. New deep learning architecture for automatic music mixing tasks
3. Exploration of stereo-invariant loss functions
4. Design of a perceptual listening test targeting highly skilled professionals
5. Listening test results showing that our approach is indistinguishable from professional human-made mixes.

Results

We designed a perceptual listening test aimed only at highly qualified professionals (avg 11.6 years of mixing experience).

Our mixes, when compared to professional mixes, scored higher in terms of Clarity and are indistinguishable in terms of Production Value and Excitement.





Production Value	WUN-S-L_b	Ours-S-L_b	WUN-L-L_b	Ours-L-L_a	Ours-L-L_b	Human
WUN-S-L_b	·	o	$\underline{\star}$	o	$\underline{**}$	$\underline{\star}$
Ours-S-L_b	o	·	o	o	$\underline{**}$	$\underline{\star}$
WUN-L-L_b	$\overline{\star}$	o	·	o	o	o
Ours-L-L_a	o	o	o	·	o	o
Ours-L-L_b	$\overline{**}$	$\overline{**}$	o	o	·	o
Human	$\overline{\star}$	$\overline{\star}$	o	o	o	·
Clarity	**WUN-S-L_b**	**Ours-S-L_b**	**WUN-L-L_b**	**Ours-L-L_a**	**Ours-L-L_b**	**Human**
WUN-S-L_b	·	o	o	o	o	o
Ours-S-L_b	o	·	o	$\underline{*}$	$\underline{**}$	o
WUN-L-L_b	o	o	·	o	o	o
Ours-L-L_a	o	$\overline{*}$	o	·	o	$\overline{*}$
Ours-L-L_b	o	$\overline{**}$	o	o	·	$\overline{\star}$
Human	o	o	o	$\underline{*}$	$\underline{\star}$	·
Excitement	**WUN-S-L_b**	**Ours-S-L_b**	**WUN-L-L_b**	**Ours-L-L_a**	**Ours-L-L_b**	**Human**
WUN-S-L_b	·	o	o	o	o	o
Ours-S-L_b	o	·	$\underline{*}$	$\underline{\star}$	$\underline{*}$	o
WUN-L-L_b	o	$\overline{*}$	·	o	o	o
Ours-L-L_a	o	$\overline{\star}$	o	·	o	o
Ours-L-L_b	o	$\overline{*}$	o	o	·	o
Human	o	o	o	o	o	·

Table Post hoc Mann-Whitney test results of pairwise comparison with Bonferroni Correction. o > 0.05, $\star < 0.05$, $* < 0.01$, $** < 0.001$. E.g. when y-axis is compared to x-axis, $\overline{*}$ or $\underline{*}$ indicate y-axis is significantly better or worse than x-axis for a p-value < 0.01, respectively.