MIREX 2024: Renewed, Revived, and Ready for the Future

MIREX 2024 Poster

Future MIREX Team

Introduction

After a three-year break, we revive MIREX starting in 2024. Our goal is to introduce new tasks and datasets that reflect the rapid advancements in computer music research. We envision the new MIREX serving the following key purposes:

1. **Modern Tasks:** New tasks aiming to push the boundaries of current research and foster innovation within the field of music information retrieval.
2. **Benchmarking**: To establish well-defined, fair, and eventually open-source benchmarks to assess music-related models and algorithms.
3. **Online Evaluation & Live Leaderboard**: An ongoing process to build an interactive evaluation environment, where researchers can test their working-in-progress models for instant feedback.

MIREX 2024 Tasks

	Tasks	# Teams	# Submissions
Traditional MIR Tasks	Audio Chord Estimation	0	0
	Lyrics-to-Audio Alignment	1	1
	Cover Song Identification	3	3
Modern MIR Tasks	Symbolic Music Generation	1	1
	Music Audio Generation	1	1
	Music Description & Captioning	4	15
	Polyphonic Transcription	3	3
	Singing Voice Deepfake Detection	4	7

MIREX 2025: An Outlook

1. More tasks & new *Call for Challenges* next year!
2. Open-sourced benchmarking.
3. Live leaderboard for more tasks.

MIREX 2024 Evaluation Results

Please scan the QR code for the results!

https://www.music-ir.org/mirex/wiki/2024:MIREX2024_Results

Task Spotlight: Codabench Evaluation

Codabench is a competition platform that incorporates *live leaderboards* and *customizable tasks*. Two new tasks this year were evaluated on Codabench and received widespread attention:

- **Music Description & Captioning**
- **Singing Voice Deepfake Detection**

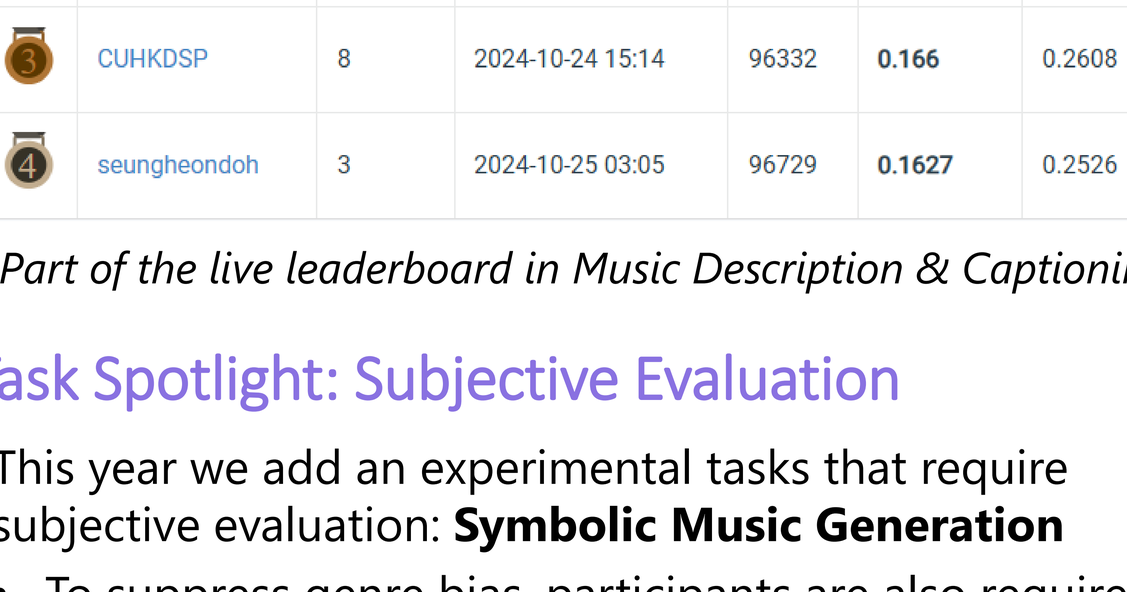
#	Participant	Entries	Date	ID	ROUGE-L	BLEU-1
1	ee895	2	2024-10-24 01:10	95923	**0.2111**	0.688
2	yuda	2	2024-10-29 21:00	99815	**0.1729**	0.3041
3	CUHKDSP	8	2024-10-24 15:14	96332	**0.166**	0.2608
4	seungheondoh	3	2024-10-25 03:05	96729	**0.1627**	0.2526

Part of the live leaderboard in Music Description & Captioning

Task Spotlight: Subjective Evaluation

This year we add an experimental tasks that require subjective evaluation: **Symbolic Music Generation**

- To suppress genre bias, participants are also required to submit their test samples
- Comparison is done by a questionnaire containing selected generated examples of submissions and baselines



Part of the subjective evaluation results

MIREX 2024: Renewed, Revived, and Ready for the Future

MIREX 2024 Poster

Future MIREX Team

Results: Singing Voice Deepfake Detection

Team	Methods Used	EER on test_A	EER on test_B
UNIBS1	Log-spectrogram + ResNet - Vocals	**2.38**	**9.81**
UNIBS2	Log-spectrogram + ResNet - Mixtures	2.70	12.19
IMS-SCU1	Ensemble - Vocals	2.70	12.95
IMS-SCU2	WavLM - Vocals	3.54	15.32
IMS-SCU3	Ensemble - Mixtures	3.61	11.00
NTU	SingGraph - Mixtures	4.31	31.82
IMS-SCU4	WavLM - Mixtures	4.94	16.72
PDL	Ensemble - Vocals	5.80	22.01
Baseline1	Wav2vec - Vocals	6.09	24.09
Baseline2	Raw - Vocals	8.84	26.11
Baseline3	Wav2vec - Mixtures	9.57	21.45
Baseline4	Raw - Mixtures	10.88	17.69

Results: Polyphonic Transcription

Team	Methods Used	Summary Onset F1	Holistic Note F1
wlazbzfll	CRNN + regression onset & offset	0.7066	0.1607
teamWLY	FiLM with CNN+ LSTM + regression onset & offset	0.9592	0.8465
Transkun V2 (Baseline 1)	ViT + Neural SemiCRF	0.9490	0.8764
Transkun V2 Aug (Baseline 2)	ViT + Neural SemiCRF + Data Augmentation	**0.9648**	**0.9081**
hFT-Transformer	Two stacks of Transformers for Frequency and Time + regression onset & offset	0.9416	0.8359

* average note onset F1 = average of note onset F1 on all three datasets
* holistic note F1 = average of note onset+offset+velocity on Maestro and SMD

Results: Music Description & Captioning

Team	Methods Used	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
ee895	llama & JMLA	**0.2111**	0.6880	0.0798	0.0007	0.0000	0.0987
yuda	llama 2 7b	0.1729	0.3041	0.0415	0.0062	0.0009	0.1484
CUHKDSP	llama & JMLA	0.1660	0.2608	0.0252	0.0034	0.0004	0.1555
seungheondoh	LP-MusicCaps	0.1627	0.2526	0.0240	0.0031	0.0003	0.1506
baseline	LP-MusicCaps	0.1190	0.1270	0.0040	0.0013	0.0000	0.1670

Results: Music Audio Generation

Team	Methods Used	Frechet Distance ↓	FAD ↓	KLD ↓	Inception Score ↑	Relative Overall ↑
S1-CodecLM	7B decoder only + 2 stage semantic tokenizer	**13.77**	2.67	**1.71**	1.52 ±0.04	**0.716**
B1-MusicGen-Large	MusicGen	19.05	**2.50**	2.11	1.57 ±0.03	0.672
B2-MusicGen-Medium	MusicGen	24.58	3.59	2.46	**1.61 ±0.06**	0.356
B3-MusicGen-Small	MusicGen	26.21	3.75	2.61	1.58 ±0.04	0.167
GT	Ground truth	0	0	0	1.65 ±0.06	-

Results: Symbolic Music Generation

Team	Methods Used	Coherecy	Naturalness	Creativity	Musicality
Chart-Accompaniment	BART	1.92 ± 0.11^{d}	1.87 ± 0.10^{c}	2.62 ± 0.13^{c}	2.01 ± 0.11^{c}
AccoMontage (BL-1)	Style Transfer	$\mathbf{3.77 \pm 0.11^{a}}$	$\mathbf{3.59 \pm 0.11^{a}}$	$\mathbf{3.65 \pm 0.11^{a}}$	$\mathbf{3.63 \pm 0.12^{a}}$
Whole-Song-Gen (BL-2)	DDPM	3.59 ± 0.11^{b}	3.24 ± 0.11^{b}	$\mathbf{3.66 \pm 0.10^{a}}$	3.47 ± 0.13^{b}
Compose-&-Embesslish (BL-3)	Transformer	3.39 ± 0.10^{c}	3.38 ± 0.12^{b}	3.13 ± 0.10^{b}	3.36 ± 0.11^{b}

Note: Results are reported in the form of mean ± sem^{s} (sem refers to standard error of mean), where *s* is a letter. Different letters within a column indicate significant differences (p-value $p < 0.05$) based on a Wilcoxon signed rank test.

Results: Cover Song Identification

Team	Methods	mAP	Rank-1	Rank-5	Rank-10
Liufeng	CoverHunter	0.783	0.890	0.925	0.937
MTG-SonyAI	Discog-VI	0.807	0.889	0.932	0.944
ByteDance	ByteCover2	**0.877**	**0.947**	**0.966**	**0.972**

Results: Lyrics-to-Audio Alignment

Jamendo dataset

Team	Methods	Average absolute error	Median absolute error	Correct segments %	Correct onsets with tolerance %
FZZ1	WavLM + Conformer	0.547	0.047	0.686	0.912
NUS (baseline)	Genre-informed Silence + Phone Model	**0.217**	**0.046**	**0.751**	**0.945**

Jamendo v2 MultiLang dataset

Team	Methods	Average absolute error	Median absolute error	Correct segments %	Correct onsets with tolerance %
FZZ1	WavLM + Conformer	**0.584**	0.252	**0.683**	**0.887**
NUS (baseline)	Genre-informed Silence + Phone Model	0.651	**0.136**	0.502	0.729