Meta

Audiobox Aesthetics

Unified Automatic Quality Assessment for Speech, Music, and Sound

Paper: https://arxiv.org/abs/2502.05139
Github: https://github.com/facebookresearch/audiobox-aesthetics
HuggingFace Demo: https://huggingface.co/spaces/facebook/audiobox-aesthetics

Andros Tjandra

Conversational AI reading group (06/12/2025)

Collaborators: Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis,
Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, Wei-Ning Hsu

AI at Meta

Self-introduction

- Name: Andros Tjandra
- Current occupation: Research Scientist @ Meta AI (FAIR), USA.
- Education:
  - Ph.D: NAIST, Japan
    - Thesis: Machine Speech Chain
  - Master & Bachelor: University of Indonesia, Faculty of Computer Science
- Recent research interests:
  - Speech SSL & multilingual ASR
    - XLS-R, XLS-R-LID
    - MMS, Meta 70-languages ASR
  - Audio generation
    - Audiobox, Audiobox-RAG
    - Moviegen-Audio generation
- Link:
  - Google Scholar: URL
  - Website: URL

Outline

- Motivation
- Data collection
- Data analysis
- Model
- Objective Evaluation
- Downstream Tasks

Motivation

- Audio Aesthetics is a framework to assess audio quality from multiple perspectives.
- We need this tools for several reasons:
  - Data curation is crucial for large scale training
    - Manual labelling is too costly and not scalable
    - Automatically quality assessment is required
  - Automatic evaluation for generative model
    - In the large scale experiment, there are tonnes of optimization involved during training.
    - Instead of evaluating every (potential) checkpoint with human evaluation, we want to have quick signal on its quality.

Motivation

- Drawbacks of conventional quality assessments
  - Signal-based distortion
    - Parallel ground-truth audio is required.
    - SI-SDR, MCD, LSD etc. are not directly related to human perception.
  - Distribution similarity
    - FAD (Adapting Frechet Audio Distance) do not provide utterance-based assessment.
  - Neural-based quality predictor
    - Most of prior works focus on speech (e.g. MOSNet, SQUIM, etc).
    - Only measuring the overall quality (e.g. mean opinion score, MOS).

Motivation

- Drawbacks of conventional quality assessments (cont.)
  - Most of prior works focus on predicting Mean Opinion Score or overall quality.
  - MOS are heavily dependent on the audio domain and affected by many different factors.
  - MOS alone is not always reliable and very noisy (i.e., corpus effect [1] or range-equalizing bias [2]).
- Therefore, instead of trying to measure only 1 axes, we want to “factorize” the score into several independent axes to reduce the ambiguity compared to standard MOS score.

[1] Generalization ability of MOS prediction networks (Cooper et al., 2022)
[2] Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech (Cooper et al., 2023)

Motivation

- Ideal quality predictors for data curation
  - ✅Non-intrusive assessment
  - ✅Utterance-based assessment
  - ✅Highly related to human perception
  - ✅Supporting arbitrary audio types (e.g. sound, music, and speech)
  - ✅Detailed information for different applications and scenarios
- To build suitable predictors
  - Human annotations on different audio types
  - New annotation guidances of data collections for different proposes

Audio aesthetics evaluation survey

- In this work, we groups each audio into several audio type:
  - Speech
  - Music
  - Sound (ambient and sound effects)

Q1. What modalities are present in the audio?

- ☐ Speech;
- ☐ Music - no vocal;
- ☐ Music - has vocal;
- ☐ Other sound events;
- ☐ Ambient sound;

Audio aesthetics evaluation axes

- Production Quality (PQ)
  - Recording quality
  - Focuses on the technical aspects of quality instead of ambiguous subjective quality
    - Clarity & fidelity, dynamics, frequencies and spatialization of the audio

Q2. What is the Production Quality of this audio? Rate from 1 to 10.
You should focus only on **technical aspects of quality** instead of subjective quality. We want you to rate the quality based on aspects including clarity & fidelity, dynamics, frequencies and spatialization of the audio.

More specifically,

1. **Clarity and Fidelity**: High quality audio should have clear, crisp sound with minimal distortion, noise, or artifacts:
   - The instruments, vocals, and other elements should be well-defined and easily distinguishable / intelligible;
   - No microphone noise / other white noise;
   - No distortions of vocals and other elements;
   - No other audio artifacts (e.g. hissing, buzzing, shrill, etc.)
2. **Dynamics**: High quality audio should maintain an appropriate dynamic range, encompassing both quiet and loud passages with clarity and impact:
   - Transitions between different dynamic levels are smooth and natural, with gradual changes in volume rather than abrupt jumps;
   - The subtle nuances and variations in volume should be well-preserved.
3. **Frequencies**: High quality audio should exhibit a balanced and natural frequency response across the entire spectrum, with each frequency range contributing harmoniously to the overall sound:
   - Low frequency bass sound should be well-defined without muddiness or boominess;
   - High frequency sound should be crisp and detailed without harshness or sibilance.
4. **Spatialization**: For multi-channel audio, the spatialization of audio elements within the stereo field should be well-defined and appropriately positioned. This creates a sense of depth and dimensionality, enhancing the listening experience.
5. **Overall technical proficiency**: During the recording, mixing, and mastering of audio, whether it exhibits skillful application of techniques and tools to achieve high-quality sound reproduction and optimal sonic results.

Audio aesthetics evaluation axes

- Production Complexity (PC)
  - Focuses on the complexity of an audio scene, measured by number of audio components
  - Different scenarios require different PC
    - Low PC with simple stem is suitable for creating synthetic data for source separation, etc.
    - HIgh PC with high diversity is suitable for training neural codecs, etc.

Q3. What is the Production Complexity of this audio?, Rate from 1 to 10.

- Complex production means that there are many audio components (may or may not from the same audio modality) mixed together
  - e.g. You can think of a piece of podcast audio with speech, music and sound effects mixed together as high complexity. Alternatively, a piece of symphony with many instruments playing together should also be considered as complex;
- Simple production means with few audio elements and components
  - e.g. Only one speaker speaking no other audio events, Piano sound only, etc.

Audio aesthetics evaluation axes

- Content Enjoyment (CE)
  - Focuses on the subject quality of an audio piece. It's a more open-ended axis, some aspects might includes emotional impact, artistic skill, artistic expression, as well as subjective experience, etc.
  - Enjoyment can be decoupled from production quality
    - A 70s rock music recording can be noisy/lower quality but very enjoyable for some listeners.

Q4. How much do you enjoy this audio? Rate from 1 to 10.
In this question we ask you to rate the subject quality, it's an open-ended question as everyone has their own preferences and tastes. However, there are some directions / aspects that you can consider when appreciate these audio pieces:

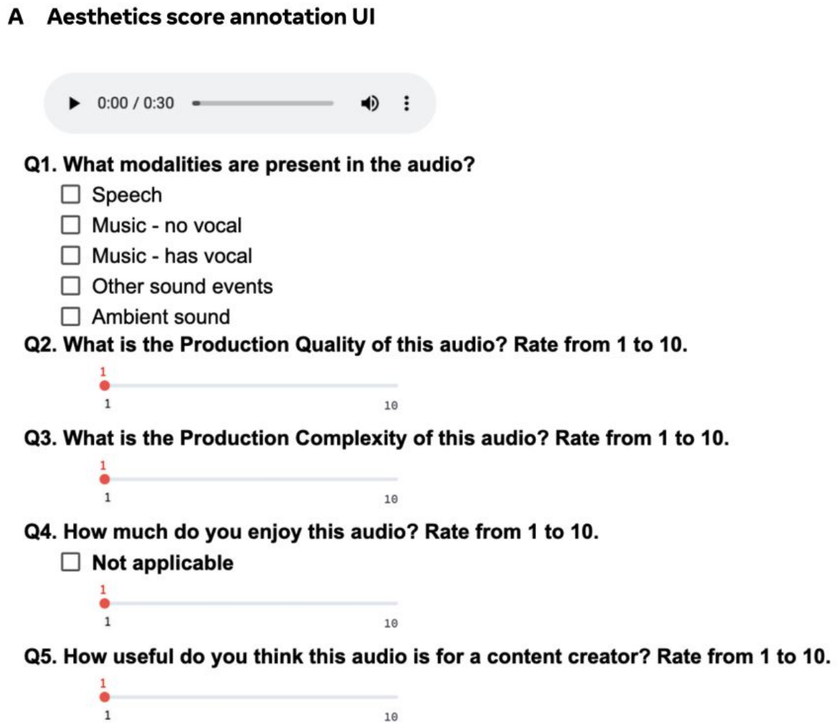
1. **Emotional Impact**: This means the ability of the audio to evoke emotions, convey mood, and connect with the listener. Are you able to resonate with the expressiveness / emotive quality of the audio piece?
2. **Artistic Skill**: If the audio is for entertainment purpose (e.g. clips of music / podcast / audiobook), then the performer / speaker should demonstrate high artistic / professional skills;
3. **Artistic Expression**: Focus on the creativity and originality in the audio. Is it innovative and gives you a unique audio experience?
4. **Subjective Experience**: Ultimately, the subjective experience of the listener is paramount when rating the aesthetic/subjective quality of audio. Your score should reflect your personal preferences, individual taste, and emotional response.

Audio aesthetics evaluation axes

- Content Usefulness (CU)
  - For audio producers
  - Focus on evaluating the likelihood of leveraging the audio as source material for content creation.
  - Sound effects is difficult to evaluate its enjoyness but easy to evaluate the usefulness

Q5. How useful do you think this audio is? Rate from 1 to 10.
For usefulness, imagine you are a YouTube or Instagram content creator, and want to generate popular and high quality audio-visual clips (movie level quality), how likely would you be able to use this audio as source material to create some contents?

AES Annotation UI (overall)



- We also reject audio if annotators reported following issues:
  - Audio doesn't load properly
  - Has violating content:
    - Hate speech - Violent
    - Sexual content
    - Strong & explicit languages (e.g., profanities, obscenities, etc.)

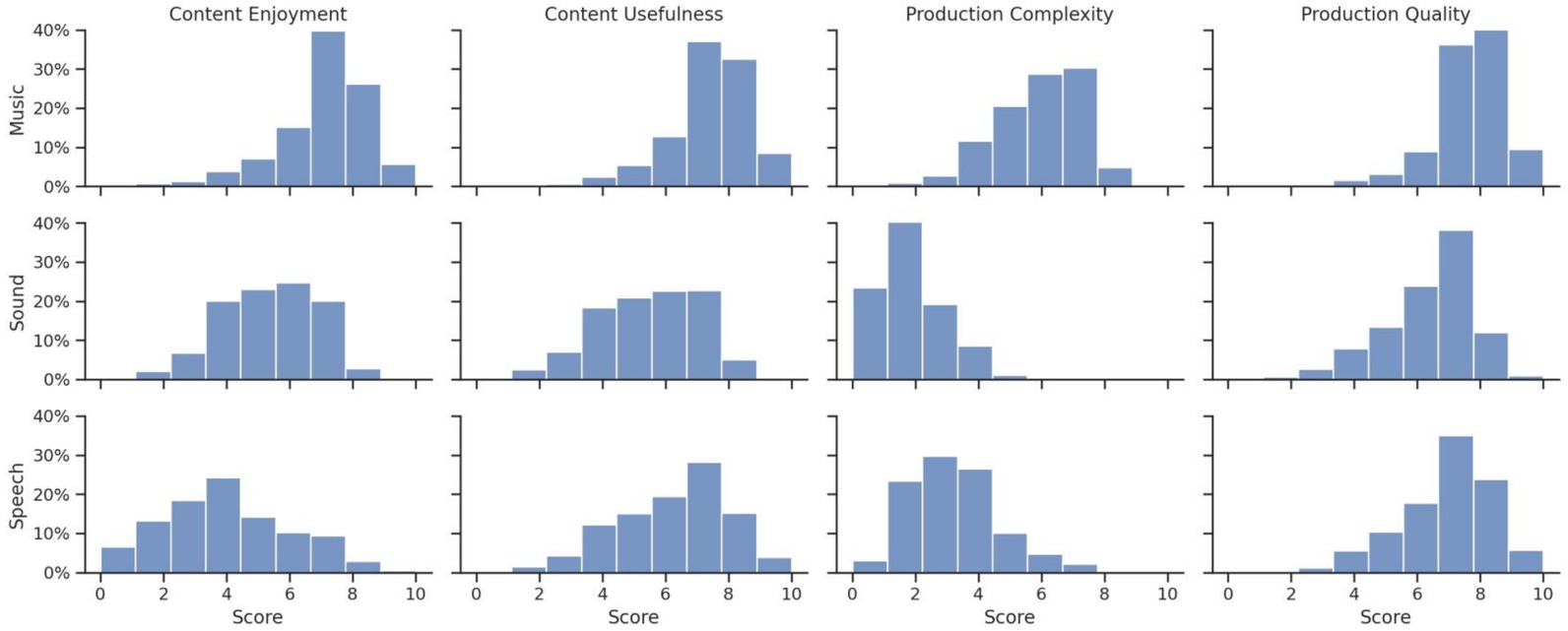
Data preparation

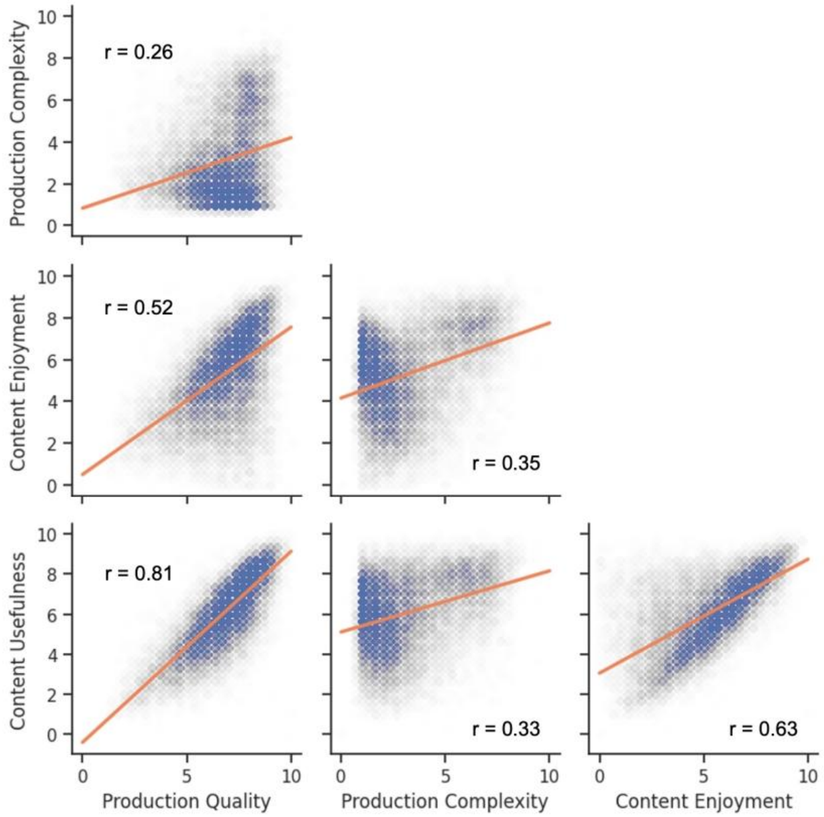
- Sample ~500hrs data including speech, music, and sounds
  - Evenly sample by different attributes
    - Speech: gender, emotion, quality etc.
    - Music: music types, quality etc.
    - Sound: AED tags, quality etc.
  - Loudness normalization
- Sample 3000 speech, music, and sound files for open-source benchmark
  - Evenly sample by different attributes
  - Without loudness normalization for easy reproduce

Data preparation

- Works with outside vendor to get 3 annotation for each audio
  - Annotator calibration is important for the annotation consistency
    - We label a small golden set by our team and filter out the samples with low agreements
    - Qualified raters are with Pearson correlation > 0.7 on production quality and complexity (more objective measurements)
  - Examples are important for annotators to well understand the measurements
    - We provides samples with different level of AES scores in the annotation guidelines (see Appendix of our papers)
  - Including all three audio types in each annotation batch to avoid bias
  - Total raters: 158

Data distribution and correlation





Aesthetic Model

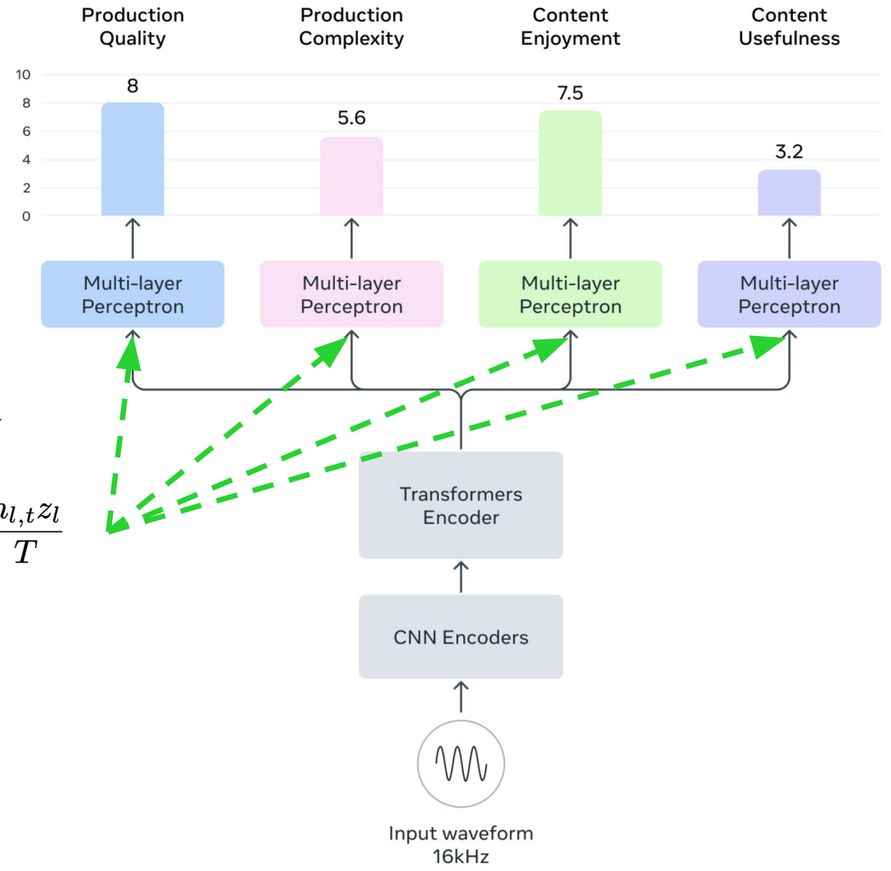
- Input: raw waveform (16kHz)
- Audio encoder (grey part) shared across different tasks
  - CNN Encoders
  - Transformers layer
- Multi-layer perceptron (colorised)
  - 5 layers of non-linear block
  - Activation function GeLU
  - Layer norm
- Loss function: Mean Absolute Error (MAE) + Mean Squared Error (MSE)

$$\mathcal{L} = \sum_{a \in \{PQ,PC,CE,CU\}} (y_a - \hat{y_a})^2 + |y_a - \hat{y_a}|.$$

$$z_l = \frac{w_l}{\sum_{l=1}^{L} w_l}$$

$$\hat{e} = \sum_{t=1}^{T} \sum_{l=1}^{L} \frac{h_{l,t} z_l}{T}$$

$$e = \frac{\hat{e}}{\|\hat{e}\|_2}$$

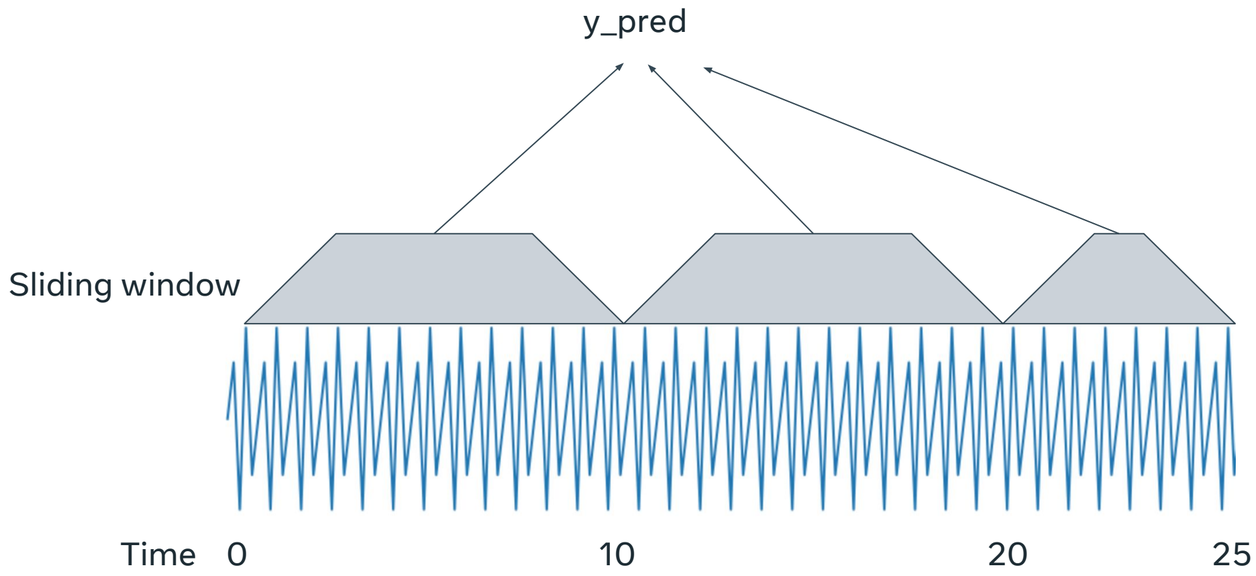


Model Inference

Algorithm 1 Audio Aesthetic Inference

Require: x: audio input, sr: sample rate
Ensure: y_pred: predicted aesthetic score

Initialize $lens \leftarrow []$, $preds \leftarrow []$
$stepsize \leftarrow sr \times 10$
for $t \leftarrow 0$ to $\text{len}(x)$ with step $stepsize$ **do**
    $x_{now} \leftarrow x[t \times stepsize : (t+1) \times stepsize]$
    Append $\text{AES}(x_{now})$ to $preds$
    Append $\text{len}(x_{now})$ to $lens$
end for
$w \leftarrow lens/\text{sum}(lens)$
$y\_pred \leftarrow \text{sum}(preds \times w)$
return y_pred



Objective Evaluation

- Dataset
  - Speech: Voice MOS Challange 2022 (Huang et al, 2022)
  - Sound & Music: PAM dataset
- Model for comparison:
  - UTMOSv2 (Kaito et al., 2024; challenge winner)
  - PAM (Deshmukh et al., 2023)
  - TorchSQUIM-PESQ (Kumar et al,, 2023)
  - Audiobox-Aesthetics-{PQ,PC,CE,CU}

Table 1 Utterance- and system-level correlations between human-annotated and predicted scores on speech test set.

Model	VMC22-main		VMC22-OOD	
	utt-PCC	sys-SRCC	utt-PCC	sys-SRCC
PAM	0.357	0.403	0.471	0.668
DNSMOS	0.612	0.773	0.459	0.615
SQUIM	0.708	0.711	0.465	0.515
UTMOSv2	0.916*	0.932*	0.634	0.707
Audiobox-Aesthetics-PQ	0.689	0.752	0.651	0.813
Audiobox-Aesthetics-PC	-0.192	-0.394	-0.315	-0.039
Audiobox-Aesthetics-CE	**0.775**	**0.813**	**0.767**	**0.876**
Audiobox-Aesthetics-CU	0.647	0.706	0.655	0.823

*Data leakage: UTMOSv2 is trained using also the BVCC test set.

Table 3 Utterance-level Pearson Correlation Coefficient between human-annotated and predicted scores on PAM-sound.

Model	OVL	GT-PQ	GT-PC	GT-CE	GT-CU
PAM	**0.650**	0.408	0.269	0.542	0.393
Audiobox-Aesthetics-PQ	0.355	**0.617**	0.071	0.406	**0.573**
Audiobox-Aesthetics-PC	0.092	-0.051	**0.654**	0.275	-0.098
Audiobox-Aesthetics-CE	0.464	0.318	0.447	**0.638**	0.279
Audiobox-Aesthetics-CU	0.396	0.583	0.058	0.413	**0.573**

Table 4 Utterance-level Pearson Correlation Coefficient between human-annotated and predicted scores on PAM-music.

Model	OVL	GT-PQ	GT-PC	GT-CE	GT-CU
PAM	**0.581**	0.568	0.377	**0.699**	**0.573**
Audiobox-Aesthetics-PQ	0.464	0.587	0.193	0.449	0.537
Audiobox-Aesthetics-PC	0.251	0.113	**0.710**	0.322	0.096
Audiobox-Aesthetics-CE	0.528	0.487	0.455	0.661	0.488
Audiobox-Aesthetics-CU	0.465	**0.594**	0.221	0.502	0.558

Objective Evaluation (AES-Natural)

- We collected 2950 audio aesthetic annotation on multiple public datasets for open-source.
- Dataset:
  - Speech:
    - EARS
    - LibriTTS
    - CommonVoice13
  - Music:
    - MUSDB18-HQ
    - MusicCaps
  - Sound:
    - AudioSet

Table 5 Utterance-level Pearson Correlation Coefficient between human-annotated and predicted scores (natural speech)

Model	GT-PQ	GT-PC	GT-CE	GT-CU
PAM	0.317	-0.292	0.250	0.284
DNSMOS	0.662	-0.462	0.598	0.632
SQUIM	0.660	-0.466	0.570	0.604
UTMOSv2	0.603	-0.358	0.574	0.588
Audiobox-Aesthetics-PQ	0.888	-0.538	0.783	0.834
Audiobox-Aesthetics-PC	-0.693	**0.700**	-0.643	-0.677
Audiobox-Aesthetics-CE	0.879	-0.544	**0.859**	**0.886**
Audiobox-Aesthetics-CU	**0.898**	-0.565	0.835	0.876

Table 6 Utterance-level Pearson Correlation Coefficient between human-annotated and predicted scores (natural sound)

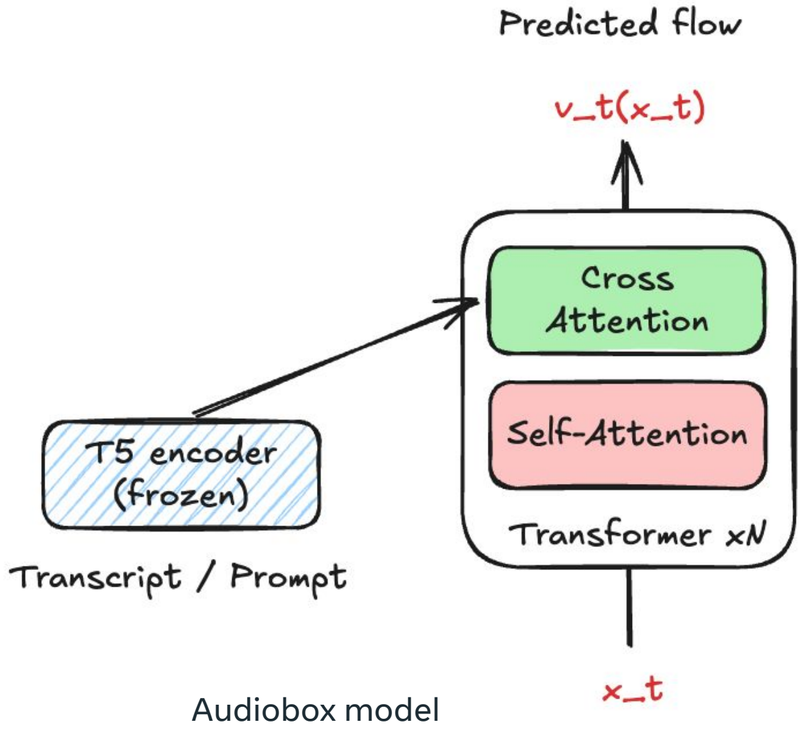
Model	GT-PQ	GT-PC	GT-CE	GT-CU
PAM	0.462	-0.022	0.438	0.443
Audiobox-Aesthetics-PQ	**0.728**	-0.014	0.552	**0.655**
Audiobox-Aesthetics-PC	0.106	**0.758**	0.297	0.017
Audiobox-Aesthetics-CE	0.492	0.288	**0.763**	0.466
Audiobox-Aesthetics-CU	0.676	0.012	0.571	0.644

Table 7 Utterance-level Pearson Correlation Coefficient between human-annotated and predicted scores (natural music)

Model	GT-PQ	GT-PC	GT-CE	GT-CU
PAM	0.656	-0.244	0.675	0.696
Audiobox-Aesthetics-PQ	**0.887**	-0.352	0.664	0.834
Audiobox-Aesthetics-PC	-0.270	**0.905**	0.001	-0.229
Audiobox-Aesthetics-CE	0.643	-0.004	**0.750**	0.697
Audiobox-Aesthetics-CU	0.852	-0.322	0.685	**0.835**

Downstream Task

- Goal: explore the application of an aesthetic model predictor in enhancing the performance of various downstream tasks: text-to-speech (TTS), text-to-audio (TTA), text-to-music (TTM)
- Setup
  - Model: Conditional flow-matching with text conditioning (Audiobox-Sound architecture)
- Experiment scenario
  - Baseline: 100% dataset, standard transcript / description
  - Filtering: filter out part of dataset with aesthetic score lower than *p* percentile. (lose *p*% of dataset)
  - Prompting: 100% dataset, append aesthetic score as prompt (“Audio quality: y”) as prefix. During inference, we explicitly set y with value from higher percentile.



Audiobox model

Effects on audio quality

- Compared to baseline, all filtering and prompting strategy shows better audio quality on pairwise human preference test.
- In addition (see bottom), prompting win compared to filtering strategy on most of experiment.
- Q: How about the text-audio alignment effects?

Table 9 This tables compares model A and model B in term of audio quality judged by human listeners. We report net win rate $[-100\%, 100\%]$ and their 95% confidence interval. Positive value means model A outperforms model B.

Model A	Model B	Speech (%)	Sound (%)	Music (%)
Filter $p = 25$	Baseline	$12.93_{\pm 9.58}$	$9.09_{\pm 8.92}$	$9.98_{\pm 7.33}$
Filter $p = 50$	Baseline	$15.77_{\pm 7.83}$	$7.04_{\pm 9.59}$	$20.99_{\pm 8.42}$
Prompt $p = 50, r = 2$	Baseline	$12.95_{\pm 9.33}$	$5.38_{\pm 8.00}$	$28.62_{\pm 7.96}$
Prompt $p = 75, r = 2$	Baseline	$28.44_{\pm 7.29}$	$10.64_{\pm 8.75}$	$46.17_{\pm 7.46}$
Prompt $p = 90, r = 2$	Baseline	$35.35_{\pm 7.59}$	$15.12_{\pm 8.92}$	$25.06_{\pm 7.92}$
Prompt $p = 50, r = 5$	Baseline	$10.66_{\pm 8.56}$	$14.52_{\pm 9.75}$	$27.66_{\pm 7.72}$
Prompt $p = 75, r = 5$	Baseline	$21.89_{\pm 8.50}$	$18.40_{\pm 8.06}$	$40.67_{\pm 8.75}$
Prompt $p = 90, r = 5$	Baseline	$45.07_{\pm 6.75}$	$18.52_{\pm 8.92}$	$23.80_{\pm 7.92}$
Prompt $p = 75$	Filter $p = 25$	$37.23_{\pm 8.08}$	$9.12_{\pm 8.50}$	$48.55_{\pm 7.83}$
Prompt $p = 75$	Filter $p = 50$	$31.05_{\pm 8.17}$	$-4.51_{\pm 9.25}$	$50.19_{\pm 7.92}$

Effects on text-audio alignment

- Filtering strategy lead us into worse alignment compared to baseline and prompting.
- It made sense since filtering removed *p*% of data during training, and this effects will be more severe if training data is limited.
- Conclusion:
  - **Prompting > filter > baseline** (in term of audio quality)
  - **Prompting == baseline > filter** (in term of audio alignment)
  - We conclude that using **prompting is more effective** than filtering.

Table 8 This table shows objective evaluation to measures text and generated audio alignment for each downstream tasks.

Metric Name		WER ↓	CLAP-sound ↑	CLAP-music ↑
Model	Train data (%)	Speech	Sound	Music
Baseline	100%	2.95	0.40	0.36
Filter $p = 25$	75%	3.37	0.37	0.36
Filter $p = 50$	50%	5.06	0.33	0.36
Prompt $p = 50, r = 2$	100%	2.87	**0.41**	0.36
Prompt $p = 75, r = 2$	100%	2.81	0.40	0.36
Prompt $p = 90, r = 2$	100%	2.83	0.39	0.36
Prompt $p = 50, r = 5$	100%	2.84	**0.41**	0.36
Prompt $p = 75, r = 5$	100%	**2.80**	**0.41**	0.36
Prompt $p = 90, r = 5$	100%	2.76	0.40	0.36

AES prompting on Moviegen-Audio

Example captions for Movie Gen Audio
This audio has quality: 8.0. This audio does not contain speech. This audio does not contain vocal singing. This audio has a description: "gentle waves lapping against the shore, and music plays in the background.". This audio contains music with a 0.90 likelihood. This audio has a music description (if applicable): "A beautiful, romantic, and sentimental jazz piano solo.".
This audio has quality: 7.0. This audio does not contain speech. This audio does not contain vocal singing. This audio has a description: "fireworks exploding with loud booms and crackles.". This audio contains music with a 0.01 likelihood. This audio has a music description (if applicable): "A grand, majestic, and thrilling orchestral piece featuring a massive symphony orchestra with a soaring melody and pounding percussion, evoking a sense of awe and wonder.".

Table 27 (Moviegen paper)

We control the model audio quality by first adding prefix “This audio has quality: x” (same prompt format used during training) followed by other sound and music content prompt.

AES prompting on Moviegen-Audio (Sound)

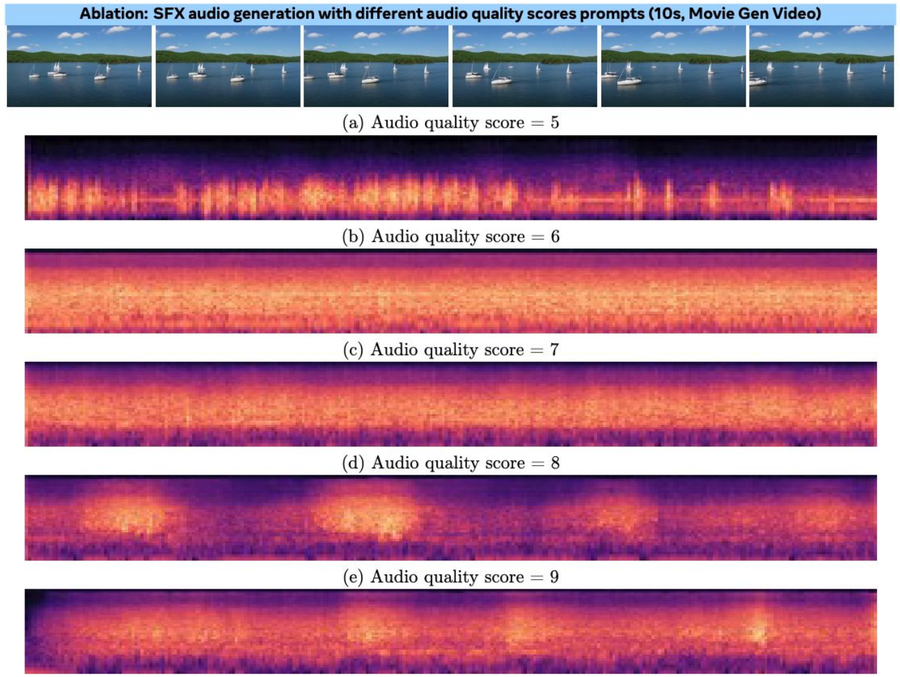


Figure 40 Examples of generated audio with different audio quality score in the text prompt with Movie Gen Audio. For these samples, higher audio quality scores tends to produced audio without wind noises compared to lower audio quality scores. Videos in this Figure found at https://go.fb.me/MovieGen-Figure40.

Audio quality 5

Audio quality 6

Audio quality 7

Audio quality 8

Audio quality 9

https://www.youtube.com/playlist?list=PL86eLlsPNfyi3fo5T74nXCIjWJYqwRx0V

AES prompting on Moviegen-Audio (Sound+Music)

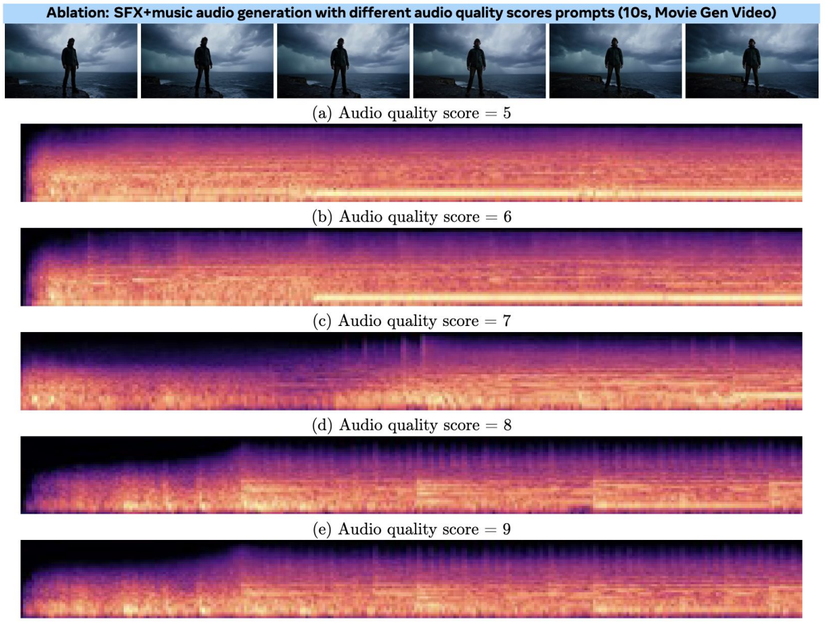


Figure 41 Examples of generated audio with different audio quality score in the text prompt with Movie Gen Audio. Movie Gen Audio controls the generated audio output quality using the input text prompts (refer to 27). As we can observed in the spectrogram plot, lower quality scores contains some high frequency noises and by gradually increasing the quality scores, we got a cleaner audio spectrogram. Videos in this Figure found at https://go.fb.me/MovieGen-Figure41.

Audio quality 5

Audio quality 6

Audio quality 7

Audio quality 8

Audio quality 9

https://www.youtube.com/playlist?list=PL86eLlsPNfyjBhGJub69uZvs-PZsvbYLN

Open-Source Model & Evaluation

Github: https://github.com/facebookresearch/audiobox-aesthetics

Simple installation & usage:

Installation

1. Install via pip

```
pip install audiobox_aesthetics
```

2. Install directly from source

This repository requires Python 3.9 and Pytorch 2.2 or greater. To install, you can clone this repo and run:

```
pip install -e .
```

How to load model using HuggingFace way (for finetuning, etc)

```
from audiobox_aesthetics.model.aes import AesMultiOutput
model = AesMultiOutput.from_pretrained("facebook/audiobox-aesthetics")
# finetune the model
...
# finished finetuning & upload the model
model.push_to_hub("<your_hf_username>/<your_hf_repo>")
```

CLI inference

```
audio-aes input.jsonl --batch-size 100 > output.jsonl
```

How to run prediction from Python script or interpreter

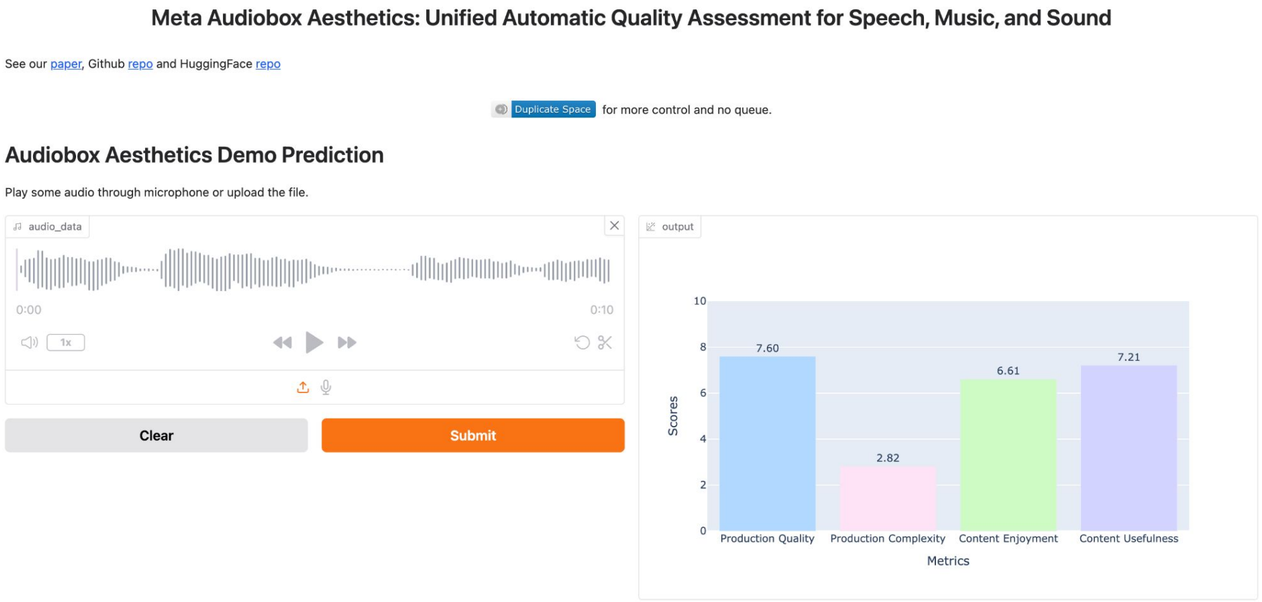
1. Infer from file path

```
from audiobox_aesthetics.infer import initialize_predictor
predictor = initialize_predictor()
predictor.forward([{"path":"/path/to/a.wav"}, {"path":"/path/to/b.flac"}])
```

2. Infer from torch tensor

```
from audiobox_aesthetics.infer import initialize_predictor
predictor = initialize_predictor()
wav, sr = torchaudio.load("/path/to/a.wav")
predictor.forward([{"path":wav, "sample_rate": sr}])
```

HF Spaces Demo



https://huggingface.co/spaces/facebook/audiobox-aesthetics

Limitation of current model

- Sampling rate:
  - Our current model finetuned on top of WavLM architecture, which always forced to resample the input audio to 16 kHz.
  - However, during annotation, we annotated audio on their original sample rate (between 8 - 48kHz).
  - This would ignore some high-frequency details which may related to the audio quality.
- Mono channel:
  - Similar to issue above, we always combine multi-channel audio into mono for our model training and inference.
  - But we use original number of channel during data annotation.
  - For certain domain like music, this might have some effects on content enjoyment and quality as well.

AudioMOS Challenge - Track 2

https://sites.google.com/view/voicemos-challenge/audiomos-challenge-2025

Motivation

- In the our previous work, we train our model on real audio data.
- However, since one of our goal is to automate generative model evaluation, we also interested to observed our model performance and improve on top of it on synthetic data.
- In this challenge, we setup a challenging task where:
  - Limited amount of training data (AES-Nature, total ~3000 samples, each samples around 10-30 seconds, 10 annotation for each samples & axis).
  - Domain mismatch between training and testing data.

END OF PRESENTATION