Choose the challenge **Background Music Generation**

Overview Discussion Leaderboard

Overview

Description

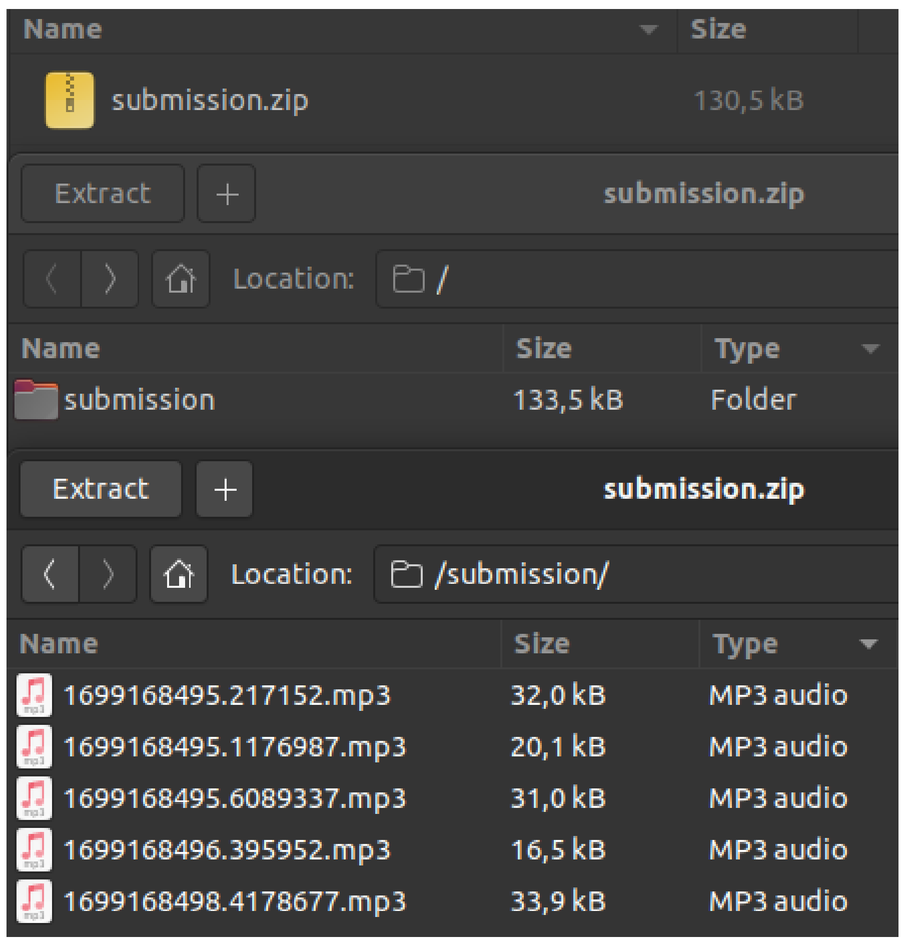
Evaluation

Data

Rule

***Submission format:**
You must provide a submission.zip file that includes 10s 16khz sample rate MP3 audio files named according to the names defined in the public.json file (or private.json for the private leaderboard)

Sample Submission: [Download]



Name	Size
submission.zip	130,5 kB

Location: /

Name	Size	Type
submission	133,5 kB	Folder

Location: /submission/

Name	Size	Type
1699168495.217152.mp3	32,0 kB	MP3 audio
1699168495.1176987.mp3	20,1 kB	MP3 audio
1699168495.6089337.mp3	31,0 kB	MP3 audio
1699168496.395952.mp3	16,5 kB	MP3 audio
1699168498.4178677.mp3	33,9 kB	MP3 audio

*** Evaluation metrics**

We will assess the generated music based on two key metrics: the CLAP score [1] and the Fréchet Audio Distance (FAD) score [2].

The CLAP score measures the resemblance between the provided track description and the generated audio, providing a quantifiable evaluation of audio-text alignment. To compute this score, a two-step process is employed: text and audio embeddings are initially generated separately using two independent encoders, and then the cosine similarity between these two embeddings is

calculated. The resulting score falls within the [0, 1] range, where a higher score indicates a superior model performance. In this challenge, we employ the official pretrained CLAP Model 6 for this computation.

FAD assesses the quality of a generative model by comparing the distance between the distribution of the embedded set of generated audios and that of the ground truth audios. FAD quantifies the similarity in distribution between the generated audio and the ground truth audio, with a lower FAD value signifying a superior model. To maintain consistency with CLAP, we convert the distance value into a similarity score, referred to as Fréchet Audio Similarity (FAS), using the following formula:

FAS = 1 / (1 + FAD)

In this challenge, we utilize the official Tensorflow implementation with VGGish Model 5 to compute the FAD value, subsequently transforming it into FAS as per the provided formula.

The ultimate score is a combination of CLAP and FAS, known as the CLAS score. The CLAS score is calculated through a linear combination of CLAP and FAS, assigning equal weight, effectively averaging the two scores. The team with the highest CLAS score will be declared the winner.

CLAS = (CLAP + FAS) /2.

*** Time constraint**

The inference time is determined on a server with the following hardware specifications:

- CPU: Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz

- RAM: 64 GB

- GPU: GeForce GTX 3090

The essential requirement for the AI model is to operate at real-time speed, which implies it should generate a 10-second music audio within a maximum timeframe of 10 seconds. If the model exceeds the stipulated time limit, it will be forcibly halted, and the evaluation will focus solely on the completed portion of the generated audio.

*** References:**

[1] https://arxiv.org/pdf/2206.04769.pdf

[2] https://arxiv.org/abs/1812.08466