UNIVERSITÉ LAVAL

Fine-Tuning MusicGen on a complex niche musical style for Text-Based Music Generation

MLSP 2025

METAIS Robin - LANIER Théo

Why this project?

- MusicGen is already proeficient in a lot of “simple” musical styles, thanks to it’s pretraining on many different tracks

- However, for some very specific niche musical style, MusicGen have some issues

UNIVERSITÉ LAVAL

Why this project?

- Moreover, MusicGen have some hard time to generate “complex” music, with a large amount of instruments (symphony orchestra)

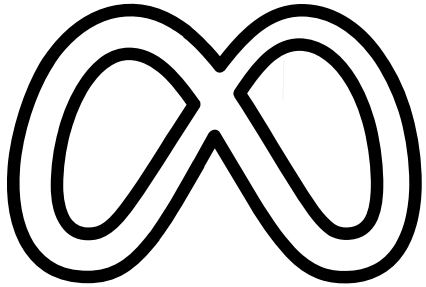
- Well, we willl combine these 2 issues, and try to fix them !

UNIVERSITÉ LAVAL

What is MusicGen ?

MusicGen

PyTorch | FlashAttention | SDPA



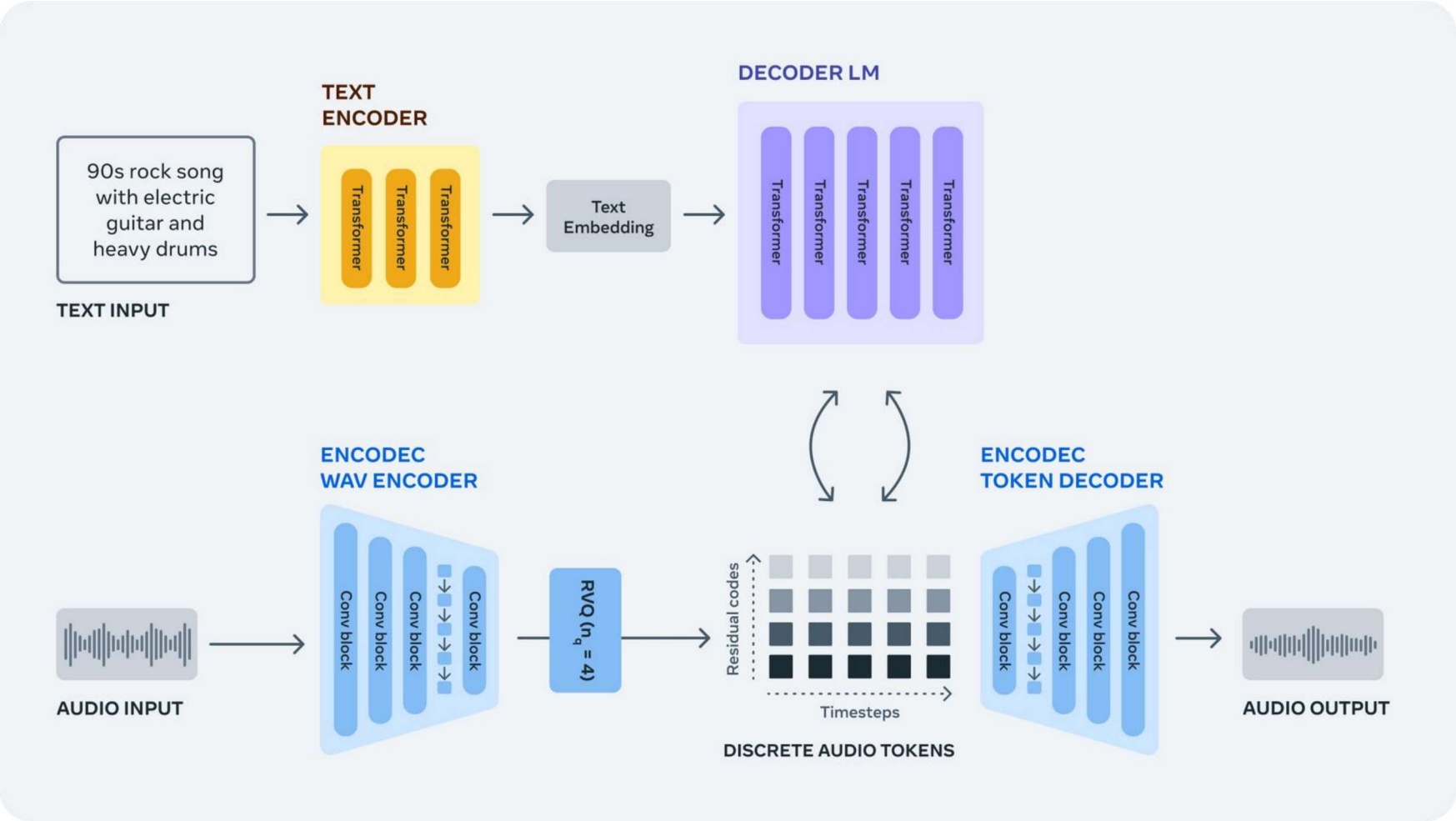
Overview

The MusicGen model was proposed in the paper Simple and Controllable Music Generation by Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi and Alexandre Défossez.

MusicGen is a single stage auto-regressive Transformer model capable of generating high-quality music samples conditioned on text descriptions or audio prompts. The text descriptions are passed through a frozen text encoder model to obtain a sequence of hidden-state representations. MusicGen is then trained to predict discrete audio tokens, or *audio codes*, conditioned on these hidden-states. These audio tokens are then decoded using an audio compression model, such as EnCodec, to recover the audio waveform.

UNIVERSITÉ LAVAL

What is MusicGen ?



UNIVERSITÉ LAVAL

Which musical style should we pick ?

First try : LoFI

(the one on our project proposal)

✓ Huge dataset with labels on HuggingFace

✗ Doubts about MusicGen being already able to generate good LoFi tracks without any specific training

UNIVERSITÉ LAVAL

Which musical style should we pick ?

First try : LoFI

(the one on our project proposal)

Result on this prompt, with MusicGen-small:

lofi hip hop beat, chill, study music, piano, rain sounds

UNIVERSITÉ LAVAL

Which musical style should we pick ?

First try : LoFI

(the one on our project proposal)

❌ Already fairly good

❌ No real room for improvement

UNIVERSITÉ LAVAL

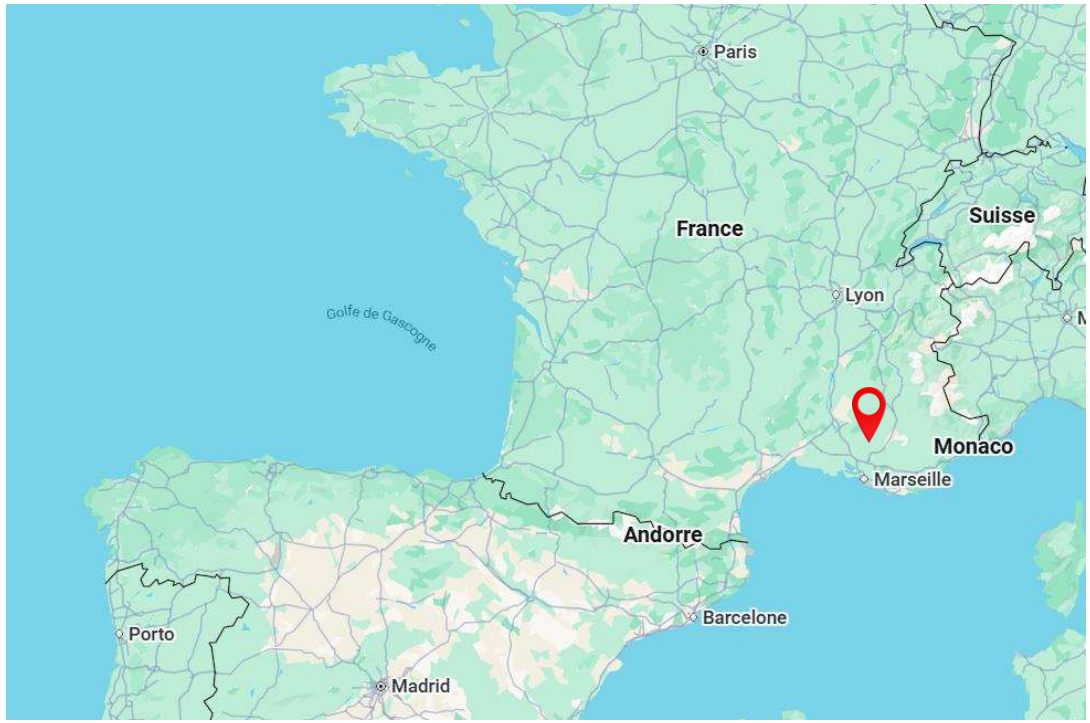
Which musical style should we pick ?

Second try : Folk music

UNIVERSITÉ LAVAL

Which musical style should we pick ?

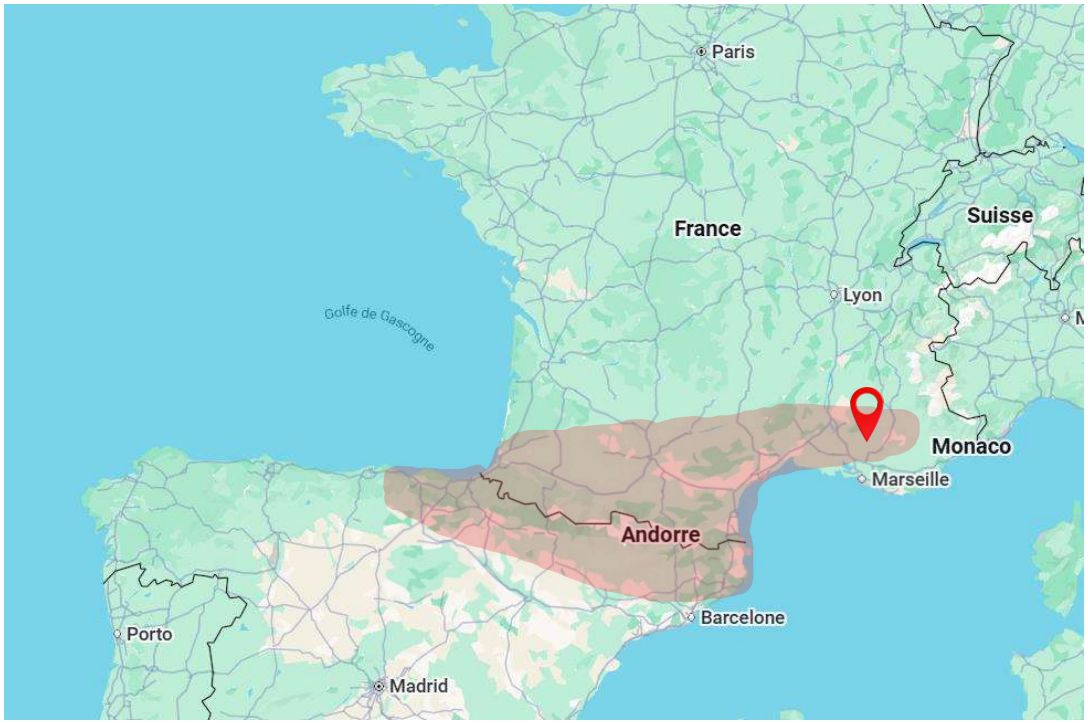
Second try : Folk music



UNIVERSITÉ LAVAL

Which musical style should we pick ?

Second try : Folk music



UNIVERSITÉ LAVAL

Which musical style should we pick ?

Second try : Folk music

Peña (or banda)

UNIVERSITÉ LAVAL

Which musical style should we pick ?

Second try : Folk music
(peña music, pasodobles)

A short example !

UNIVERSITÉ LAVAL

Which musical style should we pick ?

Second try : pasodoble music style

Pasodoble (Spanish: *double step*) is a fast-paced Spanish military march used by infantry troops. Its speed allowed troops to give 120 steps per minute (double the average of a regular unit, hence its name). This often was accompanied by a marching band, and as a result of that, the military march gave rise to a modern Spanish musical genre and partner dance form. Both voice and instruments, as well as the dance then began to develop and be practiced independently of marches, and also gained association with bullfighting due to the genre being popular as an instrumental music performed during bullfights.

UNIVERSITÉ LAVAL

Which musical style should we pick ?

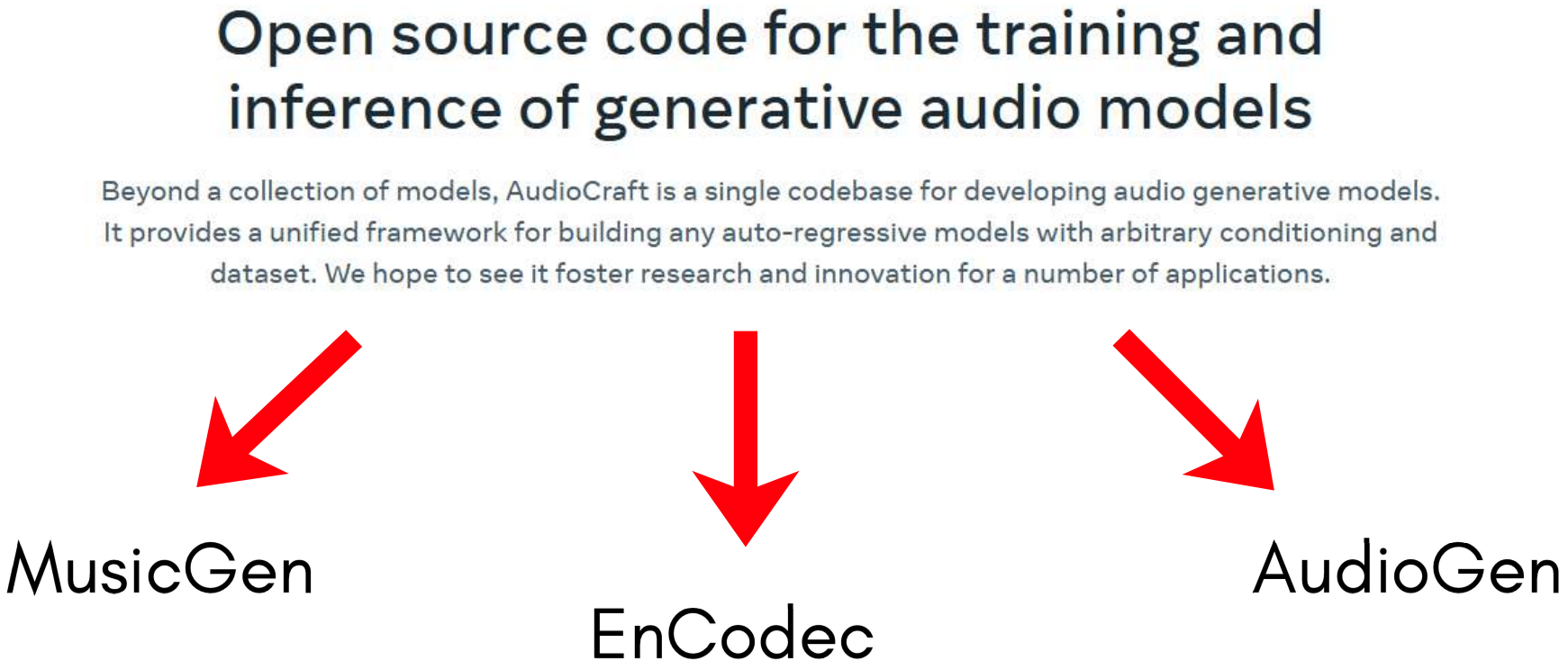
Second try : pasodoble music style

Result on this prompt, with MusicGen-small:

Traditional Paso Doble music for bullfighting, festive Spanish and South-West France brass band (Banda), majestic trumpets and trombones

UNIVERSITÉ LAVAL

What is AudioCraft ?



UNIVERSITÉ LAVAL

Step 1 - Collect chunks from our dataset

Not enough data (yet) - We have 60 extracts (30 seconds each)

UNIVERSITÉ LAVAL

Step 1 - Collect chunks from our dataset

Not enough data (yet) - We have 60 extracts (30 seconds each)

Step 2 - Let's put a label for each chunks

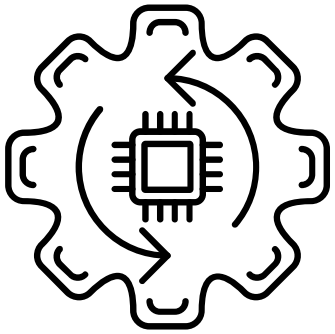
For now, our focus was on generating satisfying results, without prompt.

We attribute the same prompt for each chunks :

Peña music, pasodoble, South West France feria, trumpet, party, banda style

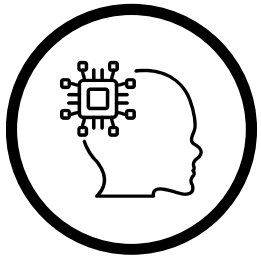
UNIVERSITÉ LAVAL

Step 3 - Let's train !



30 epochs

4 batchs



MusicGen - small

2-3 hours

UNIVERSITÉ LAVAL

Step 4 - Let's generate !

High-Guidance Specialization Strategy

```
model.set_generation_params(
        duration=30,
        temperature=1.0,
        top_k=250,
        cfg_coef=10
    )
```

The cfg_coef parameter controls how strictly the model adheres to the provided text prompt during generation.

UNIVERSITÉ LAVAL

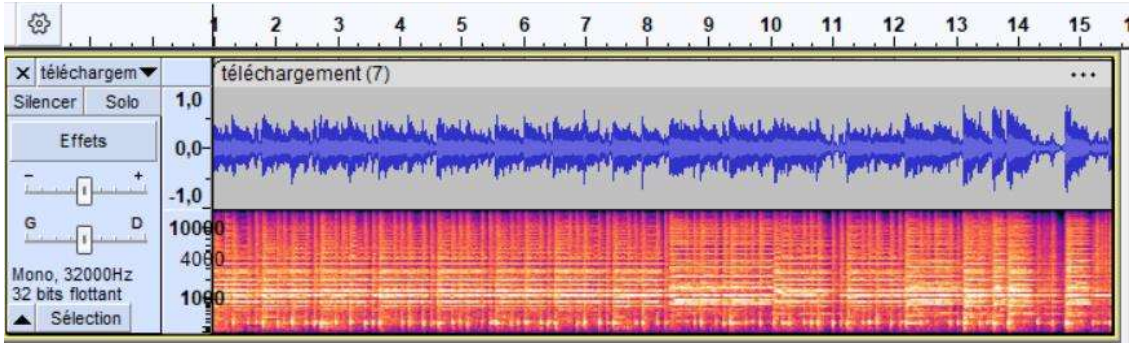
Step 4 - Let's generate !

Time to listen :

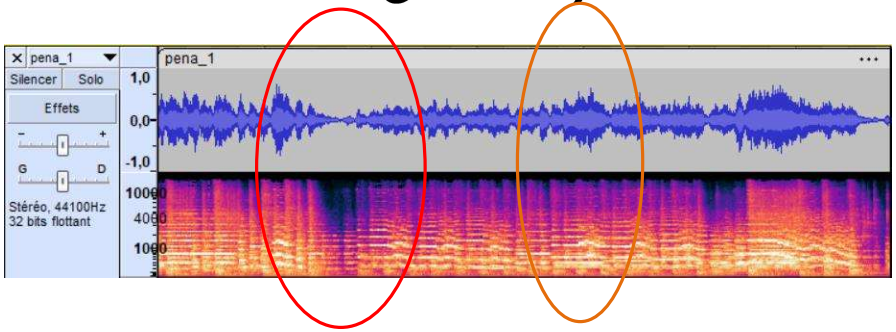
It's not perfect, but it's already better than the base model !

UNIVERSITÉ LAVAL

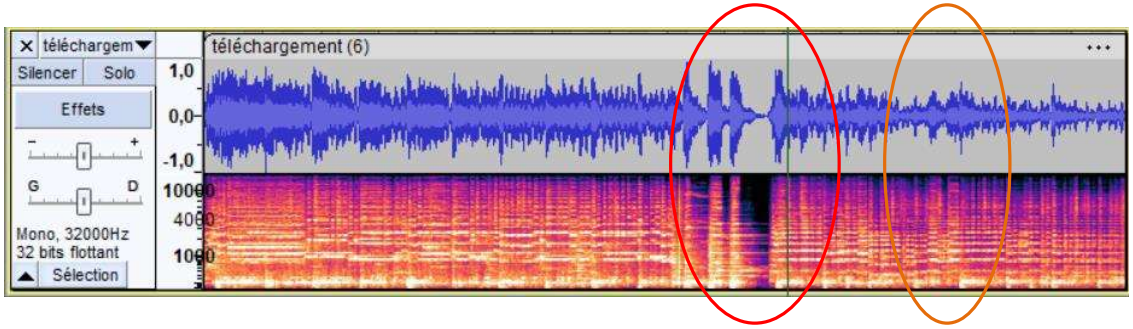
MusicGen small without training



Target style



MusicGen small with training



UNIVERSITÉ LAVAL

How to improve :

- *First of all, we need DATA*
  - *With more data, we will be able to reduce the cfg_parameter, and then the audio quality will be better*

UNIVERSITÉ LAVAL

How to improve :

- *Then, we need to labelize*

Auto-labeling ? *By hand ?*

The musical style is very similar, so we might only put the BPM in the label

UNIVERSITÉ LAVAL

How to improve :

- *Big evaluation part*

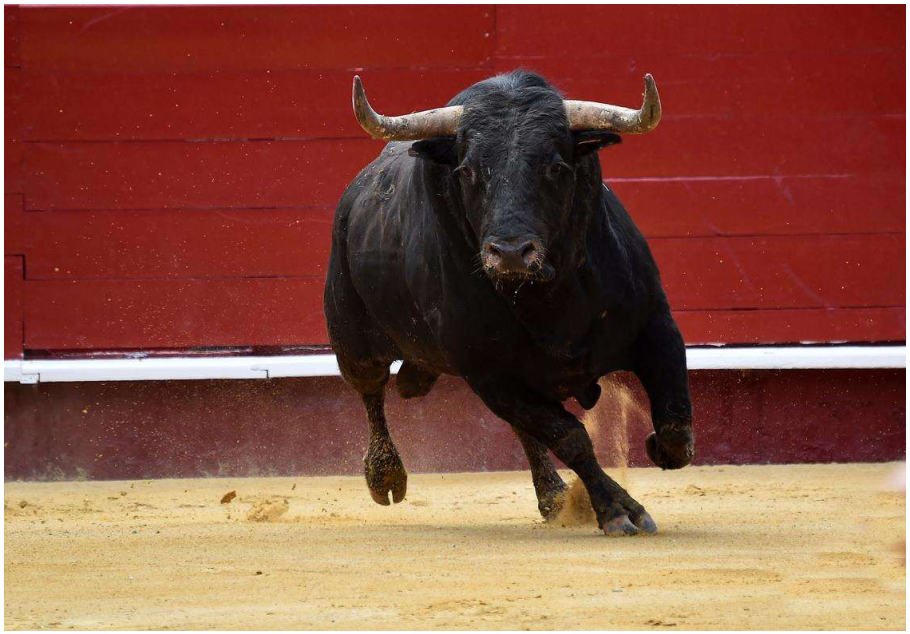
1: Fréchet Audio Distance

2: Clap Score

Subjective evaluation with confirmed musicians in two parts :

A/B Test *Mean opinion score*

UNIVERSITÉ LAVAL



Any questions ?

UNIVERSITÉ LAVAL