UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO MATEMATICA

CSC Centro di Sonologia Computazionale

VARIATIONAL AUTOENCODERS AND THEIR USE FOR SOUND GENERATION

CANDIDATE: CHIARA DE LUCA

SUPERVISIOR: PROF. SERGIO CANAZZA TARGON

CO-SUPERVISOR: PROF. ANTONIO RODÀ

RESEARCH GOALS:

DEEP LEARNING EXPLORATION → HOW TO GENERATE A GOOD SOUND USING VAE?

MUSICAL EXPLORATION → CAN THE SOUND SPACE BE EXPLORED USING VAE?

MUSIC GENERATION: AN OVERVIEW

TRENDS

- COMBINING ARCHITECTURES
- ASSISTANT SYSTEM

METHODOLOGIES

VARIATIONAL AUTOENCODERS	RECURRENT NEURAL NETWORK
RESTRICTED BOLTZMANN MACHINE	GENERATIVE ADVERSARIAL NETWORK

LIMITS

CONTROL	INTERACTIVITY
CREATIVITY	STRUCTURE

DATASETS

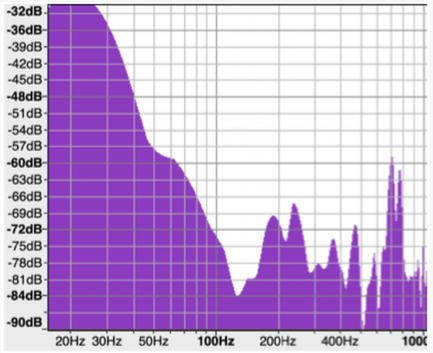
NON-HARMONIC

- BACKGROUND SOUNDS
- PERCUSSIVE SOUNDS

Some examples from the dataset:

- Deep
- Sizziness
- Glitchy

- 11000 clips
- 4 seconds per clip
- 21 classes



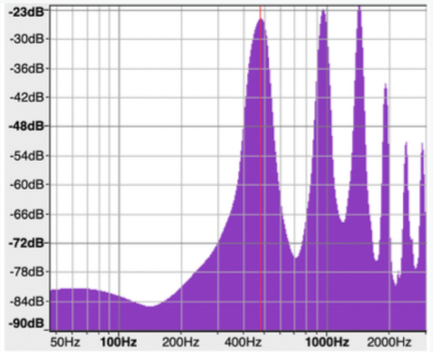
HARMONIC

- MUSICAL INSTRUMENTS

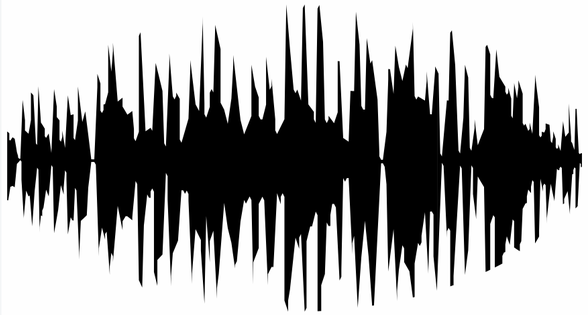
Some examples from the dataset:

- Violin
- Clarinet
- Cello

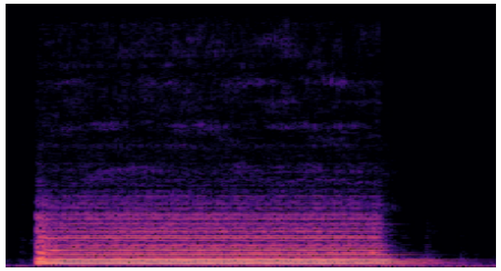
- 2000 clips
- 3-8 seconds per clip
- 7 classes



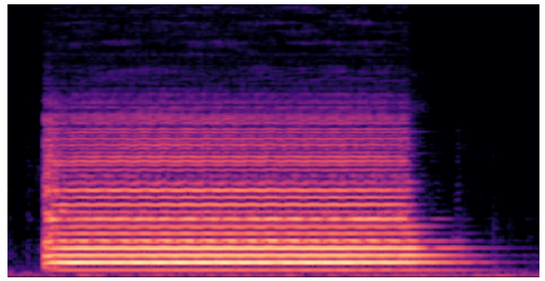
AUDIO REPRESENTATION



WAVEFORM



SPECTROGRAM



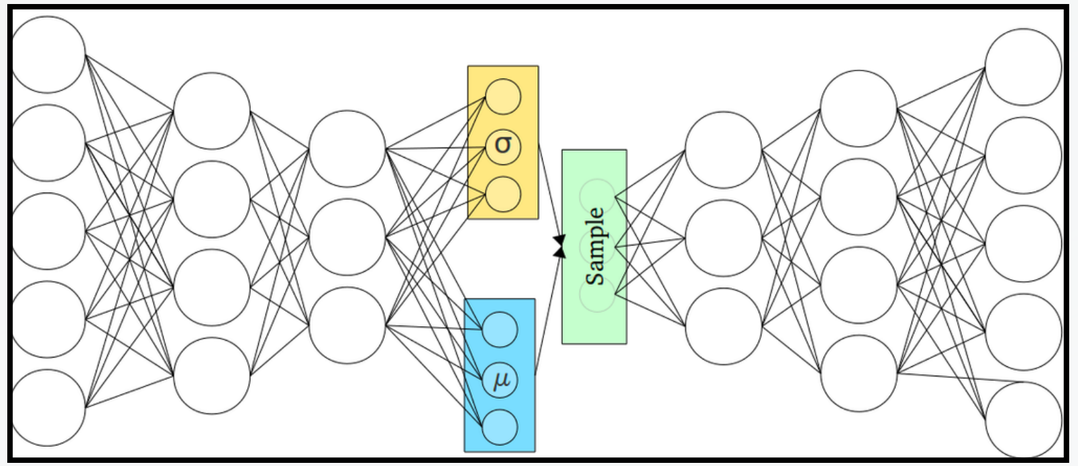
MEL SPECTROGRAM

VARIATIONAL AUTOENCODER

Goal

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{reconstruction}} - \beta \cdot \underbrace{D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z})\right]}_{\text{regularization}}$$

Structure



Why it could be a good choice

1. Compared to Autoencoder?
2. Compared to GAN?

BATCH SIZE	**64, 128, 256**
LEARNING RATE	**1e-4, 5e-4. 1e-3**
OPTIMIZER	**ADAM**
EPOCHS	**100, 250, 500, 1000**

LOSS

$$\boxed{\mathrm{RL} + \mathrm{KLD}}$$

$$\frac{1}{N}\sum_{i=1}^{N}\left\|y_{\text{target}}^{(i)} - y_{\text{predicted}}^{(i)}\right\|_2^2$$

$$-\frac{1}{2}\sum_{i=1}^{N}\left(1 + \log(\text{variance}^{(i)}) - \mu^{(i)2} - \exp(\text{variance}^{(i)})\right)$$

NON HARMONIC

ENCODING: two hidden layers
DECODING: two hidden layers
HIDDEN UNITS: 256
LATENT SPACE DIM: 128
ACTIVATION FUNCT: ReLU, Sigmoid

LEARNING RATE: 0.0001
BATCH SIZE: 128
EPOCHS: 250

TIME PER EPOCH: 6s
LOSS: 0.0008

HARMONIC

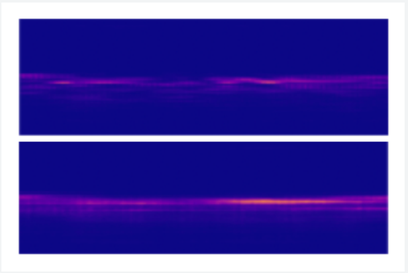
ENCODING: two hidden layers
DECODING: one hidden layer
HIDDEN UNITS: 512
LATENT SPACE DIM: 256
ACTIVATION FUNCT: ReLU, Sigmoid

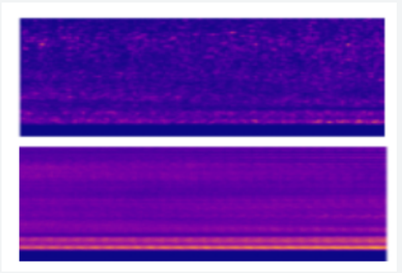
LEARNING RATE: 0.0001
BATCH SIZE: 64
EPOCHS: 100

TIME PER EPOCH: 8s
LOSS: 0.0005

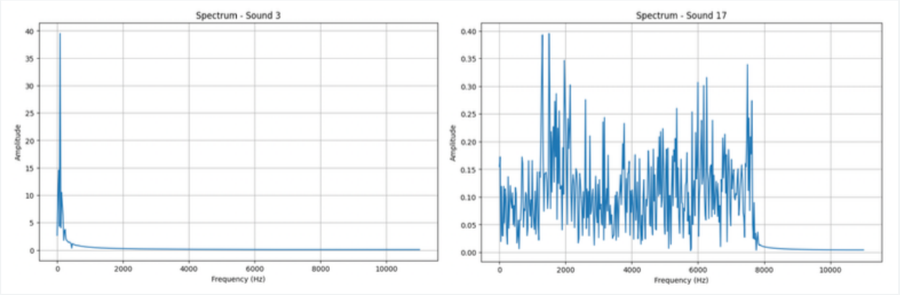
GENERAL CONSIDERATIONS

NON HARMONIC

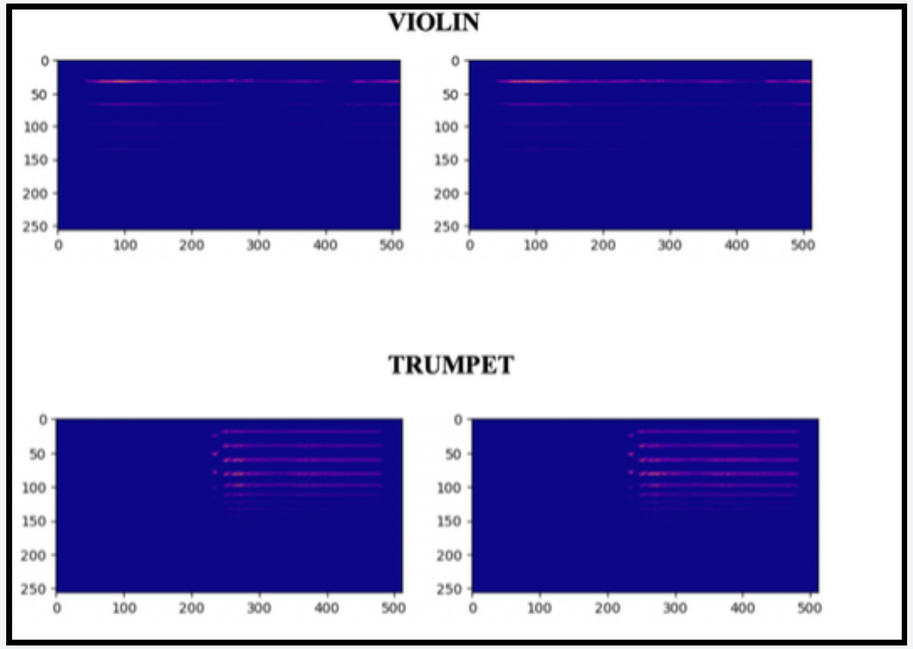




SOUND COMPLEXITY
VS
LATENT SPACE DIMENSIONALITY



HARMONIC

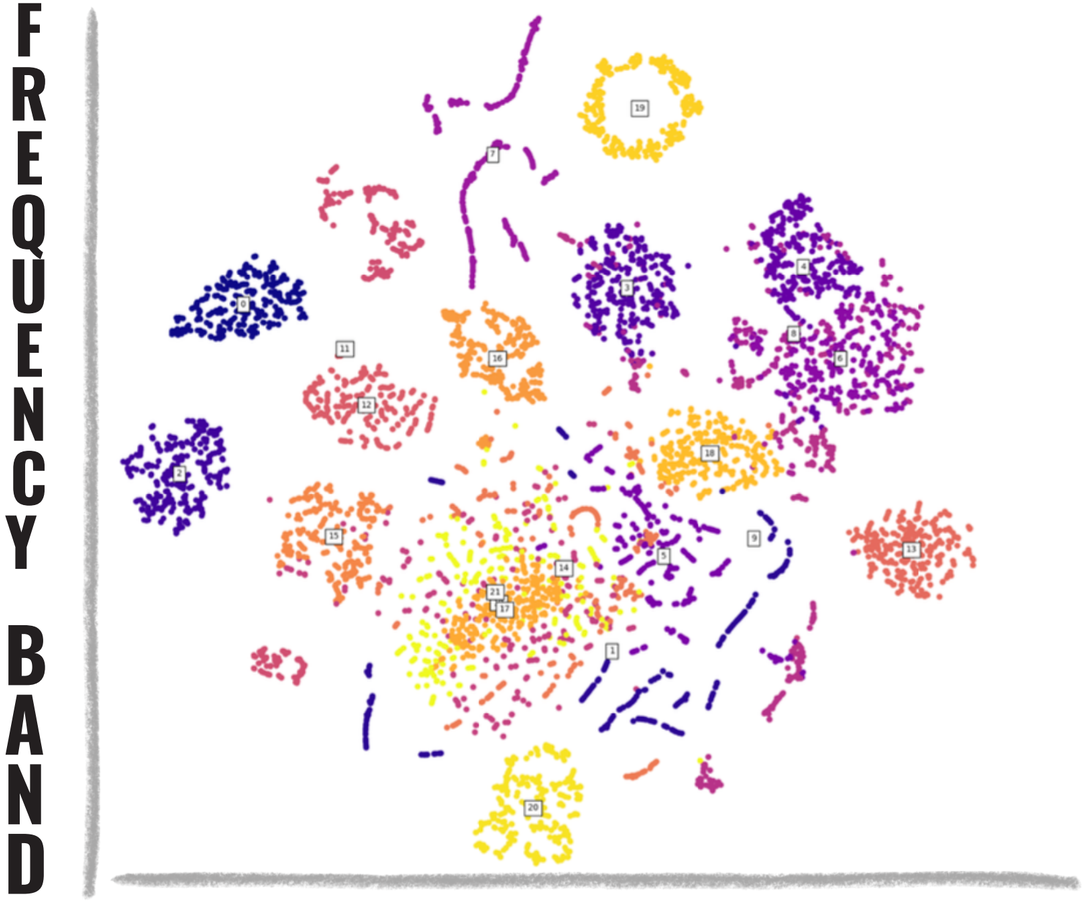


TIMBRE	INTENSITY
PITCH	ENVELOPE

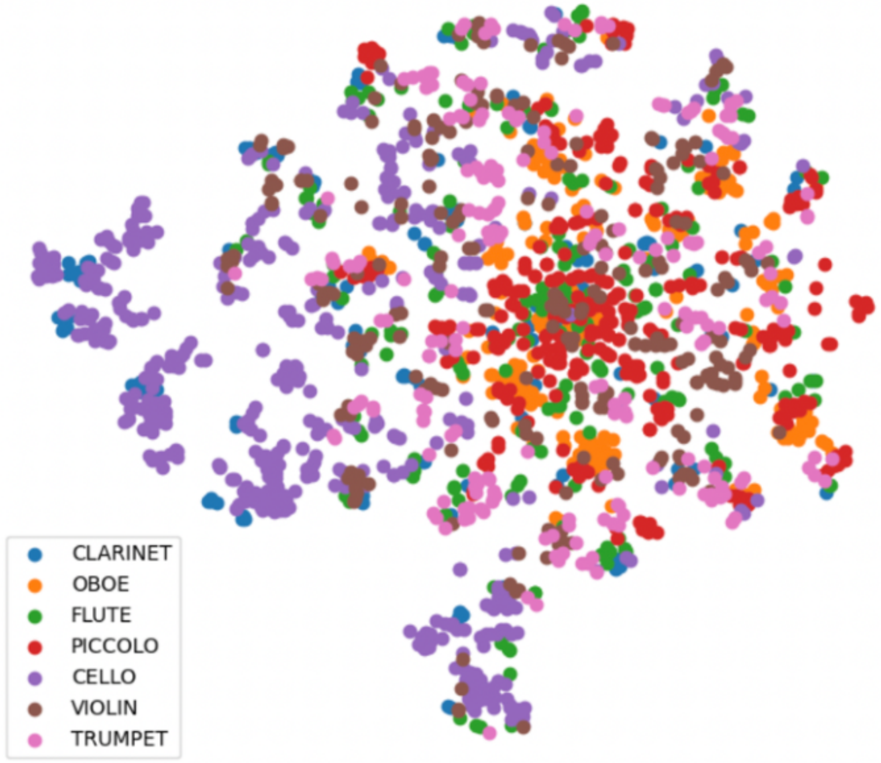
SOUND SPACE

A T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING VISUALIZATION

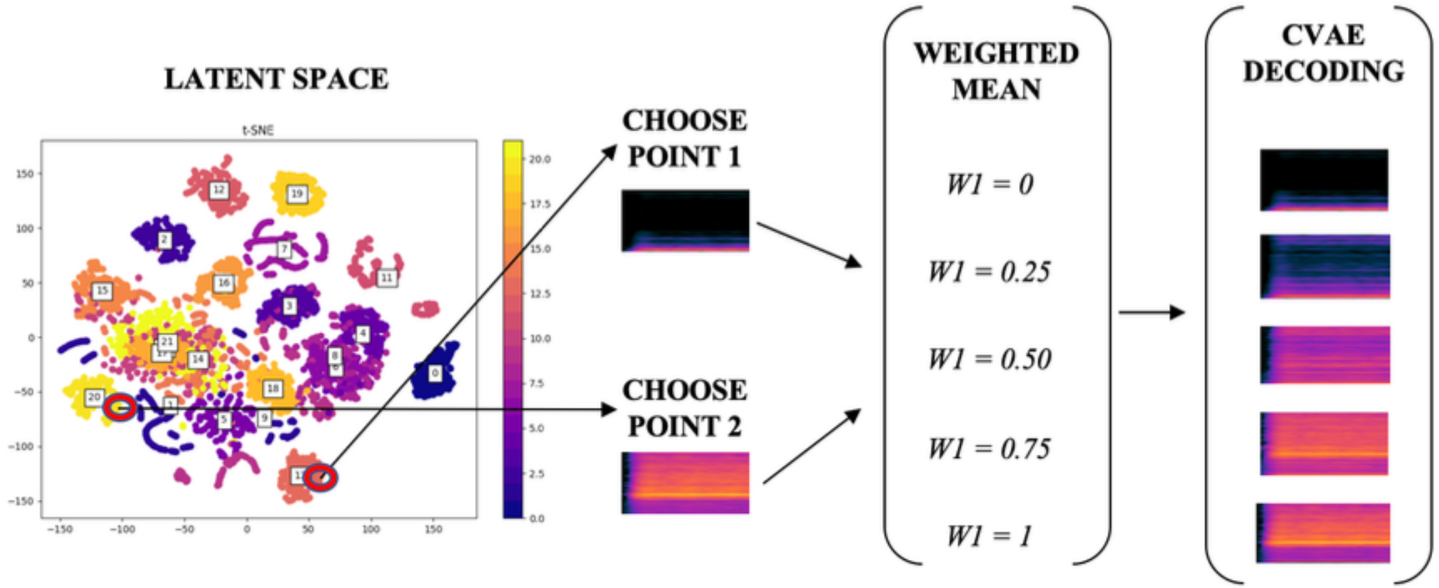
NON HARMONIC

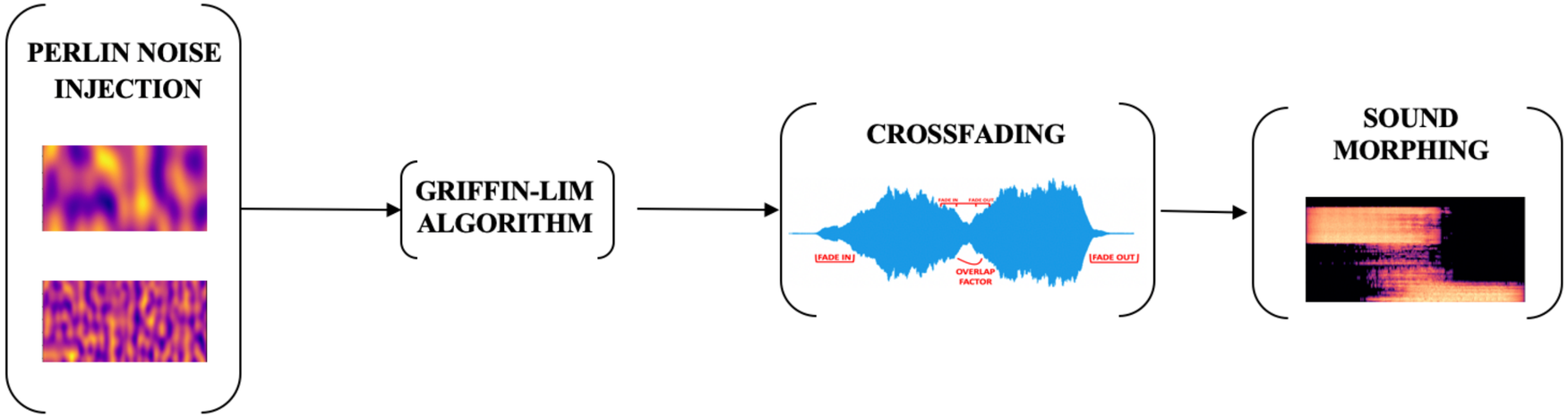


HARMONIC

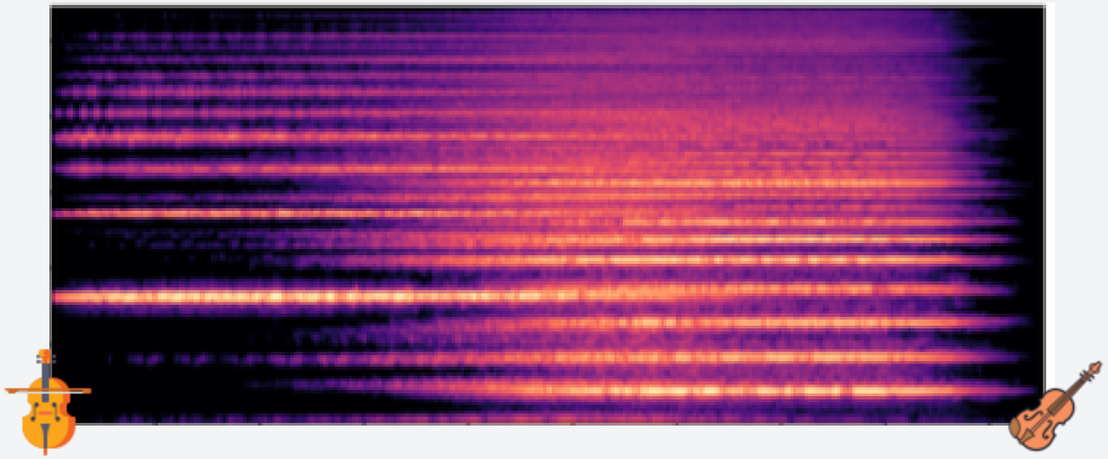


SOUND MORPHING





AN EXAMPLE

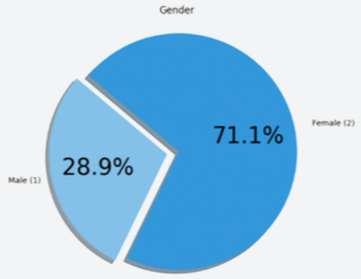


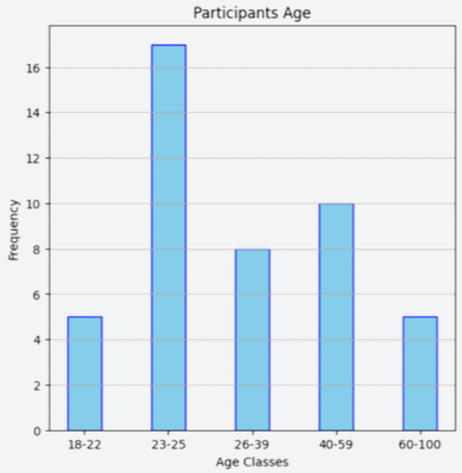
GENERAL CONSIDERATIONS

ADVANTAGES	DISADVANTAGES
HIGH LEVEL OF CONTROL	PARAMETERS AND HYPERPARAMETERS: TOO MUCH CONTROL
DIFFERENT COMBINATIONS WITH FLEXIBILITY	SMOOTHNESS DEPENDS TOO MUCH ON THE DISTANCE BETWEEN POINTS IN LS

SURVEY

PARTICIPANTS



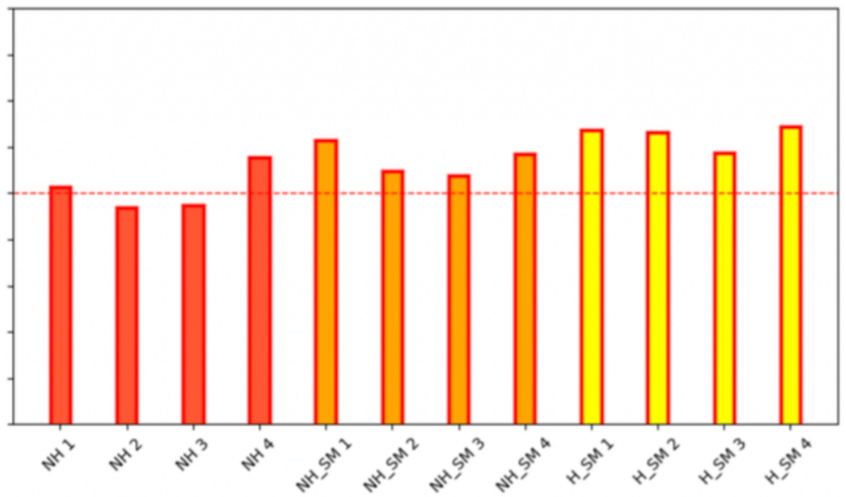


	hearing problems
yes	3
no	42

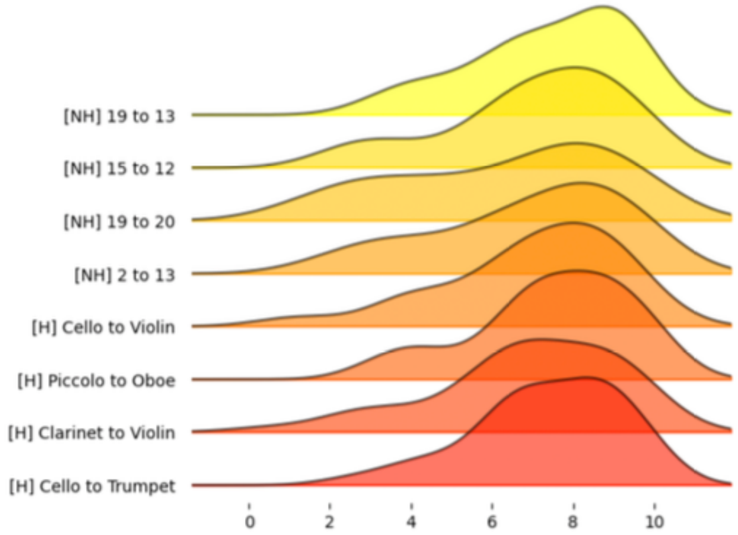
EVALUATED ASPECTS

	NON-HARMONIC (single sounds)	NON-HARMONIC SOUND MORPHING	HARMONIC SOUND MORPHING
QUALITY	✓	✓	✓
CLASS	✓		✓
SMOOTHNESS		✓	✓

QUALITY



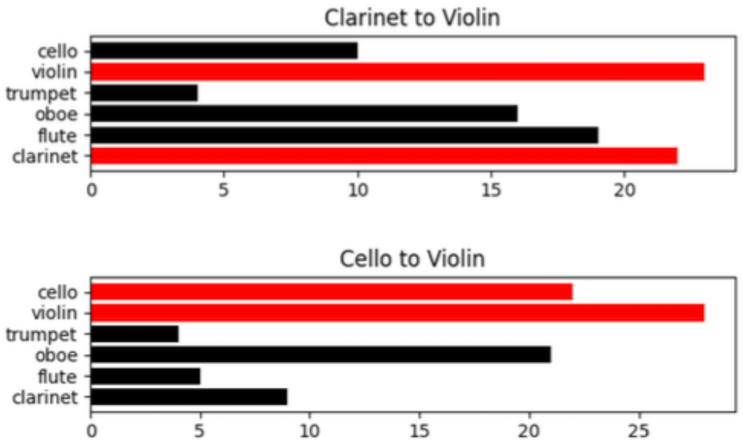
SMOOTHNESS



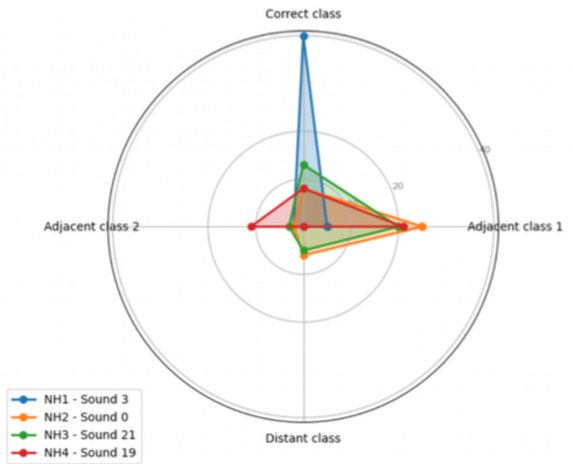
SHAPIRO-WILK
LEVENE
KRUSKAL-WALLIS
DUNN

STATISTICAL EVALUATION: USED TESTS

HARMONIC CLASSIFICATION



NON HARMONIC CLASSIFICATION



CHI-SQUARE

CONCLUSION

Generated sounds have a good quality. They are clearly recognizable and classificable.

VAE is a good choice. Sounds are good, and the level of control is high.

Generated sound morphings have good smoothness. The sound space is highly explorable.

FUTURE DIRECTIONS

Real-time Infinite and Controlled Music Generation

AI Tool for Musicians and Sound Designers

SoundFood Project at