Exploring Polyphonic Music Accompaniment Generation using Generative Adversarial Networks

Danae Charitou[3], Christos Garoufis[1,2,3], Athanasia Zlatintsi[1,2,3], Petros Maragos[2,3]

danaecharitou@gmail.com, christos.garoufis@athenarc.gr, athanasia.zlatintsi@athenarc.gr, maragos@cs.ntua.gr

[1]Institute of Language and Speech Proc., Athena Research Center, Athens, Greece
[2]Institute of Robotics, Athena Research Center, Athens, Greece
[3]School of ECE, National Technical University of Athens, Athens, Greece

ATHENA
Research & Innovation
Information Technologies

Overview

Motivation & Goal: designing a generative framework for **symbolic multi-track music generation** that is **structurally flexible** and **adaptable** to different musical configurations:

- **Unconditional Generation**: Generation of multi-track symbolic music from scratch.
- **Conditional Generation**: Generate the multi-track accompaniment, given a single track.

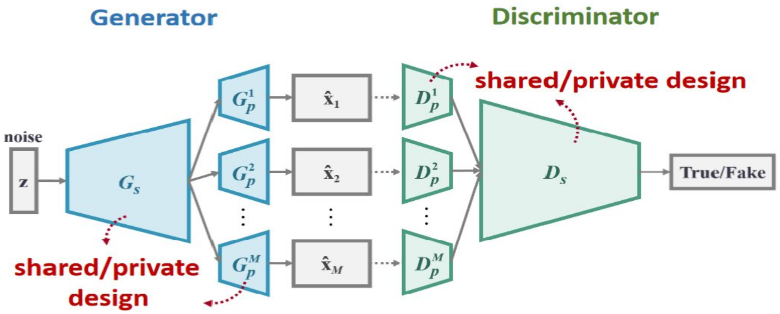
Contributions:

- Structural improvements upon the baseline **unconditional** MuseGAN architecture.
- Extension of this framework to a **cooperative human-AI setup** for the generation of polyphonic accompaniments to user-defined tracks.
- Experimental validation through both **objective** and **subjective** evaluation.

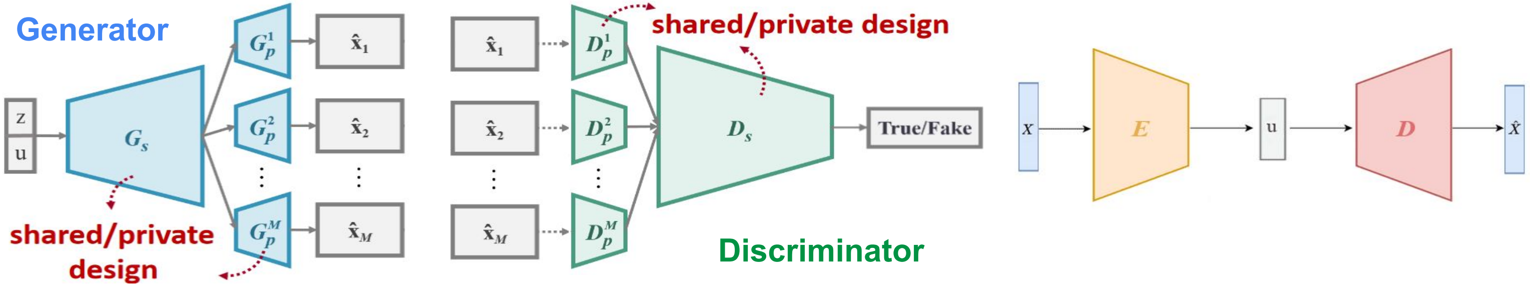
Methodology

Unconditional model: a GAN model that generates multi-track pianorolls:

- **shared-private** design for both Generator and Discriminator.
- convolutional layers developed with respect to **tonal/rhythmic parameters**.



Conditional model: Additionally incorporates an encoder to create an embedding for the input accompaniment, used by the generator as an additional input.



Experimental Setup

Data format: Multi-track **pianorolls** (binary matrices, rows $\longleftrightarrow$ notes, columns $\longleftrightarrow$ timesteps)

- Five tracks: Bass (B), Drums (D), Guitar (G), Piano (P), Strings (S)
- Lakh Pianoroll Dataset

Training: Using a Wasserstein-GP GAN loss: $\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_d}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[D(G(\mathbf{z}))]$

$$+ \mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{\mathbf{x}}}}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2]$$

Evaluation Protocol:

- **Objective Evaluation Metrics:** Empty Bars (**EB**), Used Pitch Classes (**UPC**), Qualified Notes (**QN**), Drum Pattern (**DP**), Tonal Distance (**TD**), Used Pitches (**UP**), Scale Ratio (**SR**), Polyphonic Rate (**PR**).
- **Subjective Evaluation:** Listening test (40 participants)
  - **Criteria**: Music Naturalness, Harmonic Consistency, Musical Coherence

Results: Objective Evaluation

Unconditional Setup:

- Our framework **outperforms** almost all baseline variations on fragmentation-related metrics (QN, DP) and **surpasses** all baseline architectures (generating harmonic samples).

Conditional Setup:

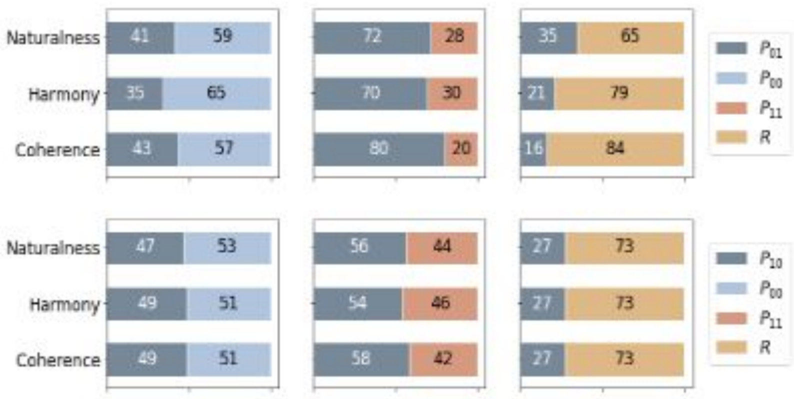
- Experimented with both piano and guitar as **condition instruments**, as well as the encoder **training scheme** (joint or 2-stage) and adding a **local discriminator**.
    - The 2-stage training scheme mostly benefits **empty bar** rate (EB).
    - The inclusion of a local discriminator helps in modeling texture elements such as **polyphonic rate** (PR).

Results: Subjective Evaluation

- **Outperforming** MuseGAN in the unconditional setup.
- Mainly obtaining improved results using either the **local discriminator** or the **2-stage training scheme** (not both) in the conditional setup.

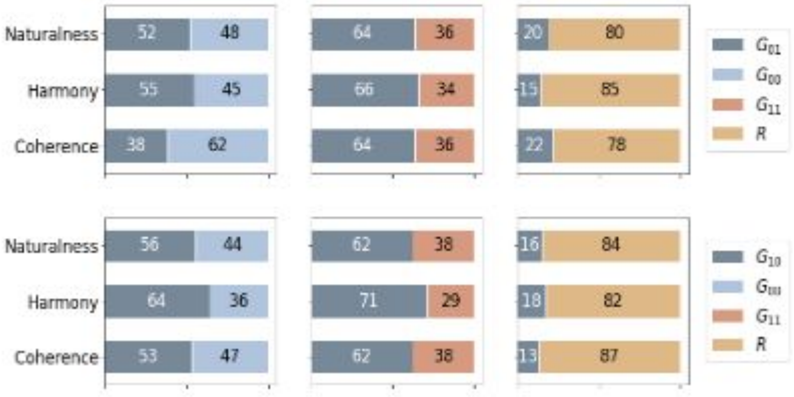
Piano - comparisons regarding:

discriminator **training scheme** **real samples**



Guitar - comparisons regarding:

discriminator **training scheme** **real samples**



Thank you for your attention!

This research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers” (Project Number: 7773). For more information: https://i-mreplay.athenarc.gr/