



Exploring Polyphonic Accompaniment Generation using Generative Adversarial Networks

D. Charitou³, C. Garoufis^{1,2,3}, A. Zlatintsi^{1,2,3} and P. Maragos^{2,3}

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

¹Institute of Language and Speech Proc., Athena Research Center, Athens, Greece

²Institute of Robotics, Athena Research Center, Athens, Greece

³School of ECE, National Technical University of Athens, Athens, Greece



For more information: <https://i-nreplay.athenarc.gr/>

1. Introduction

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:

- Proposition of structural improvements upon the **unconditional** MuseGAN architecture [1].
- Extension of this framework to a **cooperative human-AI setup** for the generation of polyphonic accompaniments to user-defined tracks:
 - Exploration of multiple **structural variants** and **training schemes**
 - Two different candidate conditional instruments: **piano** and **guitar**.
- Evaluation of the produced samples for both cases
 - **objectively**, using a set of widely used musical metrics, and
 - **subjectively**, by conducting a listening test across **40 participants**.
- The proposed modifications and experiments:
 - in the **unconditional** case lead to auditory improvements over MuseGAN, and
 - in the **conditional** case provide useful insights about the properties of the generated music.

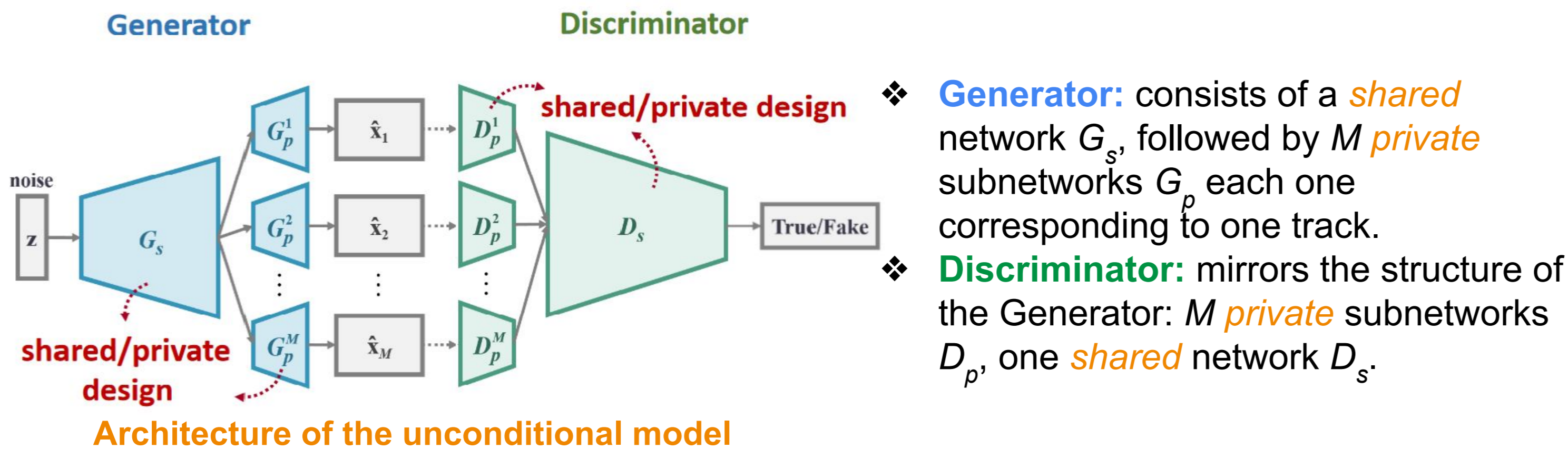
2. Methodology

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

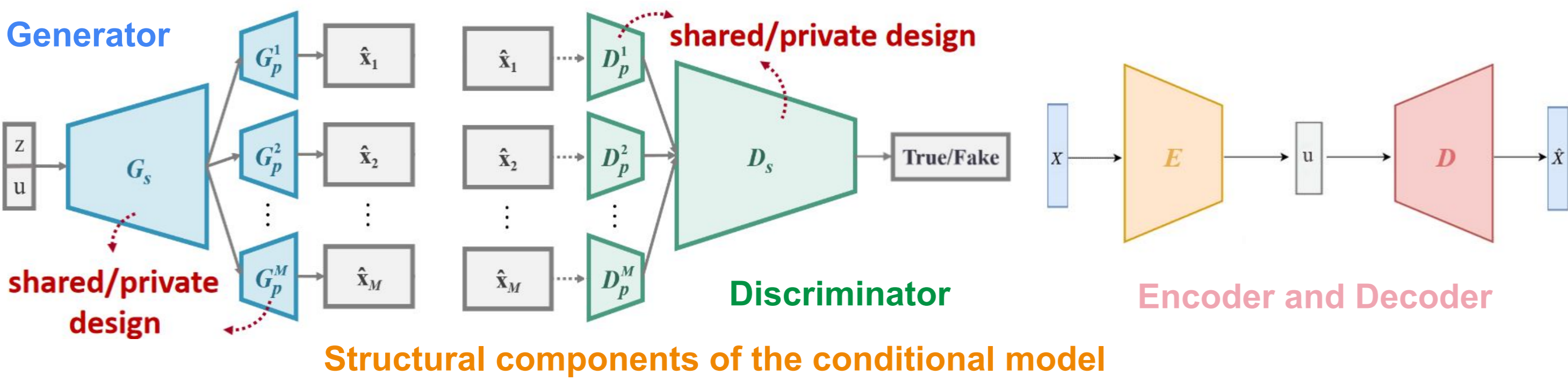
- Five tracks: Bass (B), Drums (D), Guitar (G), Piano (P), Strings (S)

Unconditional model: a GAN model that generates musical phrases of variable length

- **shared-private** design for both Generator and Discriminator [3].
- convolutional layers developed with respect to **tonal/rhythmic parameters** (i.e. bar lengths)



Conditional model: extension of the unconditional model to a co-operative setup.



Structural modifications:

- **Conditional Generator:** Generates 4 pianoroll tracks, which accompany the conditional one
 - comprises only 4 **private** subnetworks instead of 5.
- **Conditional Discriminator**
 - **Global:** incorporates 5 **private** subnetworks and evaluates all 5 tracks collectively.
 - **Local:** incorporates only 4 **private** subnetworks and evaluates only the **accompaniment** tracks as an independent musical composition.
- **Encoder/Decoder module**, produces embeddings of the conditional tracks
 - Decoder used only during training, to facilitate a reconstruction objective.

3. Experimental Setup

Dataset:

Lakh Pianoroll Dataset (174,154 multi-track **pianorolls** derived from the Lakh MIDI Dataset).

→ We employ the **LPD-5-cleansed** version, containing only the 5-track pianorolls with the higher matching confidence score to MSD entries [2], a "Rock" tag and 4/4 time signature.

Preprocessing:

- Temporal **downsampling**.
- Removal of notes outside the desired pitch range.
- Randomized selection of samples that contain an **adequate amount of notes**.
- Final dataset size: 15,600 phrases from 7,323 songs.

Training Protocol:

- Wasserstein-GAN loss function with gradient penalty: $\min_D \max_G \mathbb{E}_{x \sim p_x} [D(x)] - \mathbb{E}_{z \sim p_z} [D(G(z))] + \mathbb{E}_{z \sim p_z} [\|\nabla_x D(\hat{x})\|_2 - 1]^2$

Unconditional setup:

- The training strategy is established on consecutive interchanges between k optimization steps of the Discriminator and one optimization of the Generator.

Conditional setup:

- Updating both Global and Local Discriminators during the same training steps.
- Aggregating their feedback for the optimization of the Generator.
- **Encoder/Decoder** (2 training modes):
 - **1-phase training:** the Encoder is **trained jointly** with the GAN.
 - **2-phase training:** the Encoder is **pre-trained** along with the Decoder (with a pianoroll reconstruction MSE loss and an embedding KL divergence loss).

Musical 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**).

Configurations:

- **C1:** Pitch range: 84 notes, 24 timesteps/beat, 4 beats/bar (MuseGAN's generative setup)
- **C2:** Pitch range: 72 notes, 4 timesteps/beat, 4 beats/bar (lower resolution).

4. Objective Evaluation

Unconditional Generation (comparison to baseline/MuseGAN)

| Instruments | | EB | | | | | UPC | | | | | QN | | | | | DP | TD (↓) | | | | |
|---------------|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|--------|------|------|------|--|
| | | B | D | G | P | S | B | G | P | S | B | G | P | S | D | B-G | B-S | B-P | G-S | G-P | S-P | |
| training data | baseline | 8.06 | 8.06 | 19.4 | 24.8 | 10.1 | 1.71 | 3.08 | 3.28 | 3.38 | 90.0 | 81.9 | 88.4 | 89.6 | 88.6 | - | - | - | - | - | - | |
| | ours | 1.6 | 1.1 | 4.1 | 5.1 | 3.2 | 2.48 | 4.16 | 4.2 | 4.57 | 91.7 | 85.3 | 89.7 | 89.7 | 83.1 | - | - | - | - | - | - | |
| Baseline | jamming | 6.59 | 2.33 | 18.3 | 22.6 | 6.10 | 1.53 | 3.69 | 4.13 | 4.09 | 71.5 | 56.6 | 62.2 | 63.1 | 93.2 | 1.56 | 1.60 | 1.54 | 1.05 | 0.99 | 1.05 | |
| | composer | 0.01 | 28.9 | 1.34 | 0.02 | 0.01 | 2.51 | 4.20 | 4.89 | 5.19 | 49.5 | 47.4 | 49.9 | 52.5 | 75.3 | 1.37 | 1.36 | 1.30 | 0.95 | 0.98 | 0.91 | |
| | hybrid | 2.14 | 29.7 | 11.7 | 17.8 | 6.04 | 2.35 | 4.76 | 5.45 | 5.24 | 44.6 | 43.2 | 45.5 | 52.0 | 71.3 | 1.34 | 1.35 | 1.32 | 0.85 | 0.85 | 0.83 | |
| | ablated | 92.4 | 100 | 12.5 | 0.68 | 0.00 | 1.00 | 2.88 | 2.32 | 4.72 | 0.00 | 22.8 | 31.1 | 26.2 | 0.0 | - | - | - | - | - | - | |
| Ours | C_1 | 0.0 | 0.7 | 0.4 | 1.3 | 1.2 | 3.63 | 4.67 | 4.64 | 5.29 | 55.6 | 75.8 | 74.1 | 75.9 | 59.5 | 0.2 | 0.22 | 0.2 | 0.21 | 0.2 | 0.21 | |
| | C_2 | 0.3 | 0.0 | 0.9 | 1.9 | 2.1 | 2.89 | 4.4 | 4.88 | 5.14 | 59.0 | 58.2 | 57.2 | 60.8 | 79.6 | 0.86 | 0.91 | 0.9 | 0.98 | 0.99 | 0.97 | |

- Both models approximate adequately the statistics of the real distribution.
- **QN** and **DP:** our framework outperforms almost all baseline variations (colored cells).
- **TD:** C_1 surpasses all baseline architectures (generating harmonic samples)
 - **Shared-private design** helps in creating harmonically coherent tracks.
- C_2 is weaker than C_1 → **fine-grained resolution** assists in the generative process.

Conditional Generation

Piano models:

- 2-phase training (P_{10} and P_{11}) mostly benefits the **note density** (**EB**) of the generated samples.
- Bass more sparse than the original (**EB** equal to 17.4%) for P_{10}

- Local Discriminator (P_{01} and P_{11}) benefits **tonality** (**SR**, **UP**), **fragmentation** (**QN**) and **polyphonicity** (**PR**) of each track.

| Instruments | | EB | | | | | UPC | | | | | QN | | | | | PR | | | | |
|---------------------|----------|------|-----|-----|-----|-----|------|------|------|------|------|------|------|------|------|------|------|------|---|---|--|
| | | B | D | G | P | S | B | G | P | S | B | G | P | S | B | D | G | P | S | D | |
| Piano train | P_{00} | 1.6 | 1.0 | 5.0 | 5.6 | 3.7 | 2.47 | 4.09 | 4.19 | 4.5 | 91.6 | 85.6 | 90.0 | 89.7 | 2.71 | 5.68 | 5.85 | 6.71 | | | |
| | P_{10} | 1.8 | 0.9 | 4.3 | 5.2 | 3.6 | 2.47 | 4.21 | 4.14 | 4.49 | 91.8 | 87.5 | 91.6 | 90.5 | 2.7 | 5.85 | 5.84 | 6.75 | | | |
| | P_{11} | 0.6 | 0.0 | 2.2 | - | 2.4 | 2.71 | 3.95 | - | 4.33 | 51.4 | 56.5 | - | 58.9 | 2.94 | 5.79 | - | 6.28 | | | |
| | P_{10} | 0.2 | 0.0 | 1.8 | - | 1.5 | 2.57 | 4.09 | - | 4.76 | 58.2 | 56.1 | - | 61.7 | 2.94 | 5.77 | - | 7.17 | | | |
| Guitar train | G_{00} | 17.4 | 0.2 | 3.0 | - | 4.4 | 1.68 | 3.9 | - | 4.3 | 50.7 | 49.2 | - | 55.1 | 1.74 | 5.05 | - | 6.07 | | | |
| | G_{10} | 1.6 | 0.0 | 0.7 | - | 0.9 | 2.56 | 4.19 | - | 5.16 | 54.8 | 56.6 | - | 51.0 | 2.84 | 5.43 | - | 7.3 | | | |
| | G_{01} | 0.8 | 0.0 | - | 2.1 | 1.8 | 2.51 | - | 5.04 | 4.59 | 62.5 | - | 49.3 | 60.3 | 2.77 | - | 7.31 | 6.91 | | | |
| | G_{11} | 0.0 | 0.0 | - | 3.1 | 0.0 | 3.05 | - | 4.31 | 5.28 | 57.6 | - | 52.4 | 59.6 | 3.36 | - | 6.18 | 7.69 | | | |
| Local Discriminator | P_{01} | 1.6 | 0.0 | - | 1.8 | 3.5 | 2.35 | - | 4.28 | 4.01 | 50.2 | - | 59.5 | 58.6 | 2.59 | - | 6.13 | 5.88 | | | |
| | G_{11} | 0.4 | 0.2 | - | 3.3 | 0.6 | 2.32 | - | 4.62 | 4.66 | 55.6 | - | 47.8 | 57.9 | 2.46 | - | 6.4 | 6.68 | | | |

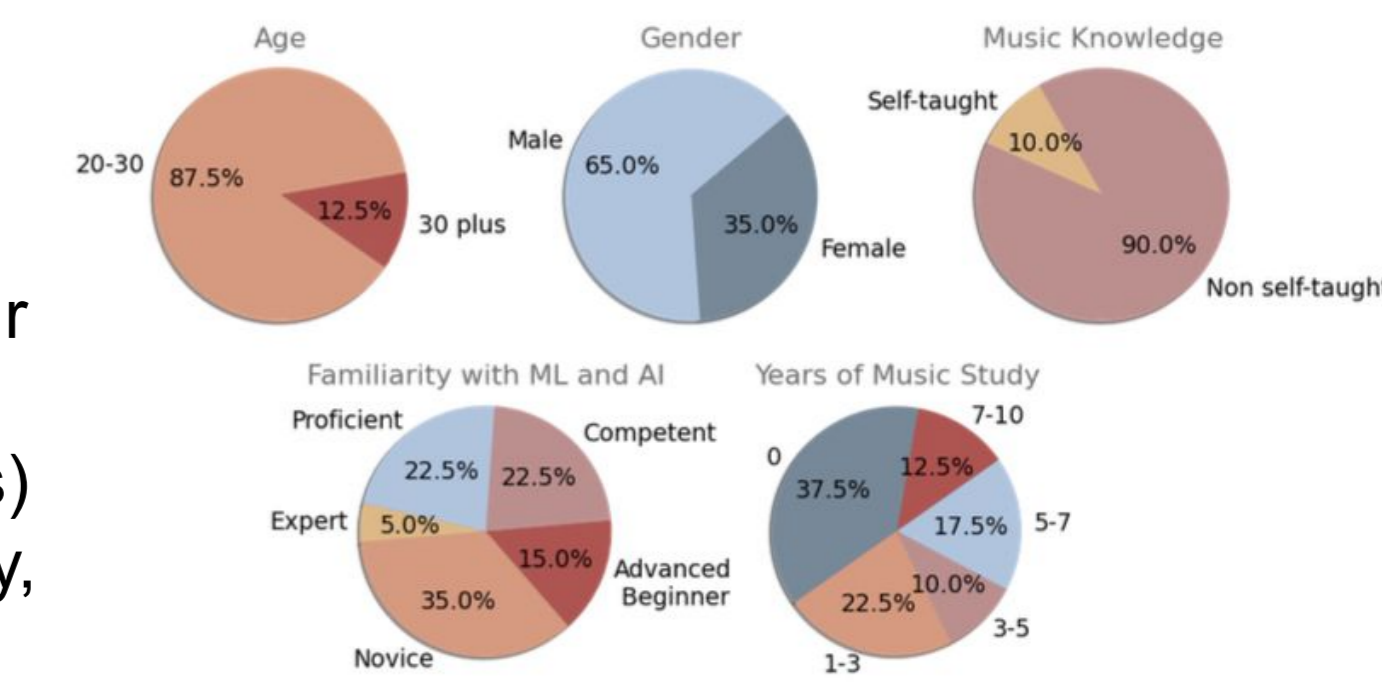
Guitar models:

- 2-phase training (G_{10} and G_{11}) benefits **note density** (**EB**) and **tonality** (**UP**, **SR**).
- Local Discriminator: stronger **harmonic relations** between the tracks (**TD**), improving also **rhythm** (**DP**) and **texture** elements such as **PR**.

| Instruments | | EB | | | | | UPC | | | | | QN | | | | | PR | | | | |
|---------------------|----------|------|-----|-----|-----|-----|------|------|------|------|------|------|------|------|------|------|------|------|---|---|--|
| | | B | D | G | P | S | B | G | P | S | B | G | P | S | B | D | G | P | S | D | |
| Piano train | P_{00} | 1.6 | 1.0 | 5.0 | 5.6 | 3.7 | 2.47 | 4.09 | 4.19 | 4.5 | 91.6 | 85.6 | 90.0 | 89.7 | 2.71 | 5.68 | 5.85 | 6.71 | | | |
| | P_{10} | 1.8 | 0.9 | 4.3 | 5.2 | 3.6 | 2.47 | 4.21 | 4.14 | 4.49 | 91.8 | 87.5 | 91.6 | 90.5 | 2.7 | 5.85 | 5.84 | 6.75 | | | |
| | P_{11} | 0.6 | 0.0 | 2.2 | - | 2.4 | 2.71 | 3.95 | - | 4.33 | 51.4 | 56.5 | - | 58.9 | 2.94 | 5.79 | - | 6.28 | | | |
| | P_{10} | 0.2 | 0.0 | 1.8 | - | 1.5 | 2.57 | 4.09 | - | 4.76 | 58.2 | 56.1 | - | 61.7 | 2.94 | 5.77 | - | 7.17 | | | |
| Guitar train | G_{00} | 17.4 | 0.2 | 3.0 | - | 4.4 | 1.68 | 3.9 | - | 4.3 | 50.7 | 49.2 | - | 55.1 | 1.74 | 5.05 | - | 6.07 | | | |
| | G_{10} | 1.6 | 0.0 | 0.7 | - | 0.9 | 2.56 | 4.19 | - | 5.16 | 54.8 | 56.6 | - | 51.0 | 2.84 | 5.43 | - | 7.3 | | | |
| | G_{01} | 0.8 | 0.0 | - | 2.1 | 1.8 | 2.51 | - | 5.04 | 4.59 | 62.5 | - | 49.3 | 60.3 | 2.77 | - | 7.31 | 6.91 | | | |
| | G_{11} | 0.0 | 0.0 | - | 3.1 | 0.0 | 3.05 | - | 4.31 | 5.28 | 57.6 | - | 52.4 | 59.6 | 3.36 | - | 6.18 | 7.69 | | | |
| Local Discriminator | P_{01} | 1.6 | 0.0 | - | 1.8 | 3.5 | 2.35 | - | 4.28 | 4.01 | 50.2 | - | 59.5 | 58.6 | 2.59 | - | 6.13 | 5.88 | | | |
| | G_{11} | 0.4 | 0.2 | - | 3.3 | 0.6 | 2.32 | - | 4.62 | 4.66 | 55.6 | - | 47.8 | 57.9 | 2.46 | - | 6.4 | 6.68 | | | |

5. Subjective Evaluation: Listening Test

- **40 participants**, recruited via social circles
- **Unconditional Generation:** Comparison to the original MuseGAN configuration, in **pairs**.
- **Conditional Generation:** Comparison between our developed configurations, as well as real samples, in **triplets** (conditional track + two accompaniments)
- **Criteria:** Music Naturalness, Harmonic Consistency, Musical Coherence



Unconditional Generation

| | Proposed | MuseGAN |
|-------------|----------|---------|
| Naturalness | 89 | 11 |
| Harmony | 85 | 15 |
| Coherence | 82 | 18 |

- The proposed framework outperforms MuseGAN with respect to all the examined musical aspects.
- Improvement in **Naturalness** & **Coherence** is attributed to our parameterized architecture that emphasizes on rhythmical attributes.
- **Stronger harmonic relations** among the tracks and **enhanced tonality** as a result of the shared/private design.

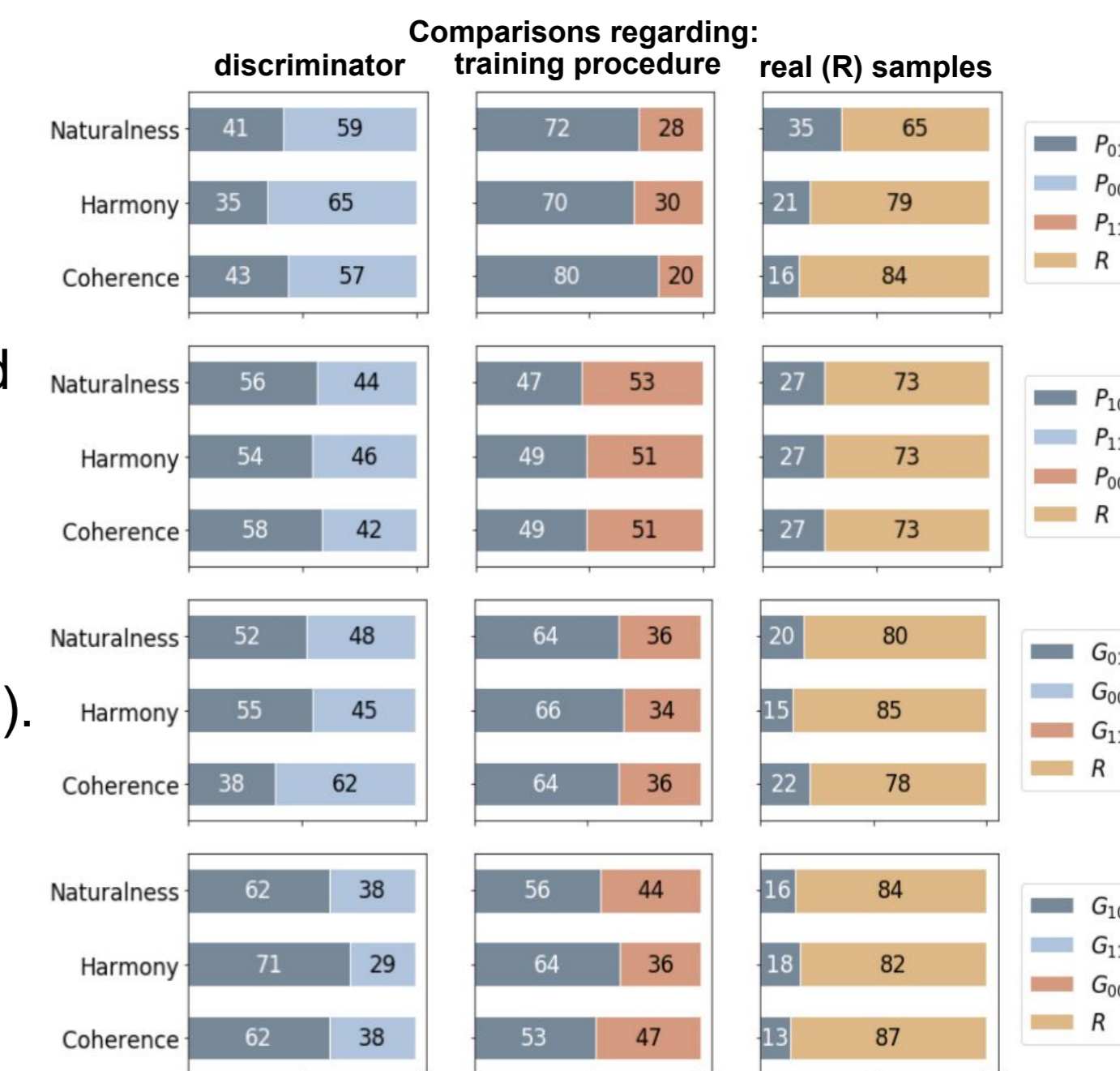
Conditional Generation

Piano models:

- Fake accompaniments are easily distinguishable
- P_{01} best compared to real on **Naturalness** (35%).
- P_{01} outperforms P_{11} with respect to all the examined musical aspects, especially **Coherence**.

Guitar models:

- Fake versions are **easily distinguishable** under all musical criteria (preference ranging from 13 to 20%).
- G_{10} outperforms G_{00} and G_{11} regarding all musical aspects (2-phase mode with Global Discriminator).
- G_{01} surpasses G_{11} , indicating that the most suitable training practice for the architecture of both Discriminators is the 1-phase mode.



6. Conclusions

- Proposed a **configurable generative framework** capable of:
 - creating multi-track polyphonic musical phrases **from scratch**,
 - generating multi-instrumental **accompaniments** for human-composed tracks.
- Hierarchical **shared/private design** for both Generator and Discriminator modules.
- **Objective and subjective evaluation:**
 - **Outperform** MuseGAN in the unconditional setup under 3 musical criteria.
 - Provide **useful insights** on training and structural schemes for conditional setups.
- **Future work:** validate our findings on transformer-based architectures and use other feature representations.

References

- [1] H.-W. Dong et al., "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment", in *Proc. AAAI 2018*.
- [2] T. Bertin-Mahieux et al., "The Million Song Dataset", in *Proc. ICWWW 2012*.
- [3] H.-W. Dong and Y.-H. Yang, "Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation", *arXiv preprint arXiv:1804.09399*, 2018.

Acknowledgements

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).