Ph.D. in Information Technology
Thesis Defenses

May 22th, 2026
At 11:30 a.m.
Room Alpha - Building 24

Gioele GRECO – XXXVIII Cycle

NEW TECHNOLOGIES FOR INTELLIGENT SPATIAL AUDIO
Supervisor: Prof. Fabio Antonacci

Abstract:

Emerging immersive technologies, including telepresence and virtual reality, require spatial audio systems that maintain coherence during user movement (6-DoF rendering), even when operating over sparse, heterogeneous sensor networks with limited bandwidth. This dissertation conceptualizes intelligent spatial audio as a cross-layer challenge, proposing a unified hierarchical architecture that integrates network-level cooperation, node-level robustness, and field-level reconstruction into a deployable pipeline. Distributed optimization facilitates online cooperation across the network layer, allowing nodes to collectively solve complex problems such as sound source localization. The proposed approach operates without a fusion center, relying solely on local direction-of-arrival estimates, while adaptive rules mitigate degradations arising from varying node reliability. Furthermore, to address multi-source scenarios, the thesis introduces a fully distributed computational strategy for online identity alignment. At the node layer, the work addresses robustness in signal acquisition using spherical microphone arrays. To stabilize spherical-harmonic representations in reverberant environments, a convolutional encoder-decoder is developed. This architecture maps noisy descriptors to anechoic-like counterparts, thereby reducing bias in local direction-of-arrival estimates and mitigating error propagation across the network. At the field layer, the challenge of sound-field reconstruction under data scarcity is addressed through two complementary frameworks. First, ParaDER offers a parametric approach that explicitly reconstructs direct sound, early reflections, and diffuse reverberation from limited measurements. Second, DAME provides a data-driven mixtureof-experts model designed to learn localized mappings where simplified physical models fail within complex geometries. Finally, this work contributes to the reproducibility and advancement of the field by releasing two open-access datasets dedicated to spatial audio and musical acoustics.

Francesca RONCHINI – XXXVIII Cycle

FROM ARTISTIC INTENT TO AUDIO GENERATION: CONTROL, EVALUATION, AND SUSTAINABILITY OF TEXT-GUIDED AUDIO GENERATIVE MODELS

Supervisor: Prof. Fabio Antonacci

Abstract:

Recent advances in generative modeling have transformed audio creation workflows. Tasks that once required specialized equipment and professional expertise can now be accomplished through simple text prompts to generative systems. Text-guided audio generation is reshaping creative industries by enabling sound designers, musicians, and non-experts to synthesize music, sound effects, and environmental audio directly from natural language descriptions. Despite their growing accessibility and capabilities, it remains unclear to what extent these models can reliably interpret and reproduce fine-grained artistic intent. Prior work indicates that achieving precise control over generated audio and capturing nuanced aesthetic attributes continues to pose significant challenges, underscoring the limitations of relying solely on textual input for detailed artistic specification. Moreover, the broader implications of integrating generative tools into cultural and creative practices remain unclear, particularly regarding their impact on artistic processes, cultural production, and human creativity. Existing evaluation practices rely on objective metrics and subjective listening tests, but these offer only a partial view of model behavior and fail to capture the complex interactions between generative systems, human interpretation, and cultural contexts. To address this gap, we developed two custom tools that extend Text-To-Audio models with optional controls, allowing us to investigate whether these features better support artists' expressive intent. The first interface enhances control at the input stage by combining text prompts with reference audio, enabling optional audio personalization and closer alignment with the artist's creative intent. The second interface integrates a source separation model to provide control over individual stems within generated tracks. We then conducted user studies to evaluate the tools' usability, creative impact, and potential integration into creative workflows. The results reveal that most participants were interested in adopting the tools, considering text-based generative models as both production aids and sources of unexpected inspiration that supported creative exploration beyond predefined goals. However, participants raised challenges regarding beat synchronization and the limited editability of the generated audio, particularly when integrating the tools into existing creative workflows. To address this limitation, we introduce an audio waveform generative model based on the Mamba Selective State-Space architecture that generates raw audio samples aligned with external temporal conditioning, achieving substantially faster generation than state-of-the-art methods while delivering higher overall quality, competitive temporal fidelity, and lower computational complexity. The community has raised concerns about copyright, potential misuse, artist authorship, and stylistic homogenization. To address these issues, we adopt Anti-Memorization

Guidance (AMG), which modifies the sampling process of pre-trained diffusion models to discourage memorization while preserving generation quality. Integrating AMG into the generative model reduces the risk of stylistic homogenization and the likelihood of reproducing copyrighted material. Ensuring realistic audio quality is also essential to support artists and creators in their workflows. To evaluate how well synthetic audio captures the characteristics of real recordings, we propose a novel evaluation approach. We generate audio datasets that mirror real data and use them to replace real recordings when training deep learning models for downstream tasks. This enables assessment of how effectively synthetic audio serves as a practical alternative in real-world scenarios. Our results show that audio produced by generative models often falls short, providing only partial replacement of real data. One source of this limitation lies in the prompt strategies used during generation. To address this, we evaluated multiple prompting approaches and identified techniques that substantially improve the quality and realism of generated audio. This work also highlights the potential of text-based audio generative models as efficient dataset generators, offering a promising path to reducing the cost and effort of real-world data collection. Beyond data quality and generation effectiveness, generative models also entail substantial computational and energy costs. To better quantify these demands, we analyze the energy consumption of multiple text-to-audio generative models across different architectures and applications, examining how inference-time settings affect energy usage. We further derive a Pareto frontier that highlights trade-offs between performance and efficiency. Together, these insights support more sustainable use of text-to-audio generative models and enable informed, application-specific model selection.

PhD Committee

Dr. Marco Marcon, **Politecnico di Milano**

Prof. Pedro Vera Candeas, **Universidad de Jaen**

Prof. Geoffroy Peeters, **Télécom Paris**