Міністерство освіти і науки України
Національний технічний університет «Харківський політехнічний інститут»
Мішкольцький університет (Угорщина)
Магдебурзький університет (Німеччина)
Петрошанський університет (Румунія)
Варшавська політехніка (Польща)
Познанська політехніка (Польща)
Софійський університет (Болгарія)
Міжнародний університет INTI (Малайзія)

Ministry of Education and Science of Ukraine
National Technical University «Kharkiv Polytechnic Institute»
University of Miskolc (Hungary)
Magdeburg University (Germany)
Petrosani University (Romania)
Politechnika Warszawska (Poland)
Poznan Polytechnic University (Poland)
Sofia University (Bulgaria)
International University INTI (Malaysia)

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ: НАУКА, ТЕХНІКА, ТЕХНОЛОГІЯ, ОСВІТА, ЗДОРОВ'Я

Наукове видання

Тези доповідей
XXXIV МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ
MicroCAD-2026

INFORMATION TECHNOLOGIES: SCIENCE, ENGINEERING, TECHNOLOGY, EDUCATION, HEALTH

Scientific publication

Abstracts
XXXIV INTERNATIONAL SCIENTIFIC-PRACTICAL CONFERENCE
MicroCAD-2026

Харків 2026

Kharkiv 2026

І 74

УДК 004(063)

Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXXIV міжнародної науково-практичної конференції MicroCAD-2026, 13-16 травня 2026 р. / за ред. проф. Сокола Є.І. – Харків: НТУ «ХПІ». – 2029 с.

Подано тези доповідей науково-практичної конференції MicroCAD-2026 за теоретичними та практичними результатами наукових досліджень і розробок, які виконані викладачами вищої школи, науковими співробітниками, аспірантами, студентами, фахівцями різних організацій і підприємств.

Для викладачів, наукових працівників, аспірантів, студентів, фахівців.

Тези доповідей відтворені з авторських оригіналів.

ISSN 2786-9253 (Online)

IMAGE TO AUDIO AGENT FOR CREATING AUDIO SEQUENCES

Volokitin V.G., Selivanova K.G.
Kharkiv National University of Radio Electronics, Kharkiv

Recent advancements in multimodal systems enable the direct transformation of images into soundscapes. To achieve this, we propose a three-stage Image-to-Audio pipeline (scene analysis, intelligent interpretation, and acoustic synthesis) with the primary objective of developing a highly effective method for converting visual data into descriptive text prompts for audio generation models [1].

The process of converting a visual signal into an audio sequence is divided into three sequential technological stages, each of which is based on a distinct type of neural network:

– analytical level (Vision): responsible for extracting semantic information from the image. At this stage, objects, their spatial relationships, lighting and the overall atmosphere of the scene are identified. The result is a structured set of features or a technical description;

– cognitive level (SLM): uses a small language model to "make sense" of the data received. SLM transforms technical descriptors into a creative, technically accurate description of the sound scene. This block is critical for establishing the correct "mood" and dynamics of the resulting audio;

– synthesis level (Audio): takes the generated text and converts it into an audio track, using diffusion or transformer architectures trained on a wide range of sounds and music.

To demonstrate the viability of the architecture and test its hypotheses, an optimal set of modern neural networks was selected: Florence-2 was chosen for the vision component, providing accurate segmentation and detailed scene description; Phi-3.5-mini was selected as the SLM for fast and efficient text processing and prompt generation; and Stable-Audio-Open-1.0 handles the audio synthesis, generating high-quality stereo sound.

The practical value of this development could be realized by addressing specific social and humanitarian needs. As a potential assistive technology for people with visual impairments, the system could create atmospheric audio "snapshots" of the surrounding world, allowing users to intuitively perceive their space through ambient sounds like rustling leaves or distant traffic. Additionally, in the fields of psychology and therapy, such an architecture could be used to generate personalized binaural soundscapes based on pleasant imagery to promote relaxation, aid meditation, and reduce anxiety levels.

The proposed modular architecture demonstrates the advantages of using SLM as an intelligent "bridge" between computer vision and sound generation. From the examples of models cited above, which served as the key components of the concept, we can highlight the potential for creating compact and efficient conveyors capable of operating with high precision.

References:

1. A. Sokolov, S. Nataliia, A. Sokolov and K. Selivanova, "Overview Of Modern Augmented Reality Capabilities For Creating A Navigation Aid For The Blind," *2023 IEEE 4th KhPI Week on Advanced Technology (KhPIWeek)*, Kharkiv, Ukraine, 2023, pp. 1-4, doi: 10.1109/KhPIWeek61412.2023.10311579.