Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Deliverable index:	D1.1
Version:	1.0

KICK-OFF MEETING REPORT

of the Project "Multimodal multilingual human-machine speech communication" (AI-SPEAK).

The kick-off meeting took place in Novi Sad, on the premises of the Speech Technology Group at the Faculty of Technical Sciences, University of Novi Sad, on **January 29th 2024**, with participation of all team members.

The project kick-off meeting started with a brief summary of the project objectives, particularly its specific objectives:

- Recording and processing of an audio-visual speech corpus in Serbian and English with accompanying video recordings of lip movements, recorded in strictly controlled conditions;
- Collection, sampling, and processing of a similar corpus, but obtained from existing videos published on the Internet, providing the flexibility needed in real-world scenarios (speech recognition “in the wild”);
- Design and implementation of beyond-state-of-the-art deep learning algorithms for:
  - speech recognition using visual cues from facial movements (audio-visual speech recognition), as well as speech recognition from video without audio (lip-reading);
  - direct conversion of video-only recordings into audio based on soft speech recognition (video-to-speech synthesis);
  - synthesis of facial movements from speech using visual representations of speakers (speech-to-lip), as well as from text (text-to-lip);
- Planned integration of the abovementioned algorithms into existing ASR and TTS systems for the Serbian language.

The discussion that ensued focused on project implementation particularly in view of the Gantt chart, structure and relations between workpackages, as well as project deliverables. Particular team members assumed their roles as leaders of workpackages and discussed the interplay between project tasks, their dependence on existing and new data as well as the equipment that is to be procured in the first phase of the project. New scientific methods and technical frameworks, related to the approach to particular project objectives were discussed in detail, as an introduction into the series of technical meetings

aimed at the definition of a detailed **project implementation plan** (deliverable 1.2, due in April 2024). The research direction related to speech recognition using visual cues from facial movements (audio-visual speech recognition) was discussed in particular detail, as the project external collaborator announced the impending availability of an audio-visual speech recognition corpus and the corresponding research framework, which would be very useful until aprropriate corpora are collected and annotated within this project.

Furthiermore, dissemination activities at the project were discussed, as regards interaction with other similar projects and plans for joint publications, in line with the principles of open data and open research, which will be expanded in more detail in the **project dissemination plan** (deliverable 1.2, due in April 2024).

The **project website** has been created in January 2024: https://www.ktios.ftn.uns.ac.rs/ai-speak/AI-SPEAK_sr.html, and its public section will be used for Project presentation and dissemination, as well as repository of relevant scientific papers that the members of the project team have published before or during the Project, as well as the implementations of the algorithms (computer code) developed within the project. The private section of the Project website will be used as a platform for communication between team members, including a repository of relevant scientific papers published by others, as well as a repository of computer code shared between team members.