Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Milestone index:	M2.1
Version:	1.0

PROJECT MEETING REPORT

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

The 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 May 16th 2024, with participation of all team members. The focus of the project meeting was the definition of the methodology for collection of multimodal data, most notably including two corpora referred to as **AI-SPEAK speech corpus** and **Internet speech corpus**.

I. Methodology for collection of AI-SPEAK speech corpus

AI-SPEAK speech corpus is planned to contain recordings of speech in both Serbian and English from 25 adult speakers of both genders, together with video recordings of the movements of their lips. The intended quantity of speech data per speaker is 10 minutes. The corpus will be recorded in the IAC Mini anechoic chamber of the University of Novi Sad. The speakers will deliver a fixed number of utterances in both Serbian and English, including spoken digits, names of letters, simple commands (“up”, “down”, “back”...) as well as a number of short sentences, including a set of sentences which will be the same for all users, and another one which will be different for each speaker. The corpus will be phonetically aligned and semi-automatically annotated for prosodic events (accents, phrase breaks, sentence emphasis).

The pipeline for the efficient corpus production is presented in Figure 1. The basis for the production of the corpus is the repository of sentences in Serbian and English, which serves as a point for fast production of personalized scripts for each individual speaker. The fixed part of the repository, identical for all speakers, in both languages will consist of the spelled letters of the alphabet, names of 10 digits, 7 days as well as 13 command words (forward, back, left, right, up, down, confirm, cancel, delete, send, next, home, end). In this part there will be a 1-to-1 correspondence between items in Serbian and English, except in the case of alphabet. The part of the repository intended for corpus personalization will consist of a large number of random sentences in both Serbian and English, short enough to be easy to render (most often) without mistakes, but long enough as to ensure that each speaker delivers approximately 10 minutes of speech in total in both languages. We plan 25 sentences in either language

for each speaker, without a modest level of control of the phonetic coverage. These will contain approximately 350 words for Serbian and 400 words for English, having in mind that average word length in English is lower and consequently, average word rate in English is higher, which is why it is expected that 350 words in Serbian will correspond to approximately the same net speech duration as 400 words for English.

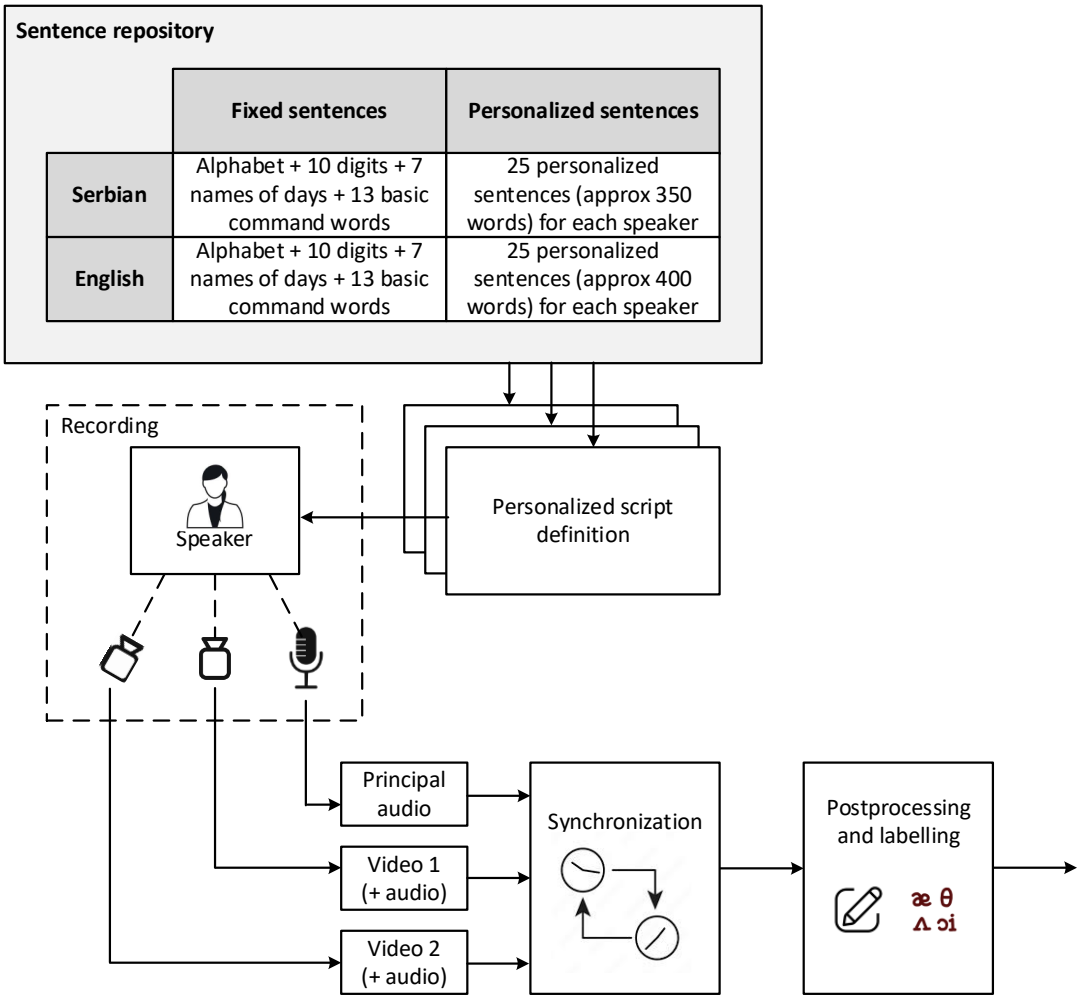

Figure 1.

As shown in Fig. 1, every speaker will deliver (i.e. pronounce) their own personalized script and will be recorded by a high quality microphone Rode Podmic, which has been obtained for this purpose, as well as Sony VLOG camera ZV-1, also obtained at this project, which can capture multimodal data (audio+video). Furthermore, to obtain auxiliary low quality audio and video recording we will use a standard quality cellphone as a source of additional information which can be considered very useful from the aspect of training of AI systems.

Audio/video streams obtained in this way from different sources need to be synchronized, and to do this we will augment the PPT presentation which will be used to prompt the speaker with a pilot tone that will be played out at each slide (at each sentence). This will (1) enable the speaker to navigate through prompts at his/her own pace, and (2) enable us to synchronize the audio component coming from different sources. Clearly, the video component will be automatically synchronized as a consequence. All video recordings will be postprocessed (cropping, level adjustment etc.). Only the audio component from the high-quality microphone will be submitted to phonetic and prosodic annotation, and the obtained transcriptions and other annotation data will be automatically propagated to complementary recordings. An efficent recording setup will enable the recording process to take no more than one hour per recording session (per speaker), possibly only half as much, since the speakers will be able to control the recording session and the estimated time for production of 10 minutes of net speech is reasonably estimated (from previous experience) at three times net speech duration.

II. Methodology for collection of Internet speech corpus

Internet speech corpus, which will include a far greater quantity of data (principally in the Serbian language, as there is sufficient coverage of English), but with little or no control over textual content or recording conditions, with far greater diversity in terms of factors such as speaker characteristics, sound quality, acoustic ambience, recording equipment, lighting or background.

- We will **collect video recordings** from sites like YouTube, using specific search queries, look for YouTube channels dedicated to Serbian content, like news channels (RTS, N1), entertainment, or education channels. For instance, the official streaming service of Radio Television of Serbia (RTS) offers a variety of Serbian content. It should also be noted that Google's advanced search features can be used to find videos specifically in Serbian (we will search for video file types (like .mp4) using filetype:mp4). We will check for playlists that might compile multiple videos, and filter them according to audio/video quality and the degree of spontaneity of speech. Video search engines like Bing Video or DuckDuckGo Video to find Serbian videos will also be used.

- We will explore **existing video resources**. We will explore Serbian news websites which often have video sections, as well as many news outlets who also post video content on their social media channels.We will check for any Serbian national or regional video archives that might have public access to recordings.

- We will explore **alternative approaches** including use of educational resources and university archives.

We will always ensure that downloading and using videos comply with copyright and legal requirements.

Colledted videos will be post-processed similarly to videos produced through the pipeline related to the AI-SPEAK corpus. The focus will be on extracting the region around the lips for each speaker as well as level adjustment. The videos will be phonetically transcribed in order to produce a valuable speech resource usable for training multimodal large language models, which will be capable of multimodal speech recognition as well as multimodal speech synthesis. Automatic phonetic transcription will be

obtained through the system for (audio-only) **automatic speech transcription of very high accuracy** in Serbian, trained on large quantities of existing data at the disposal of the project team, as explained in the project proposal (most notably **large multi-speaker corpora** of elicited or spontaneous speech (a total of nearly 400 hours), which are phonetically annotated and to a great extent manually inspected for errors, as well as **large text corpora** containing different functional styles (a total of over 26 million words), also manually inspected for errors, and used for language modelling).