Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Milestone index:	M2.2
Version:	1.2

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 August 30th 2024, with participation of all team members. The focus of the project meeting was the definition of the methodology for processing of multimodal data, i.e. two corpora referred to as **AI-SPEAK speech corpus** and **Internet speech corpus**.

I. Methodology for processing of AI-SPEAK speech corpus

AI-SPEAK speech corpus will 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 average quantity of speech data per speaker is 10 minutes (silent segments excluded). The corpus was recorded in the IAC Mini anechoic chamber of the University of Novi Sad. Each speaker has delivered the following items:

- alphabet spelling in Serbian
- 4 fixed sets of words in Serbian
    - names of digits (*nula jedan dva tri četiri pet šest sedam osam devet*)
    - names of days (*ponedeljak utorak sreda četvrtak petak subota nedelja*)
    - spatial directions (*napred nazad levo desno gore dole*)
    - command words (*potvrdi odustani obriši pošalji dalje početak kraj*)
- a phonetically balanced set of fixed 25 sentences in Serbian, identical across all speakers (e.g. "*Kod glodara je nađena povećana koncentracija olova*.")
- a phonetically balanced set of 50 sentences in Serbian, different for every speaker

- alphabet spelling in English
- 4 fixed sets of words in English
    - names of digits (*zero one two three four five six seven eight nine*)
    - names of days (*Monday Tuesday Wednesday Thursday Friday Saturday Sunday*)
    - spatial directions (*forward back left right up down*)
    - command words (*confirm cancel delete send next home end*)

- a phonetically balanced set of fixed 25 sentences in English, identical across all speakers (e.g. "*When two straight lines come together, they make an angle*.")
- a phonetically balanced set of 50 sentences in English, different for every speaker

The corpus will be phonetically aligned and semi-automatically annotated for prosodic events (accents, phrase breaks, sentence emphasis), using the following methods:

The pipeline for the efficient corpus production is presented in Figure 1, given also in project report M2.1, and the following is the description of the processing procedure in more detail.

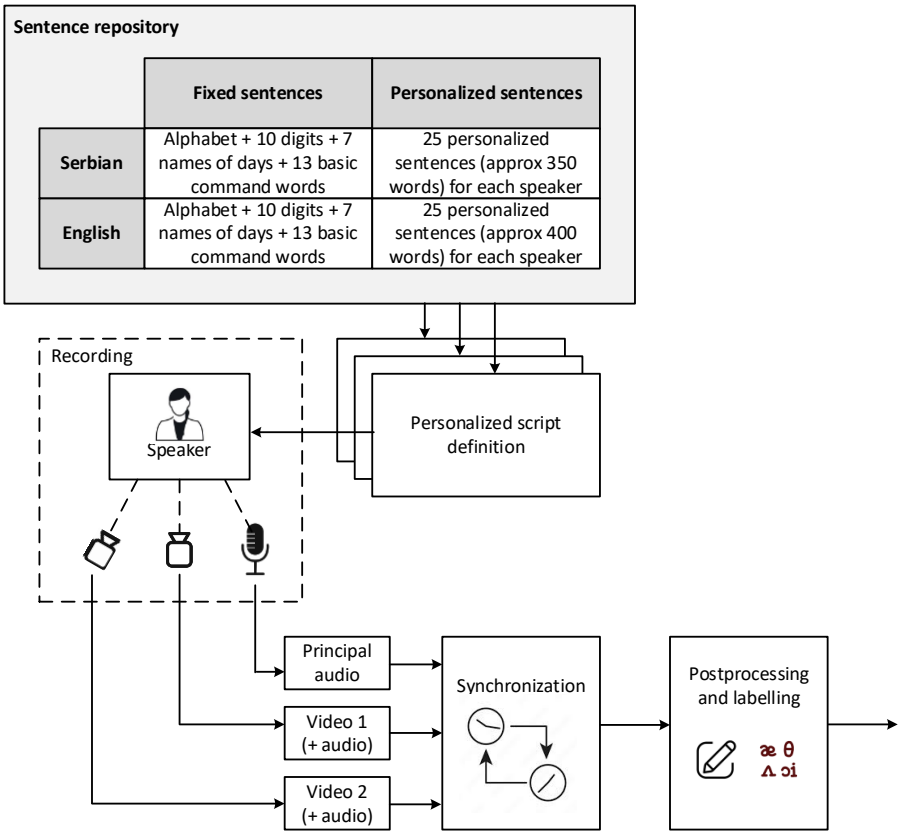

Figure 1.

As shown in Fig. 1, every speaker has delivered (i.e. pronounced) their own personalized script using a high quality microphone Rode Podmic, as well as Sony VLOG camera ZV-1, able to capture multimodal data (audio+video). Furthermore, to obtain auxiliary low quality audio and video recordings we have used standard quality smartphones (Samsung Galaxy A33 5G and Samsung Galaxy S10+).

The plan is to use an audio recording obtained from a high-quality Rode Podmic microphone. However, it has been noticed that there is interference in the form of constant noise originating from a nearby transformer. One of the preprocessing steps for the database will be removing this noise. This will be done by Audacity Noise Reduction Effect. In the first phase, a chunk of an audio signal containing only noise is selected, in order to calculate the power spectrum of noise. The second phase is noise removal using

information (statistics) about the noise power spectrum. During the noise reduction phase, the gain for each frequency band is set such that if the power of sound exceeds the previously-determined threshold, the gain is set to 0 dB, otherwise the gain is set lower, in order to suppress the noise. After that time-smoothing is applied to obtain slow changes of the gain for each frequency bin. It is followed by frequency-smoothing in order to achieve that a single frequency is never suppressed or boosted in isolation.

Except for the audio recorded with the dedicated microphone, the database will contain three video recordings acquired by three different video cameras. The audio from these recordings will be completely removed after synchronization, but audio content from the video recordings will be used in process of synchronization of these recordings with the audio recording made by main recording microphone (Rode Podmic). For this reason, 500 milisecond sinusoidal signal segments (referred to as the *separation signal* in the rest of the text) were used after each slide of the PPT presentation displaying the text the participant is supposed to read. Participants were instructed to repeat entire sentence after a short pause, in case mistake was made during recording process.Therefore, the segment between two consecutive separation signals, contains either one correct sentence or several incorrect attempts followed by the correct sentence. In any case, the correct sentence should always be at the end of the segment, followed by the separation signal.

The first step in recorded database post-processing procedure is the annotation of the database. The initial manual annotation will be performed only on the audio recordings from the high-quality microphone. The annotation procedure will be performed using Audacity software. This procedure requires setting marks at the beginning and end of each segment in the database, i.e., the positions of the separation signals, and add a label with the text of what was said in the marked segment. An example is shown in Fig. 2.

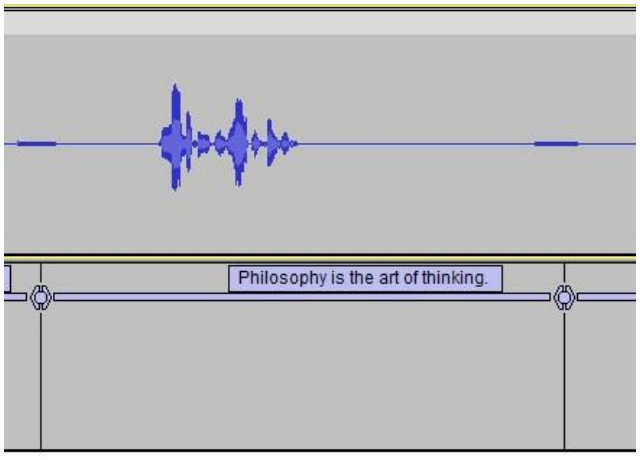

Fig.2 Appearance of one segment with manually annotated start and end

In process of marking the segments there could be several possible cases, which are also illustrated in Fig. 3.

1. If the entire segment between two separation signals is useless (nothing was said, or all attempts were wrong - for example, if something was just tested or it is the very end), the segment will be labelled by label **###D**.
2. If the segment between two separation signals contains some wrong attempts, but at the end there is a correctly spoken sentence, the segment will have a label containing the text of the correct sentence, along with a so-called marker, labeled as **###M** with a duration of 0s (the start and end are in the same position).
3. If the segment between two separation signals contains only a correctly spoken sentence, the segment will contain only a label containing the text of the spoken sentence.

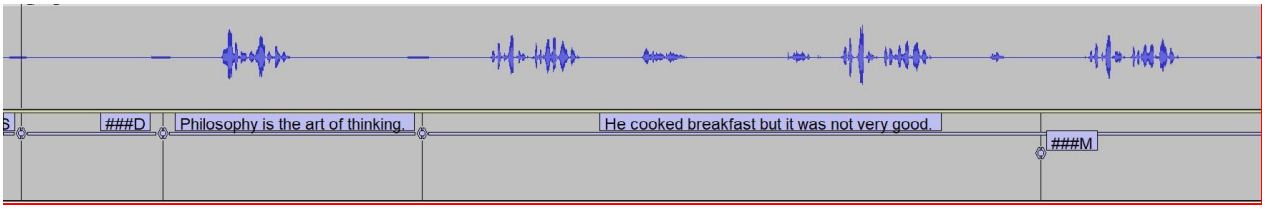

Fig.3 Appearance of all cases of text in labels: ###D, ###M and correct text of pronounced sentence

Additionally, the annotators will be instructed to note all observed issues that are required to be handled manually. An initial review of one portion of the database revealed several characteristic problems that will need to be resolved appropriately. The identified problem could be separated into several classes:

1. The participant misread a word but then repeated only that word, not the entire sentence.
2. The audio recording was damaged while the participant was pronouncing the correct sentence.
3. The participant uttered some noticeable sigh or onomatopoeia (e.g., "hmmm," "aaa") just before pronouncing the correct sentence.
4. The pronunciation of the sentence overlaps with the separation signal.
5. The participant pronounced an incorrect word and did not correct themselves (e.g., said "feel" instead of "fell" or "ujedno" instead of "uljudno").
6. The participant added a word that was not in the transcript.

The plan for problems of type 1) and 2) is to discard the problematic segment for that particular participant. If these are sentences from a set of sentences common to all participants, it will be treated as a missing value. If the set is unique to each speaker, it will not be marked as having existed in the database. Problems of type 3) and 4) will not be addressed, meaning they will present a challenge in recognition - problem 4) only complicates the audio recording. Problems of type 5) and 6) will be handled by correcting the transcript itself, which may pose an issue if they occur in the common sentence set. Depending on the number of such errors, it will be considered whether to remove the sentence from the database for all participants or to exclude the specific participant if a high percentage of their sentences contain errors.

An initial review of the video recordings identified the following problem: the participant frequently changed positions or moved their head. Although there was very little space in the anechoic room itself, some participants felt uncomfortable and thus often moved significantly during the recording, even while

pronouncing a single sentence. This will be resolved by excluding the participant. The initial review suggests that this pronounced issue appears in only 2 out of 35 recorded participants. Such a problem could significantly complicate both the preprocessing of the database and the training of ML models in later stages of research, so we believe it should not be present in a database recorded under strictly controlled conditions. Additionally, during recording sessions for 3 participants one of the auxiliary phone cameras turned off during the session, so the recording from that specific camera is missing. This will be noted in the database that the recording is unavailable, but the participant will not be excluded as there is a main video and a main audio recording.

After the database annotation, the labells will contain following information: start timestamp of the segment (*segStart*), end timestamp of the segment (*segEnd*), and thecorresponding textual transcription. The precise timestamps for extracting only useful audio content (and not the parts of the separation signal) will be automatically set by the dedicated algorithm. Labellers will be instructed to mark the approximate position of separation signal. The algorithm will extract 1 second segment around this mark (500 ms before and 500ms after) ensuring that the whole separation signal is included in these segments. The maximum value of the correlation between extracted segment and original 500-milisecond sinusoidal segment will enable the precise positioning of the separation signal in the extracted segment. *segStartCorrected* will be set 20 miliseconds after separation signal end, while the *segEndCorrected* will be set 20 milisecons before next separation signal start. The positions of the labels after manual annotation and after automatic procedure is shown in Fig. 4. Additionally in the segments with some erroneous attempts *segStartCorrected* will be set to position of **###M** label

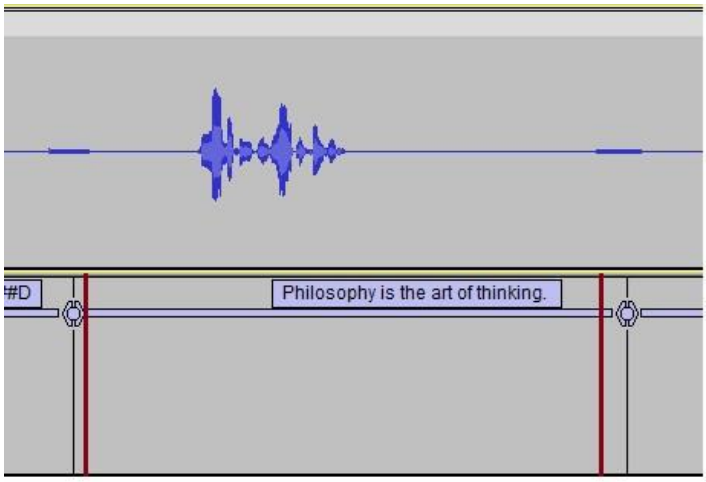

Fig.4 Illustration of corrected borders of one segment to be aligned with separation signal

The audio recording will then be processed as follows. At the determined timestamps for the start and end of the segment (*segStartCorrected* and *segEndCorrected*), the recording will be cut. The process of extracting only parts of audio and discarding the rest is shown in Fig. 5. In accordance with the determined timestamps for cutting (and possibly discarding) segments from audio recording, the video recordings will also be processed. It is implied that if a segment is deleted from the audio recording, it must also be deleted from all video recordings. Since not all recordings on all three cameras and the microphone

started simultaneously, to avoid synchronization issues, the position of the last separation signal in each of the video recordings will be manually determined before cutting the audio recordings. The timestamps for cutting (and discarding) will be adequately delayed/adjusted accordingly.

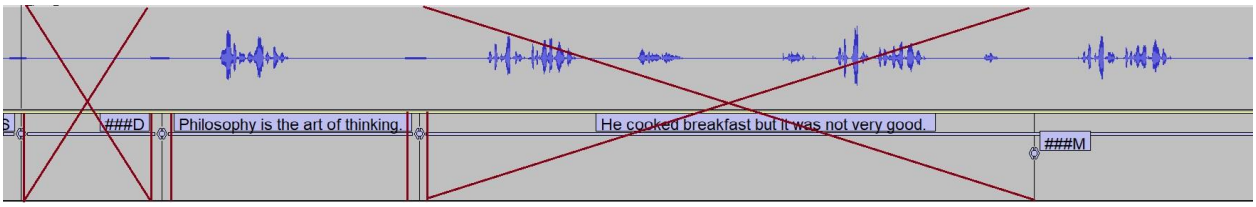

Fig.5 Illustration of what should be erased

Finally, for each participant, we will obtain N folders, where N corresponds to the number of sentences for that specific participant. Each folder will contain an audio recording and three video recordings (without audio) for the same sentence, along with a label with the textual content of the sentence. There will also be a metafile listing all folders and all information, including:

1. File path and filename or an indication of a missing value if the sentence is from the common sentence set and is missing for a specific speaker from some camera or was entirely removed.
2. The textual content of the sentence.
3. Any accompanying notes if they exist.

II. Methodology for processing of AI-SPEAK speech corpus

The creation of a Serbian **Internet speech corpus** represents a crucial step in advancing multimodal language and acoustic models capable of speech recognition, synthesis, and other innovative linguistic applications. This resource will address the lack of Serbian-language datasets by focusing on news broadcasts, which provide a rich source of natural, formal, and spontaneous speech. The corpus will integrate audio, text, and video while prioritizing diversity in speaker characteristics, sound quality, and recording environments. By adhering to rigorous legal and ethical standards, it will become a foundational tool for researchers and developers working on advanced multimodal systems.

Multimodal language models have achieved significant advancements in tasks such as automatic transcription, sentiment analysis, real-time translation, and speech synthesis. These systems rely on datasets that combine text, audio, and visual data to capture linguistic and contextual nuances. While English dominates due to abundant resources, Serbian remains underserved, hindering the development of AI systems tailored to its unique phonetic and grammatical structures. The Serbian Internet speech corpus will address these challenges by focusing on news content, a domain rich in linguistic variety, ensuring its relevance for both academic research and practical applications.

Collecting Data from Online Sources

YouTube is one of the most significant resources for collecting Serbian-language video data due to its extensive repository of accessible content. Several Serbian news outlets maintain active channels on the

platform, including N1, Radio-Television Vojvodina (RTV), and Nova S. These channels regularly upload news bulletins and reports, providing a continuous stream of content suitable for developing a robust dataset. While certain channels, like Nova S, do not organize their content into structured playlists, their individual uploads remain readily available.

Efficient management of the downloading process will be achieved using Python libraries such as *pytube* and *yt_dlp*. The latter provides advanced functionality, including the ability to select specific video and audio quality settings. For example, downloading a single *Dnevnik* news bulletin from N1 in the highest quality results in a file size exceeding 300 MB, necessitating substantial storage capacity. To accommodate these requirements, the faculty servers will be used to host the collected data, with careful planning to ensure secure and scalable storage solutions.

Serbian news websites and social media platforms (e.g., RTS, B92, Pink, Prva Srpska TV, Tanjug, K1, Blic, Kurir, Informer, Insajder, Euronews Srbija, NewsMax Balkans) offer additional avenues for data collection, hosting unique video content such as breaking news, specialized reports, and in-depth interviews, not all of them shared on YouTube. To access this material, web scraping tools like BeautifulSoup and Selenium will be employed. These tools are capable of navigating dynamically generated web pages, identifying embedded video elements, and downloading them programmatically, ensuring comprehensive coverage of diverse sources.

Publicly accessible archives, including those maintained by universities or national libraries, serve as additional sources for historical and educational Serbian-language video content. Collaborations with these institutions may further enrich the dataset.

Automated scraping methods will be combined with manual verification to ensure the accuracy and relevance of the collected data. This dual approach helps maintain alignment with the corpus's goals of representing natural and formal speech in varied contexts. Throughout the process, strict adherence to copyright laws and privacy regulations remains a priority, ensuring the project complies with ethical standards.

One of the central goals of the Serbian Internet speech corpus is to embrace the diversity inherent in real-world speech. This involves collecting data under varying conditions, including differences in speaker characteristics, sound quality, recording environments, and background noise. Such diversity enriches the dataset, making it a valuable resource for training robust multimodal systems capable of handling a wide array of real-world scenarios. However, achieving this diversity presents several challenges. Variability in recording conditions can lead to inconsistencies in data quality, requiring post-processing to standardize the corpus. Additionally, managing large volumes of data from multiple sources demands careful coordination to ensure storage and accessibility. By leveraging these methods and tools, the Serbian Internet speech corpus will encompass a wide range of linguistic and acoustic conditions, offering a valuable resource for researchers and developers in the fields of speech recognition, synthesis, and multimodal system development. The corpus will adhere to the highest standards of legal and ethical compliance, ensuring its usability for academic and technological innovation.

Processing and Annotating the Corpus

The development of the Serbian internet speech corpus involves meticulous processing and annotation to ensure its suitability for training advanced multimodal language and acoustic models. Key stages in this workflow include segmenting video clips, extracting essential features, annotating and tagging the dataset (see Fig. 6). These steps ensure the final corpus meets the quality and diversity requirements needed for cutting-edge applications.

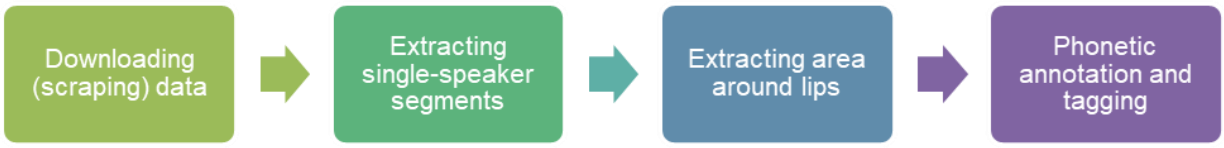

Fig.6 Internet speech corpus collection and processing workflow

A crucial task in building the corpus is *isolating segments with a single speaker*, as these are essential for applications such as phonetic transcription, speaker diarization, and multimodal speech synthesis. In order to refine the segmentation process, ensuring that video segments with single speakers are accurately identified and aligned for further analysis, we plan to explore the use of advanced processing libraries such as *MediaPipe Face Detector* and *MediaPipe Face Landmarker* to enhance the quality of our dataset.

The *MediaPipe Face Detector* identifies faces in video frames and tracks them across sequences. By pinpointing key facial features - such as eyes, nose, and mouth - the tool ensures that each segment contains clear, unobstructed views of the speaker's face. Its ability to operate in real-time and adapt to varying lighting and acoustic conditions makes it a robust solution for segmenting speech-focused clips.

Building on the capabilities of the Face Detector, the *Face Landmarker* adds a layer of granularity by identifying and tracking 3D facial landmarks. These include subtle facial movements and expressions, which are invaluable for tasks such as lip-reading and synchronizing speech with facial activity. Additionally, it outputs transformation matrices that align facial features across frames, enabling precise synchronization between audio and visual data. This capability is critical for creating training datasets for multimodal models, especially those requiring accurate lip-synchronized speech synthesis.

To enhance the quality of the extracted segments, we may apply supplementary algorithms to detect and filter out background noise or overlapping speech, ensuring that only clear and relevant single-speaker clips are retained. This preprocessing step ensures the corpus is optimized for tasks that demand high-quality, unambiguous speech data.

Each single-speaker segment will be further divided into subsegments containing single words or sentences. The *transcription* process will employ a fine-tuned version of *OpenAI's Whisper* model, specifically tailored for Serbian. This model will be trained on extensive Serbian-language datasets, enabling it to generate highly accurate phonetic transcriptions. This process will create a detailed linguistic

resource, allowing for precise analysis and annotation of the speech data. By segmenting the videos in this way, we will ensure that the transcriptions are highly accurate and aligned with the specific linguistic structure of the Serbian language. Video frames will be cropped to focus on the *speaker's lip region*, enhancing the multimodal capabilities of downstream models.

To maintain the highest standards, the automatically generated transcriptions will be manually reviewed for accuracy. Given the large scale of the dataset, this review process will focus on key segments or a representative sample to ensure the transcriptions meet the required quality standards. In addition to transcription, the dataset includes comprehensive *speaker tagging*. This process involves identifying and labeling individual speakers across the dataset, an essential feature for tasks like speaker diarization and identification. By tracking unique speaker attributes, the corpus becomes a valuable resource for training AI systems that need to handle multi-speaker scenarios effectively.

A distinguishing feature of this corpus is its focus on multimodal data integration. By aligning audio and visual streams with transcription and speaker tagging, the dataset is prepared for tasks requiring synchronized inputs. This integration is particularly beneficial for applications such as:

- Lip-Reading Models - Training systems capable of interpreting speech from facial movements, enhancing accessibility and usability in noisy environments.
- Multimodal Speech Synthesis - Developing systems that generate natural speech synchronized with facial expressions and lip movements.
- Context-Aware Speech Recognition - Enabling systems to improve recognition accuracy by considering visual cues alongside audio inputs.

Processing and annotating a diverse dataset pose unique challenges, including variations in sound quality, acoustic environments, and speaker diversity. These challenges will be met through a combination of advanced tools, robust algorithms, and meticulous human oversight. The resulting dataset will ensure coverage of a broad range of real-world scenarios, making it a versatile resource for academic research and practical applications.

The annotated and segmented Serbian Internet speech corpus will be poised to support groundbreaking advancements in multimodal AI. Its integration of audio, visual, and linguistic data will make it a foundational resource for researchers and developers, addressing challenges in areas such as speech recognition, synthesis, and human-computer interaction. By ensuring quality and diversity, this corpus will set a benchmark for linguistic resources in underrepresented languages, facilitating the development of more accurate and context-aware AI systems tailored to Serbian and other South Slavic languages.

Ethical Considerations

Compliance with copyright and data protection laws will be a cornerstone of the Project. Each video will be carefully assessed for legal permissions, and, where necessary, explicit consent will be sought from rights holders. Ethical considerations, including the privacy of individuals appearing in the videos, will be rigorously observed to ensure the corpus adheres to the highest standards.

The creation of the Serbian Internet speech corpus will involve innovative methods and a strong commitment to ethical data practices. By focusing on news content and leveraging state-of-the-art tools for data collection and processing, the project aims to build a valuable resource that will advance the capabilities of multimodal language models in Serbian. This effort addresses a critical gap in linguistic resources and lays the foundation for new possibilities in AI-driven communication technologies.