Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Milestone index:	M3.1
Version:	1.1

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 January 10th 2025, with participation of all team members. The focus of the project meeting was (1) defining implementation plans after the evaluation of the data which was recorded or collected from the Internet for the purposes of developing the **AI-SPEAK speech corpus** and the Internet speech corpus, suitably renamed to **AI-SPEAK Internet Video Database** (2) the evaluation of the outcome of initial stages of data processing and problems encountered in this work phase for both corpora.

I. Evaluation of AI-SPEAK speech corpus, defining implementation plans and correction measures

AI-SPEAK was initially planned to contain speech in both Serbian and English from 25 adult speakers of both genders, together with video recordings of the movements of their lips, with the average quantity of speech data per speaker being around 10 minutes, including, for each speaker and for both languages:

- alphabet spelling
- 4 fixed sets of words (names of digits, names of days, spatial directions and command words)
- a phonetically balanced set of fixed 25 sentences, identical across all speakers
- a phonetically balanced set of 50 sentences, different for every speaker

Each speaker was recorded 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+) positioned approximately 30 degrees to the left and right. All participants provided written informed consent for the recording and public release of this dataset, with the option to withdraw their recordings from the public version at any time.

I.1 Text/audio data evaluation

During the recording stage, we obtained recordings from **30 speakers, including 15 female and 15 male participants**, having in mind that some of them may have to be removed due to errors. Each speaker contributed 160 recordings: 80 in Serbian and 80 in English. Out of the 80 sentences per language, 30 are shared across all speakers, while the remaining 50 were speaker-specific (personalized).

Alignments between text, audio and video were generated automatically and they have been found to contain occasional inaccuracies; this component was not manually verified. Some recordings were found to contain brief content before or after the spoken sentence (e.g., facial expressions, laughter, sync signals, gestures such as covering the mouth) due to non-trimmed segments. However, the core utterance corresponding to the transcript is verified to be uninterrupted.

In case of errors in pronunciation, if the resulting utterance can still pass as a reasonably valid Serbian or English utterance, the transcript was updated accordingly. Consequently, some utterances may seem semantically odd due to transcript adjustment based on actual speech. Only recordings with clearly invalid words or major issues were removed. In some cases, recordings initially intended for the shared subset were reassigned to the personalized subset if the speaker mispronounced a word, even slightly (e.g., missing articles, incorrect number or pronouns), having in mind that while transcript modifications are not an issue for personalized sentences, they are for shared ones. Specifically, in Serbian, for the 30 "common" sentences, we have recordings from all 30 speakers for 20 of them, while the transcripts of the remaining 10 sentences were modified for 1–3 speakers (14 in total are modified out of 900). In English, out of 30 "common" sentences, we have 17 of them from all 30 speakers, while the remaining ones were modified for 1–5 speakers (21 in total are modified out of 900).

Audio recordings were found to contain a hum at 50 Hz and higher harmonics, and a task within Subactivity 2.2 was launched to define and implement a strategy for its removal, combining existing AI-based solutions (good performance but poor explainability) and/or a custom made digital signal processing solution. The general procedure for database annotation and audio file editing does not differ from the one described in detail in the report related to M2.2.

I.2 Video data evaluation

As specified in the Introduction, video recordings of each speaker were obtained using three video cameras—one high-quality front-facing (frontal) camera and two auxiliary mobile phone cameras positioned approximately 30 degrees to the left and right. Some recordings were found to be faulty (recording interrupted due to battery failure or human error), and corresponding recordings were omitted from consideration. Specifically, out of 4,800 recordings per camera, the following were deleted in total:

- 84 recordings across all cameras and audio (less than 2%);
- Between 0 and 14 audio and front-facing video recordings per speaker were deleted, totaling 84;

- The left camera is entirely missing for 5 speakers; for the others, 0–17 recordings per speaker were deleted, totaling 86 in addition to the 5 fully missing ones;
- For the right camera, more than one-third of the recordings are missing for one speaker (62 recordings), while for the rest, 0–14 were deleted per speaker, totaling 86 in addition to the missing ones for that speaker.

In the following text, we will outline the procedure for editing video files that we have agreed upon. This workflow will require an additional manual annotation of the last beep occurring in each video file. That timestamp will be used to extract the audio segment containing the beep from the video recording. A second beep will be extracted from the main-microphone audio track using the existing annotations (again, we will take the most recent beep in that signal).

By measuring the difference between the two beep positions—and refining it with the maximum of the cross-correlation between the two signals it will be possible to obtain the offset between the audio and video tracks. This offset will be subtracted from the previously established audio labels to produce the corresponding video-cut labels. This procedure will also be applied to find offset for two other video recordings.

For lip region extraction across all camera views, we propose to utilize the MediaPipe Face Mesh framework, which provides automated detection of 468 predefined 3D facial landmarks per frame. After detecting these landmarks, a masking procedure will be applied to isolate and retain only the region containing the lips and their immediate surroundings, based on a subset of landmarks relevant to mouth articulation. Our initial testing of this software enviroment has confirmed that it will be adequate for this task.

To ensure the region of interest (ROI) encompasses sufficient contextual information beyond just the inner lip contour, we will expand the ROI boundaries by including the following reference points: landmark 200 (lower nose), landmark 2 (chin midpoint), and landmarks 214 and 434 (left and right mouth periphery). These points (shown in Fig. 1) define a bounding box large enough to consistently capture the lips along with surrounding facial features essential for visual speech modeling, such as cheek and chin movements.

The resulting ROI will be computed dynamically for each frame based on the real-time positions of the selected landmarks. It should be noted that the ROI dimensions vary across speakers and over time, reflecting natural variability in facial structure and expression. Nonetheless, the content of each ROI per frame is consistent: it includes the lips and their predefined neighboring region, which are critical for accurate modeling of lip movements during speech.

All masked video recordings will be manually inspected to ensure that no irrelevant regions have been retained and that no relevant portions of the lip region have been inadvertently excluded. In addition, a manual verification will be performed to confirm that the video and audio recordings correspond accurately to the provided transcripts, and to identify any potential pronunciation issues or deviations from the intended utterances.

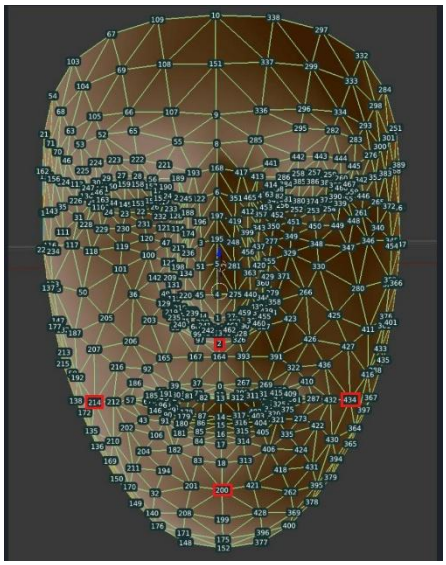

Figure 1. MediaPipe Face Mesh landmarks.

Finally, we propose to perform automatic alignment of the transcripts with the audio recordings using a Whisper-based ASR system for both Serbian and English. This process yields word-level time stamps, providing start and end times for each word in the recordings, expressed in milliseconds. Metadata for each speaker will indicate the filename, the availability of recordings from the frontal, right, and left cameras, and the microphone, respectively, the transcript of the spoken utterance corresponding to the recordings, the language of the utterance (ser/eng), and a "common" marker specifying whether the sentence is part of the common subset shared across all speakers (true) or part of the speaker's personalized subset (false). All recordings with the same filename (from the audio, video_a, video_l, and video_r folders) will be time-synchronized, i.e., they will share the same transcript and alignment information. Alignments will be generated automatically. All audio files in the database will be mono, 22.05 kHz, in WAV format (PCM_S16LE). All video recordings are in MP4 (MPEG-4) format. The frontal camera videos are recorded at 100 fps, while the auxiliary cameras record at 30 fps.

II. Evaluation of the progress in the development of VideoBase: AI-SPEAK Internet Video Database

VideoBase internet video database is envisioned as the large scale, structured database of internet videos in Serbian language that have adequate characteristics for subsequent extraction and processing of individual speakers appearing in the scene, aimed to facilitate different research activities that are part of **AI SPEAK project**, and in particular experimental studies aimed at the tasks of lip reading and speech resynthesis from video for Serbian language. Our plan is to cover wide range of person appearances and speaker characteristics, in terms of both number of unique high quality audio/video recordings as well as the number of unique speakers and recording environments. Due to requirements to have high quality content, all the recordings are limited to studio production and TV broadcast formats. However, such

deliberate choice of video sources does not affect the content diversity and its potential applications in the real world settings of uncontrolled recording environments.

Since collected internet videos are protected by copyright, collected video database will mainly be used as internal project resource, available for extensive analyses and testing of developed algorithms and procedures on collected audio-video streams. However, the database will also include additionally generated meta-data about appearances of different speakers in the video, which can be made public in the form of anonymized log-files. At the time this report was prepared, according to the methodology described in M2.2, 2400 unique videos with characteristics listed in Table 1 were collected according to the methodology presented in detail in the report related to M2.2, using Python libraries *pytube* and *yt_dlp*, from a variety of sources. Individual videos fall into either of two categories: (1) regular news broadcasts, produced by 2 Serbian TV stations with significant market share, (2) discussion talk shows in TV format with several guest speakers. All 2400 high quality internet videos were recorded by programming scripts that ensure desired level of quality and storage efficiency, without re-encoding of the originally uploaded content. The videos can be partitioned into four groups corresponding to 4 different content providers (video productions), with 600 videos from each source, as summarized in Table 2.

Number of unique videos	2400
Total video duration	~1363 hours
Videos duration	~30 minutes
Video streams	1920x1080 @ 25 fps, high bit rate
Audio streams	~100 kbps ABR, compressed
Video container formats	.mp4; .mkv
Number of videos per content provider	600
Number of unique speakers	> 100

Table 1. Properties of downloaded video recordings.

Video content provider	Video duration [mm:ss]				Total time [hh:mm:ss]
	Minimum	Maximum	Mean	Median	
Dnevnik_RTV	1:37	44:31	25:33	25:40	255:31:54
Dnevnik_N1	25:22	168:53	37:12	34:54	369:01:04
Dobro_Jutro_TANJUG	2:47	86:50	35:26	36:07	354:28:25
Uranak_K1	4:34	81:29	38:26	39:14	384:22:46

Table 2. Content providers and durations of corresponfing videos.

For each original input video file from the downloaded collection, additional metadata about persons appearing in the video will be produced based on their face identities. Such meta data, in the form of corresponding anonymized log-files, can be stored alongside original video files in the database for further post-processing, i.e. downstream tasks. By subsequent loading of original video and its log-file pair, produced meta-data allow for creating different newly generated output videos, e.g. video frames containing only the selected speaker, or the video containing only the person's face or mouth region, including the accompanying audio sequences from the original input video. A task within Subactivity 2.4 was established to develop such a software, with the aim of post-processing of anonymized log-files associated with video entries in the core database. This assumes controlled extraction and concatenation of all specific parts of the original input video that correspond to the: 1) selected speaker (unique face identity in the log-file), and 2) defined spatial region of interest (face or mouth region that can dynamically change in the original unconstrained video input). Such video outputs or video stories are sometimes also called **talking face videos**, or mouth region videos, and in both cases require that the produced output contains only the selected single speaker with face partially or fully oriented towards the camera. The video analysis tool that will be developed assigns anonymized person identities based only on the analysis of their face images. It is assumed that neither the identities nor the number of people in a scene are known in advance. In the case of regular news broadcasts as well as discussion talk shows, some speakers are regularly appearing in all video files from the same source (content provider), which allows for speaker dependent corpora development. At the same time there is also a large variety of other native speakers in each of the videos. In particular, recordings also include short reportages and outdoor recordings in forms of interviews, which are particularly suitable for providing talking face videos from uncontrolled environments. The proposed workflow of the video processing tool used for the production of talking face videos from VideoBase is shown in Fig. 2.

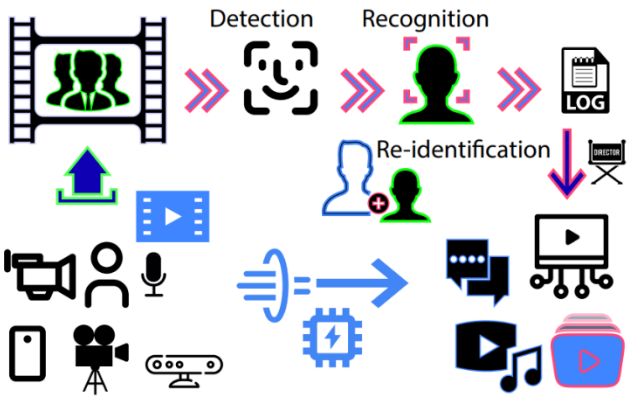

Figure 2. The proposed workflow of the video processing tool used for production of talking face videos.

During the initial video analysis the system will automatically assign unique a numerical identifier (personID) to each identified person in the scene (valid detection), which will allow for efficient post processing in the later stage. Besides face re-identification, the process also involves spatial and temporal localization of each face image, extraction of specific face landmark points and person's age and gender estimation. Based on a agenerated log-file, users will be able to select a specific person for further processing (audiovisual corpora extraction), and if necessary at the same time also filter out multiple

identities which correspond to the same person (resolve multiple identities assigned to each person due to high sensitivity of face recognition embeddings used in the face re-identification task). For example, “Person1” in one frame can also appear as “Person2” in another frame during the same video, due to different scale and orientation of the face images. On the other hand, “Person3” can be correctly recognized throughout the whole video. People who are correctly identified as single persons throughout the video which allow for most efficient production of audiovisual corpora (generation of the corresponding talking face video, or mouth region video) through selection of this identity in the later post processing, i.e. log-file analysis stage.

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.