Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Milestone index:	M3.2
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 August 19th 2025, with participation of all team members. The focus of the project meeting was (1) defining further research directions and implementation plans based on the data recorded or collected from the Internet for the purposes of developing the **AI-SPEAK speech corpus** and **AI-SPEAK Internet Video Database** (2) defining further activities on exploitation of both corpora in the next phases of the project, particularly in the context of Workpackage 4, which has begun in July 2025 (M19).

I. Implementation report and plans related to WP3 - ASR Implementation and experiments

I.2. Subactivity 3.2 - Speech recognition from audio and video

Our efforts on audio-visual speech recognition task started with a systematic review of the relevant literature in order to identify established methodologies, open challenges, and research gaps. Existing research in automatic speech recognition demonstrates that modern deep-learning-based systems achieve high accuracy in controlled acoustic conditions (even unimodal audio models provide high performance) but remain sensitive to background noise and speaker and speech style variability. The literature contains relatively few studies that examine audio-visual speech recognition of Slavonic languages (e.g. Russian and Czech), but there is a lack of literature, pre-trained models and open datasets for South Slavic languages.

Although AV-HuBERT is currently recognized as one of the most widely used multilingual, audio-visual, general-purpose state-of-the-art models, it is not pre-trained for Serbian, and fine-tuning it would require substantial computational resources as well as a large amount of labeled audiovisual data. Given these constraints, our current research activities are focused on experiments with the Zero-AVSR model (Zero-Shot Audio-Visual Speech Recognition with LLMs by Learning Language-Agnostic Speech Representations). This approach is attractive for low-resource languages, as it aims to learn language-agnostic audio-visual speech representations and leverages large language models for decoding, enabling zero-shot transfer to previously unseen languages. In the backbone, it utilizes AV-HuBERT

architecture to romanize detected speech. Our current activities include model fine-tuning using our internally recorded dataset, as well as further data preparation aimed at fully aligning the processing pipeline with the Zero-AVSR framework. We performed initial fine-tuning of the model using our internally recorded dataset, and the results are promising. Further activities will include additional data preparation to fully align with the processing pipeline described in the Zero-AVSR paper.

II. Implementation report and plans related to WP4 - Avatar development and adaptation to Serbian

II.1. Subactivity 4.1 - Avatar adaptation to Serbian

One of the primary objectives of this project is the development of an automated system for animating a speaking avatar in the Serbian language. The envisioned system relies on a deep learning model that takes audio and textual input and predicts blendshape coefficients, which are then applied to a 3D avatar to generate realistic speech-driven facial animation. This approach aims to replace manual animation with a data-driven method capable of modeling subtle and complex facial movements associated with speech production.

The initial phase of the work relied on an existing audiovisual dataset recorded in an anechoic chamber. The acoustically controlled environment ensured high-quality speech recordings, and the corresponding video data was considered a suitable source for extracting facial motion information. The original assumption was that high-resolution 2D video recordings would be sufficient for deriving facial motion parameters required for training the animation model.

Facial landmark extraction was performed using MediaPipe Face Mesh, a real-time computer vision framework capable of detecting a dense set of 468 three-dimensional facial landmarks. These landmarks cover the entire facial region, including the eyes, eyebrows, nose, lips, and jawline. For speech animation, particular importance lies in the lip contours, mouth corners, and jaw region, as these areas encode the most relevant articulatory movements. From these landmarks, blendshape-like coefficients were derived using MediaPipe Blendshape V2. Visual inspection of the extracted coefficients showed patterns consistent with the speaker's articulation, suggesting that these parameters might serve as suitable targets for neural network training.

In parallel, a 3D avatar was created based on the facial features of one of the recorded speakers. The avatar was rigged according to the ARKit 52 blendshape standard, a widely used system in facial animation where each blendshape corresponds to a predefined facial muscle movement or expression. This standardization allowed a direct mapping between extracted coefficients and avatar control parameters.

However, when the blendshape coefficients derived from 2D video were transferred to the 3D avatar, the resulting animation quality was unsatisfactory. Lip movements lacked depth, jaw motion appeared unnatural, and articulation did not convincingly reflect speech production. This outcome revealed a key limitation: facial motion parameters inferred from 2D video are insufficient for realistic 3D speech animation. The underlying issue stems from the fact that MediaPipe primarily operates on RGB imagery, making it difficult to accurately capture depth-related movements such as lip protrusion and forward

jaw motion. Speech articulation is inherently three-dimensional, and these missing cues significantly degrade animation realism.

To address this limitation, a new dataset was recorded using the FaceCap application on an iPhone, which utilizes the front-facing depth camera to directly capture ARKit blendshape coefficients in 3D. This method provides physically meaningful facial motion data without the need to reconstruct depth from 2D images. A total of 180 sentences were recorded from two speakers: one speaker whose facial identity was used to create the avatar, and another independent speaker. Including multiple speakers allows the system to learn speaker-independent motion patterns, ensuring that facial animation does not depend on a specific voice timbre and can be driven by synthesized speech in future real-time applications.

When these depth-based blendshape coefficients were applied to the avatar, the resulting animation quality improved significantly. Lip and jaw movements appeared more natural and consistent with speech articulation, confirming the importance of depth information in speech-driven facial animation. This dataset therefore forms a reliable basis for training deep learning models to predict blendshape coefficients from speech signals.

Beyond audio input, the system also incorporates textual input, which motivates the use of visemes — visual counterparts of phonemes. Serbian has a near one-to-one grapheme-to-phoneme correspondence, which simplifies phoneme extraction from text. However, viseme definitions for Serbian have not been previously established.

In contrast, English has several predefined viseme systems, typically containing between 12 and over 20 viseme categories, while English itself includes around 40 phonemes. Because multiple phonemes share similar visual realizations, viseme sets are smaller than phoneme inventories. These English viseme systems are widely used in lip-reading research and speech animation.

Directly applying English viseme sets to Serbian presents challenges. Although Serbian contains approximately 30 phonemes, certain Serbian phonemes — such as LJ, NJ, Đ, Ć, and DŽ — lack clear visual equivalents in English viseme systems. Their articulation involves specific palatal or affricate features whose visual characteristics are not distinctly represented in English-based groupings. Consequently, mapping Serbian phonemes to English visemes would result in loss of articulatory detail and reduced animation realism.

We therefore investigated the possibility of deriving viseme groupings directly from data using clustering techniques. Methods such as K-means, hierarchical clustering, DBSCAN, and spectral clustering were applied to the extracted blendshape coefficients. However, these unsupervised approaches did not yield visually or articulatorily meaningful groupings.

As a result, a linguistically guided approach was adopted. The proposed Serbian viseme system is based on place of articulation and visual analysis of recorded speech. The resulting viseme categories include bilabial, labiodental, dental/alveolar, postalveolar/affricate, and velar consonant groups, as well as four

vowel-based groups and a category for approximants. This system is considered a working hypothesis and will be further evaluated and refined in future research.

Table 1. Visemes and corresponding phonemes

Visemes	Phonemes
V0 – bilabial	p, b, m
V1 – labiodental	f, v
V2 – open vowel	a
V3 – front vowel	e, i
V4 – vowel o	o
V5 – vowel u	u
V5 – alveolar	t, d, s, z, n, c
V6 – postalveolar	č, ć, dž, đ, š,ž
V7 – velar	k, g, h
V8 – other	j, lj, nj, l, r

III. Postprocessing and follow-up activities related to AI-SPEAK Internet Video Database

AI-SPEAK Internet Video Database or VideoBase has been analyzed in the context of content diversity and scene scenario settings. It was concluded that parts of the database related to news broadcasts and reportages, including one-on-one TV interviews, are particularly suitable for the extraction of high quality model training data. On the other hand, it was discussed how to overcome the remaining challenges related to scenarios in which there is a presence of multiple persons in the scene and need to identify the active speaker who is producing the corresponding audio recording. Namely, a developed custom video analysis tool for training data generation proved to be a robust solution for the extraction of individual face images of the same person throughout the analyzed video. However, in cases when multiple persons are present in the scene it can happen that some of them are not speaking, although their facial images are still being recorded, which leads to audio recordings paired with the wrong person in the scene. These situations are usually coming from video recordings in which multiple persons are participating in talk show or round table formats, or outdoor interviews in one-on-many interaction format. Although common, it was concluded that these scenarios are not substantially affecting the applicability of designed internet video database, since many of the included videos always have a significant amount of content that is not affected by described scenarios. On the other hand, conclusion was that these anomalies in the extracted model training data could be corrected by development of an additional post processing module that would perform additional analysis of extracted face video and detect situations in which the person is not moving the mouth or lips, but there is still active speech in the accompanying audio recording. By removing such video frames, it would be possible to also utilize these types of recordings and extracted face videos.

When it comes to further post processing of the video analysis results, it should be noted that for each video in the internet video database, the designed analysis tool is producing a generic log of person appearances and localization of person's face and face landmarks. These generic data allow multiple rounds of post processing and extraction of different face related information in custom video format for each identified person. Thus it was concluded that besides extraction of face videos, there should

also be post processing options for the extraction of mouth region for each selected identity. These should also include additional options like adding an additional spatial margin around each region of interest, as well as the choice of an adequate scaling of these images in each frame. Namely, the same person can have quite different visual appearance over the whole video duration, leading to face or mouth region images of varying scale and number of pixels. Since it makes the later model training and feature extraction more challenging, it was decided that the post processing tool should also implement additional options for scaling adaptation of output results. In case of such post processing there are usually two different approaches. Either to keep the original aspect ratio of each face image throughout the face video recording, leading to additional padding of face image signal by appropriate number of zeros or by using other edge effects mitigation mirroring techniques; or to scale each image to some common dimension characteristic for the whole video duration. It was decided that both options should be implemented, since it would allow to produce generic face and mouth region videos in original image resolution, as well model training ready images with predefined scale corresponding e.g. to average size of face or mouth region appearance of the selected person throughout the whole video recording.

Since the created internet video database can be considered as large scale, after discussion of the computational and time requirements for video analysis of the whole database and post processing of computed results it was concluded that further follow-up activities should be directed in two complementary directions. One is the design of a systematic and unified approach for labeling and storing of video analysis results, which should be reused multiple times by different post processing techniques. This also includes development of additional programming interfaces and batch processing scripts that would allow automated data curation and browsing of generated data. An important aspect of this design direction is also development of code implementations that would allow full utilization of available hardware resources, by enabling simultaneous running of multiple processing models in parallel. The second identified direction was the need to re-design some of the elements of the original video analysis tool and post processing programmes in order to utilize hardware accelerated video encoding and decoding, when such is available or applicable. Namely, during post processing it is possible to generate a video story for each selected person or identity in the original video. This post processing output format assumes full frame snapshots of the original video and their export as a visualization of the person's appearance throughout the video. Since post processing does not include any additional scaling or change of the input frame it is particularly suitable to be produced without additional re-encoding of the original input video. On the other hand, any output involving e.g. only face or mouth image region requires additional scaling or processing of original input frames, i.e. re-encoding of the input video. Since these operations are quite common in all planned post processing scenarios, it was concluded that special attention should be given to design of efficient output video generation pipelines, which could be designed using custom compiled development libraries.

Developed post processing tools were tested using developed interactive control graphical user interface, it was also concluded that application of such GUI tool that requires human interaction would not be the best solution for processing of numerous recognized identities in analyzed videos in internet video database. Therefore, it was concluded that it would be good to develop an additional post processing tool that would allow automation of this process based on machine interpretation of video analysis results and automated selection of most valuable speakers in the video, e.g. ones that are the most present or among the top three by video presence duration. This tool would allow greater efficiency of developed post processing, as well as easier extension of the existing internet video database in the future.

At the end, it was discussed how designed post processing tools preserve privacy of the speakers in the original input video. It was concluded that by providing anonymized video analysis results in the form of the corresponding log files it is possible to easily overcome the potential problem of video content redistribution. Thus, in any model training scenario the users that would like to distribute their training data could do that by providing lists of urls of publicly available videos and accompanying video analysis log files, which could be reused and utilized multiple times in later post processing. Besides the privacy preserving nature of these output formats, it was also concluded that they enable easier distribution of memory demanding video database and corresponding video analysis results, without affecting their usability and later post processing steps.