Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Deliverable index:	D3.2
Version:	1.0



Modules for audio-visual speech recognition

Executive Summary

This deliverable presents the research and development activities conducted within the AI-SPEAK project on audio-visual speech recognition (AVSR) for the Serbian language. AVSR transcribes spoken language from both the acoustic signal and the video of the speaker's lips, which makes recognition substantially more robust in noisy conditions than audio alone. The work builds on the Zero-AVSR framework of Yeo et al. (2025) [1], which enables speech recognition for languages with little or no audio-visual training data by predicting a language-agnostic Roman text representation and converting it into the target script with a Large Language Model (LLM).

The work addresses two challenges. First, Serbian is an extremely low-resource language for this task: it appears in the training corpus developed by the authors of AVSR with only about 0.9 hours of unlabeled video, so the pre-trained model has barely been exposed to it. Second, Serbian is digraphic, written in both Cyrillic and Latin, which complicates the conversion of the model's Roman output into a single target script. To address these, a dedicated Serbian audio-visual corpus was recorded from multiple camera angles and used to fine-tune the speech model, and a constrained Serbian-specific prompt was developed for the GPT-4o-mini de-romanizer used in the cascaded configuration.

Evaluated on unseen speakers, the fine-tuned system reduces the Serbian transcription error from 64% word error rate (WER) for the model to 32.2% WER, a level comparable to the figures the original authors report for substantially higher-resource languages such as Russian and German. Under additive acoustic noise the audio-visual system consistently outperforms an audio-only Whisper-medium baseline, with the advantage widening as conditions deteriorate. The activities described here establish a working Serbian AVSR pipeline and a reusable evaluation framework for future AI-SPEAK activities.

1. Introduction and Objectives

1.1 Motivation

Audio-visual speech recognition combines two complementary information sources: the acoustic speech signal and the visual motion of the speaker's lips. Because lip motion is unaffected by acoustic noise, jointly modeling both modalities improves recognition accuracy, particularly in challenging real-world environments. Despite considerable progress, AVSR has been developed and evaluated almost exclusively for high-resource languages, mainly English, because labeled audio-visual data with transcriptions is scarce for most languages.

Serbian is a clear example of this limitation. High-quality labeled audio-visual data for Serbian is essentially unavailable, and the language is present only marginally (about 0.9 hours of unlabeled, automatically transcribed video, alongside 2.6 hours of closely related Croatian) in the multilingual corpus used to train the framework adopted here. This places Serbian in a zero-shot, very-low-resource regime and motivates both a dedicated data-collection effort and a recognition approach that can exploit knowledge transferred from other languages.

1.2 Objectives

The main objective of this task was to develop and evaluate an audio-visual speech recognition system for Serbian on top of the Zero-AVSR framework of Yeo et al. (2025). Specific objectives included:

- adaptation of the Zero-AVSR pipeline to Serbian in a cascaded configuration that uses an LLM as de-romanizer;
- fine-tuning of the audio-visual speech model on Serbian data recorded from multiple camera angles (AI-SPEAK multimodal multilingual corpus recorded within this project);
- design of a Serbian-specific de-romanization prompt that resolves the Cyrillic/Latin ambiguity;
- objective evaluation of recognition accuracy on unseen speakers, including robustness to acoustic noise and comparison with an audio-only baseline.

2. The Zero-AVSR Framework

Conventional multilingual AVSR systems predict language-specific letters (graphemes) directly and can therefore only recognize languages seen during training. Zero-AVSR removes this restriction through a two-part idea.

2.1 Roman Text as a Language-Agnostic Representation

Instead of predicting native letters, the speech model predicts Roman text, a phonetic spelling of how the speech sounds, using a small fixed alphabet shared across all languages. Because the alphabet is fixed and pronunciation-based, the model can transcribe what it hears even for a language it has barely seen; it only needs to have learned a sufficiently broad range of speech sounds. The component performing this is the Audio-Visual Speech Romanizer (AV-Romanizer).

2.2 Large Language Models as De-Romanizers

The Roman text must then be converted into the target language's real script. The key observation of the framework is that pre-trained LLMs already know this mapping for hundreds of languages and can convert a romanized sentence into native graphemes with no additional training. This removes the need for a separate language model per language, i.e. there is no need to train a separate language model since the foundation model (or other models performing this part of the pipeline) can be used for this task.

2.3 Two Framework Variants

The framework given in Yeo et al. (2025) can be used in two configurations. In Cascaded Zero-AVSR, the AV-Romanizer produces Roman text that is inserted into a prompt and sent to an off-the-shelf LLM (here GPT-4o-mini), which returns the native-script transcription; the LLM is never fine-tuned, so even an API-only model can be used. In the unified Zero-AVSR model, the AV-Romanizer's speech features are fed directly into a LLaMA-3.2-3B model through a trained adapter, and the LLM is fine-tuned with LoRA. The unified model performs better but requires the open LLM weights and multi-GPU training. This project uses the cascaded configuration.

2.4 The MARC Dataset

The AV-Romanizer in Yeo et al. (2025) was trained on the Multilingual Audio-Visual Romanized Corpus (MARC): approximately 2,916 hours of audio-visual speech across 82 languages, with both native-script and Roman transcriptions. MARC combines four sources: the labeled LRS3 [2] and MuAViC [3] corpora, and the much larger unlabeled VoxCeleb2 [4] and AVSpeech [5] datasets, which are automatically labeled using a language-identification model and an ASR model. MuAViC contributes only nine languages; the remaining languages, including the small amount of Serbian (about 0.9 hours), come from the automatically labeled VoxCeleb2 and AVSpeech data. The amount of data per language is therefore highly uneven, from hundreds of hours for English, German and Portuguese down to under one hour for Serbian, which is precisely why Serbian behaves as a low-resource case in this work.

3. System Architecture and Training

The framework is organized in two stages, corresponding to the two model components.

3.1 The AV-Romanizer

The AV-Romanizer is built on a pre-trained AV-HuBERT Large encoder. The audio is converted to log-filterbank features and the video to lip-region frames; the two streams are fused by concatenation and passed through a 24-layer Transformer encoder (embedding dimension 1024). A single linear layer with 32 outputs (Roman characters plus a blank symbol) predicts the Roman text, and the model is trained with the Connectionist Temporal Classification (CTC) objective. Training uses modality dropout, so the model copes when one stream is missing, and acoustic-noise augmentation, which contributes to the system's later noise robustness.

3.2 The Unified Zero-AVSR Model

In the unified model, the frozen AV-Romanizer is connected to LLaMA-3.2-3B through a 1-D convolutional length compressor, which halves the frame rate, and a two-layer adapter that maps the 1024-dimensional speech features into the LLM's 3072-dimensional embedding space. Only the adapter, the length compressor and LoRA weights are trained, using a next-token language-modeling loss. This stage was not used in the present project, where GPT-4o-mini replaces the fine-tuned LLM entirely. The cascaded configuration was chosen for this work because it requires no training of the language model and no local LLM weights: a strong foundation model accessed through an API performs the de-romanization directly. This let us reach recognition quality comparable to other languages without the multi-GPU fine-tuning the unified model requires.

4. Performance Reported in the Original Work

The framework was evaluated by its authors on the MuAViC benchmark, reporting word error rate (WER) averaged over languages. MuAViC covers nine languages and does not include Serbian, so it is not a benchmark for the language addressed here. To the authors' best knowledge, no audio-visual speech recognition model has previously been adapted to Serbian; the system described in this deliverable is therefore the first of its kind for Serbian, and no directly comparable Serbian AVSR baseline (such as AV-HuBERT) is available. The principal results are reproduced below.

Table 1. Comparison of different model performances on the MuAViC dataset.

Method	AV data (h)	# Langs	Avg WER	Type
AV-HuBERT	1,200	9	35.0%	non zero-shot
XLAVS-R 2B	1,200	9	23.6%	non zero-shot
MMS Zero-shot (audio only)	~40K	1,078+	38.9%	zero-shot ASR

Method	AV data (h)	# Langs	Avg WER	Type
Cascaded Zero-AVSR	2,916	82+	30.2%	zero-shot AVSR
Zero-AVSR (LLaMA-3.2-3B)	2,916	82+	25.2%	zero-shot AVSR

Two findings of the original work are directly relevant here. The choice of de-romanizer strongly affects the cascaded configuration: average error improves with LLM size, and GPT-4o-mini gave the best results among the LLMs evaluated, which is why it is used in this project. In addition, linguistic similarity helps: adding data from languages in the same family as the target improves performance more than adding unrelated languages, which is encouraging for Serbian given the presence of related Slavic data in MARC.

5. Serbian AVSR System

This section describes the data, pre-processing, recognition pipeline, model fine-tuning, de-romanization prompt and evaluation methodology developed for Serbian.

5.1 Serbian Audio-Visual Corpus (Training Data)

Fine-tuning used the Serbian audio-visual database recorded earlier in the AI-SPEAK project and released publicly on Kaggle[1]. As this corpus is documented in deliverable D2.1, only its scale is summarized here. It was recorded from 30 speakers and three synchronized camera angles (a frontal view and left and right side views), in both Serbian and English, at roughly five hours per language. Throughout this deliverable, the Serbian portion of this corpus used to adapt the model is referred to as the fine-tuning set (FT set).

5.2 Evaluation Data (Unseen Speakers)

The recognition results reported in this deliverable were not measured on a test split taken from the corpus above. So that the entire corpus could be used for fine-tuning, no speakers were held out from it for testing. Instead, to assess performance in a realistic scenario on speakers the model has never seen, evaluation used four additional speakers (three female and one male) who were recorded under the same conditions, background and setup as the corpus speakers but were not included in the corpus or the fine-tuning data. From these speakers, 51 clips were used, with twelve to thirteen Serbian sentences per speaker. Because none of these speakers appear in the training data, the reported figures reflect performance on unseen speakers, which is a

[1] https://www.kaggle.com/datasets/tijananosek/ai-speak-database-video-and-audio

more demanding evaluation setting than testing on speakers seen during training, since recognition accuracy is known to degrade on unseen speakers [6, 7].

5.3 Mouth-Region Pre-Processing

The visual backbone expects each video frame to be a tightly cropped 96×96 image of the speaker's mouth rather than the full face; producing these crops was a necessary step for the visual stream to function. The cropping procedure uses the MediaPipe FaceMesh facial-landmark detector: for every frame it locates the mouth landmarks, computes their center, cuts a fixed-size square region centered on the mouth, and resizes that region to 96×96 pixels. An earlier version of this step used the dlib 68-point landmark predictor with the same principle of a padded bounding box around the lip landmarks; the project subsequently adopted the MediaPipe-based version.

5.4 Recognition Pipeline

Each Serbian utterance is processed in three automated steps. First, the AV-Romanizer transcribes the cropped video and audio into Roman text using greedy CTC decoding. Second, the hypotheses (the candidate Roman-text transcriptions produced by the decoder) are re-ordered by utterance index. Third, each romanized line is sent to GPT-4o-mini, which returns the Serbian Cyrillic transcription. The complete sequence is orchestrated end-to-end by a single driver script.

5.5 Fine-Tuning on Serbian

To improve on the pre-trained model, the AV-Romanizer was fine-tuned on the Serbian corpus. Fine-tuning starts from the pre-trained checkpoint rather than from scratch, resets the optimizer and trains at a reduced learning rate. Two fine-tuned models were produced: one trained on frontal views only, and a stronger one trained on frontal plus left and right side views.

5.6 Serbian-Specific De-Romanization Prompt

The instruction sent to the LLM was rewritten specifically for Serbian. The original generic instruction simply asks the model to "convert Roman text into the target language," which is ambiguous for Serbian: the language is written in both Cyrillic and Latin, and the romanized input already resembles Latin Serbian without diacritics. The adapted prompt fixes a single unambiguous target (Cyrillic), forbids changing or reordering words, and limits the model to restoring diacritics and transliterating. GPT-4o-mini was chosen deliberately over larger models. De-romanization is a constrained transliteration task rather than open-ended generation: the model must not reorder or rewrite words, only map the romanized Latin text to the correct Serbian Cyrillic and restore diacritics. This is well within the capability of a small, fast and inexpensive model, and prior work has shown that models of this class handle Serbian text-

processing tasks reliably [8]. Larger models (e.g. GPT-4 or comparable) would add latency and cost without a corresponding benefit for this narrow task. The full text of both prompts is given in Appendix A.

5.7 Evaluation Methodology

A driver script discovers every test condition (noise type and signal-to-noise ratio), builds the corresponding manifest, and runs the full romanize, sort and de-romanize pipeline. A scoring script then normalizes reference and predicted text, transliterates the Cyrillic predictions to Latin for consistent comparison, and computes WER with the jiwer library. As an audio-only reference, a Whisper-medium model was evaluated over the same conditions, with noise mixed at controlled SNR levels. Whisper-medium offers a good balance between recognition accuracy and model size, which keeps it practical to deploy and makes it a representative, reproducible audio-only baseline. It is also an established model for Serbian ASR, having been used as the basis for a fine-tuned Serbian recognition system in prior work [8].

6. Experimental Evaluation

All figures below are measured on the held-out unseen-speaker evaluation set described in Section 5.2.

6.1 Effect of Fine-Tuning

Table 2 reports Serbian WER before and after fine-tuning, for both the raw Roman output of the AV-Romanizer and the final Serbian Cyrillic produced after GPT de-romanization. Here, *Roman WER* denotes the word error rate of the AV-Romanizer's raw romanized output, measured before de-romanization, while *Serbian (GPT) WER* is the error of the final Serbian Cyrillic transcription after the LLM step.

Table 2. WER for fine-tuned models

Model	Roman WER	Serbian (GPT) WER
Pre-trained model (no Serbian fine-tuning)	83.9%	64.0%
Fine-tuned, frontal only	41.2%	43.9%
Fine-tuned, frontal + side views	29.5%	32.2%

Even the pre-trained model, exposed to under one hour of unlabeled Serbian, produces a recognizable transcription (64% WER), confirming that the language-agnostic representation

transfers to Serbian. A modest amount of dedicated fine-tuning data halves the WER, and the inclusion of side-view camera angles gives a further substantial gain, bringing the final Cyrillic WER down to about 32% on speakers the model never saw during training.

6.2 Comparison with Reported Languages

Placing the fine-tuned Serbian result alongside the per-language figures reported by the original authors (full 82-language models) gives useful context, together with the amount of MARC data each language had.

Table 3. WER on MuAViC reported for different languages in the original work versus this project's fine-tuned Serbian result

Language	MARC hours	Cascaded WER	Zero-AVSR WER	Family
Portuguese	408.0	20.2%	15.4%	Romance
German	327.5	29.3%	27.8%	Germanic
Russian	123.8	42.9%	32.6%	Slavic
Greek	21.5	47.2%	38.4%	Hellenic
Serbian (ours, fine-tuned)	0.9 + FT set	32.2%	n/a	Slavic

The comparison is encouraging: the fine-tuned Serbian system reaches a 32.2% WER comparable to Russian (32.6% with the full model) and close to German (27.8%), despite those languages having roughly 124 and 327 hours in MARC against well under one hour for Serbian plus a small fine-tuning set. Two points should be kept in mind when reading this comparison: the Serbian figure is measured on the project's own unseen-speaker set rather than on MuAViC, and it uses GPT-4o-mini, the stronger of the two de-romanizer options.

6.3 Robustness to Acoustic Noise

Evaluating speech recognition under additive noise at a range of signal-to-noise ratios is a standard robustness protocol in the audio-visual speech literature, and is one of the main motivations for using the visual modality [9]. The same methodology is followed here so that the results are comparable with established practice. The best Serbian model was evaluated with three noise types (babble, pink, white) mixed into the audio at signal-to-noise ratios from −7.5 dB (very noisy) to +12.5 dB (mildly noisy). The three noise signals were taken from the standard NOISEX-92 noise database, allowing the experiment to be reproduced easily.

Table 4. Serbian AVSR WER (%) under babble / pink / white noise across SNR levels

SNR (dB)	Babble	Pink	White	Avg
−7.5	82.16	82.16	81.87	82.06
−2.5	64.33	62.57	66.67	64.52
+2.5	40.94	46.20	50.88	46.01
+7.5	34.80	35.09	44.44	38.11
+12.5	27.49	35.67	33.33	32.16

The behavior is as expected: accuracy degrades as noise increases, and the three noise types behave similarly, with white noise marginally the most damaging at moderate SNRs.

6.4 Audio-Visual versus Audio-Only Recognition

To demonstrate the value of the visual stream, the same noisy conditions were also transcribed by an audio-only Whisper-medium model. Table 5 compares the two systems, averaged across the three noise types at each SNR.

Table 5. Mean WER across babble/pink/white noise: the Serbian AVSR model versus audio-only Whisper-medium

SNR (dB)	AVSR (audio+video)	Whisper-medium (audio only)	Δ (abs.)
−7.5	82.1%	106.0%	−23.9
−2.5	64.5%	90.8%	−26.3
+2.5	46.0%	68.4%	−22.4
+7.5	38.1%	49.6%	−11.5
+12.5	32.2%	42.9%	−10.7

This is the clearest illustration of the project's central point. At mild noise the two systems are within about eleven points, but as the audio worsens the audio-only model collapses: at −7.5 dB its WER exceeds 100% (its output is effectively noise), while the audio-visual model degrades far more gradually, because the lip-movement information is unaffected by acoustic noise. Whisper-

medium here is a general multilingual model without Serbian fine-tuning, so part of the gap also reflects that difference.

7. Conclusions

This deliverable presented the research activities conducted within the AI-SPEAK project on audio-visual speech recognition for Serbian. The Zero-AVSR framework was adapted to Serbian in a cascaded configuration using GPT-4o-mini as de-romanizer with a Serbian-specific prompt that resolves the Cyrillic/Latin ambiguity, and the audio-visual speech model was fine-tuned on the project's Serbian corpus.

Evaluated on unseen speakers, fine-tuning reduced the Serbian transcription error from 64.0% to 32.2% WER, a level comparable to the original authors' results for Russian (32.6%) and German (27.8%). This is notable because those two languages were far better resourced in the training corpus than Serbian: MARC contains about 124 hours of Russian and 328 hours of German audio-visual speech, whereas Serbian was represented by only 0.9 hours in MARC, together with roughly five hours of fine-tuning data recorded for this project. Evaluation under additive acoustic noise confirmed that the audio-visual system is markedly more robust than an audio-only baseline, with the advantage increasing as conditions deteriorate.

Overall, the work establishes a functioning Serbian AVSR pipeline and a reusable evaluation framework. It provides a solid foundation for the wider Serbian speech-technology community to build on: further gains can be expected by recording additional data and adapting the system to specific application scenarios of interest.

Appendix A. De-Romanization Prompts

The de-romanization step sends the AV-Romanizer's predicted Roman text to GPT-4o-mini and asks it to return the native script. The exact instruction matters and had to be adapted for Serbian.

A.1 Original prompt (authors)

```
This is the romanized <LANG> transcription '<TEXT>'
Convert Roman Text into <LANG>. Don't answer additional text.
```

A.2 Adapted prompt (this project)

```
You are a Serbian orthography expert.
Task: Convert the following Serbian sentence written in Latin
WITHOUT diacritics into Serbian CYRILLIC.

Rules:
- Output MUST be Serbian Cyrillic only (no Latin letters).
- Preserve the exact same words/order and punctuation.
- Only add the correct diacritics when needed (č, ć, š, ž, đ),
  but do NOT change words beyond that.
- Do NOT paraphrase, do NOT add or remove words.
- Output only the converted sentence, nothing else.

Input: <TEXT>
Output:
```

A.3 Rationale

Serbian is digraphic: it is written in both Cyrillic and Latin, fully interchangeably. The romanized text the AV-Romanizer produces already resembles Latin Serbian with the diacritics stripped (for example, "c" for both "č" and "ć"). Under the generic instruction the model frequently had no clear target: it might leave the text in Latin, mix scripts, or treat the already-Serbian-looking input as a paraphrasing task. The adapted prompt fixes a single unambiguous target (Cyrillic), forbids changing the words, and limits the model's task to restoring diacritics and transliterating. In practice this reduced spurious rewrites and produced more consistent output.

References

1. Yeo, J.H., Kim, M., Kim, C.W., Petridis, S., Ro, Y.M.: Zero-AVSR: Zero-Shot Audio-Visual Speech Recognition with LLMs by Learning Language-Agnostic Speech Representations. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6693–6703 (2025). https://doi.org/10.1109/ICCV51701.2025.00630.
2. Afouras, T., Chung, J.S., Zisserman, A.: LRS3-TED: a large-scale dataset for visual speech recognition, http://arxiv.org/abs/1809.00496, (2018). https://doi.org/10.48550/arXiv.1809.00496.
3. Anwar, M., Shi, B., Goswami, V., Hsu, W.-N., Pino, J., Wang, C.: MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation. In: INTERSPEECH 2023. pp. 4064–4068. ISCA (2023). https://doi.org/10.21437/Interspeech.2023-2279.
4. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep Speaker Recognition. In: Interspeech 2018. pp. 1086–1090. ISCA (2018). https://doi.org/10.21437/Interspeech.2018-1929.
5. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Trans. Graph. 37, 1–11 (2018). https://doi.org/10.1145/3197517.3201357.
6. Huang, X., Lee, K.F.: On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. IEEE Trans. Speech Audio Process. 1, 150–157 (1993). https://doi.org/10.1109/89.222875.
7. Kim, M., Kim, H., Ro, Y.M.: Speaker-Adaptive Lip Reading with User-Dependent Padding. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., and Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 576–593. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_33.
8. Stanojev, V., Nosek, T., Suzić, S., Pekar, D., Delić, V., Sečujski, M.: Improving Whisper-Based Serbian ASR Using Synthetic Speech. In: Speech and Computer. pp. 118–129. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-07959-6_9.
9. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M.: Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096889.