Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Deliverable index:	D3.3
Version:	1.0



Modules for speech re-synthesis from video

Executive Summary

This deliverable presents a review of state-of-the-art lip-to-speech (also referred to as video-to-speech) synthesis frameworks and assesses their applicability to the Serbian language within the AI-SPEAK project. Lip-to-speech synthesis aims to reconstruct an intelligible and natural speech signal from a silent video of a speaker, relying solely on visual cues such as lip and facial movements. The task is intrinsically difficult: the mapping from video to audio is one-to-many, because several distinct phonemes share an identical visual appearance (homophenes), and silent video carries no information about speaker-specific acoustic attributes such as pitch and timbre.

Six representative frameworks are reviewed, spanning the main directions of the field: speech-unit reconstruction, multi-task learning with automatic speech recognition (ASR) supervision, diffusion-based generation, lightweight feature-based reconstruction, and the most recent speech-decomposition approaches based on rectified flow and source-filter modeling. For each framework, the architecture, training data, evaluation methodology, and reported performance are summarized.

The central finding of this deliverable is that none of the reviewed frameworks could be adapted to Serbian under the resources currently available within the project. The reasons are systematic rather than incidental: every competitive framework depends on hundreds of hours of in-the-wild English audio-visual data and on self-supervised and ASR components that are pre-trained on English. The available Serbian audio-visual material (the AI-SPEAK corpus) is far smaller, was recorded under controlled conditions, and is anonymized in a way that removes information required by several of the frameworks. An attempt to apply the only lightweight, low-data candidate did not reproduce the results reported by its authors even for English. The deliverable closes by identifying the Serbian phoneme-to-viseme mapping established within the project as the most concrete foundation currently available for any future Serbian lip-to-speech effort.

1. Introduction and Objectives

1.1 Task Definition and Motivation

Lip-to-speech (L2S) synthesis is the task of generating an audible speech signal directly from a silent talking-face video. It is the inverse of the speech-driven facial animation task addressed in Deliverable D4.2: rather than predicting facial motion from speech, an L2S system predicts speech from facial motion. Reliable L2S has clear applications in assistive communication for people who have lost their voice, in speech enhancement under severe acoustic noise, in audio restoration for silent or corrupted recordings, and in multimodal human-computer interaction.

The task is fundamentally ill-posed. First, the visual-to-acoustic mapping is one-to-many: many phonemes are produced with visually indistinguishable lip configurations, so the same lip movement may correspond to several different sounds. Second, a silent video provides no direct evidence of the speaker's pitch, timbre, or prosody, which are essential for natural rather than robotic speech. Modern systems therefore rely on strong contextual modeling and on large quantities of training data to resolve this ambiguity statistically.

1.2 Objective of this Deliverable

The objective of this activity was to review existing lip-to-speech synthesis frameworks and to determine whether any of them could be adapted to Serbian using the resources available within the AI-SPEAK project. The work therefore comprised three components: a survey of representative frameworks, an analysis of the data and pre-trained components each framework requires, and an assessment of the feasibility of adapting these frameworks to Serbian given the available audio-visual material.

2. Review of Lip-to-Speech Synthesis Frameworks

This section reviews six frameworks that together span the principal directions of contemporary lip-to-speech research. The first four represent the established design space; the last two reflect the most recent developments (2024-2025) and are included to ensure the review remains current.

2.1 Speech-Unit Reconstruction

Intelligible Lip-to-Speech Synthesis with Speech Units [1] reconstructs speech without relying on text transcripts. A visual front-end (ResNet-18 or AV-HuBERT applied to the cropped lip region) feeds a Conformer encoder that captures temporal context and articulation. The model produces two complementary outputs through separate predictors: a speaker-dependent predictor that estimates mel-spectrograms, and a speaker-independent predictor that estimates discrete speech units. The speech units are obtained by clustering HuBERT representations with k-means into 200 clusters and capture the linguistic content of the utterance. Both outputs drive a HiFi-

GAN vocoder that synthesizes the final waveform. The system is trained on LRS2 (223 h) and LRS3 (433 h) and evaluated using STOI/ESTOI (intelligibility), PESQ (quality), word error rate (WER), and mean opinion score (MOS). Reported WER is 50.4% with a ResNet-18 front-end and 29.7% with AV-HuBERT.

2.2 Multi-Task Learning with ASR Supervision

Lip-to-Speech Synthesis in the Wild with Multi-Task Learning [2] processes the lip video with a ResNet-18 front-end followed by a Conformer that models temporal context and articulation. The network is trained with multiple objectives simultaneously: it predicts the spoken text (a lip-reading task), predicts mel-spectrograms, and uses an ASR model as an additional source of supervision. Only the mel-spectrograms are used to reproduce speech, via a Griffin-Lim vocoder; the auxiliary tasks serve to stabilize training and improve alignment. The approach targets unconstrained (in-the-wild) data and reports comparatively high error rates, with WER of 60% on LRS2, 66% on LRS3, and 13.86% on the word-level LRW benchmark.

2.3 Diffusion-Based Generation

LipVoicer [3] generates speech using a diffusion model. A mel-spectrogram generator (MelGen) is conditioned on the cropped lip region together with a still image of the speaker's face, while a separate lip-reading model predicts the spoken text. An ASR model is then used as a classifier-guidance mechanism that synchronizes the diffusion process with the predicted text. The system is trained on LRS2 and LRS3 and evaluated with MOS (covering intelligibility, naturalness, quality, and synchronization), WER, DNSMOS, STOI-Net, and SyncNet-based synchronization measures (LSE-C and LSE-D). It reports substantially lower error rates than earlier work, with WER of 17.8% on LRS2 and 21.4% on LRS3.

2.4 Lightweight Feature-Based Reconstruction

Lip2Speech: Lightweight Multi-Speaker Speech Reconstruction with Gabor Features [4] departs from heavy end-to-end models. The authors argue that convolutional and adversarial architectures are computationally expensive, opaque, and weak in speaker independence, and instead build a lightweight pipeline around interpretable Gabor-based visual features to enable multi-speaker reconstruction with reduced data and compute requirements. Because of its low-resource orientation, this framework is the most plausible candidate for a language such as Serbian. However, as discussed in Section 4, attempts to reproduce it within the project did not match the results reported by its authors, even for English.

2.5 Recent Developments

V2SFlow: Video-to-Speech Generation with Speech Decomposition and Rectified Flow [5] addresses the degradation of earlier systems on unconstrained data by decomposing the speech

signal into separate subspaces (content, pitch, and speaker information), each predicted directly from the visual input. A rectified-flow matching decoder built on a Transformer then maps noise to the target speech distribution. The authors report performance that surpasses prior state of the art and, on some measures, the naturalness of ground-truth utterances.

RESOUND: Speech Reconstruction from Silent Videos via Acoustic-Semantic Decomposed Modeling [6] (Interspeech 2025) applies source-filter theory, splitting the problem into an acoustic path that predicts prosody and a semantic path that extracts linguistic features, so that each representation can be optimized independently. Notably, it also integrates discrete speech units, the same unsupervised representation used in [1], alongside mel-spectrograms during waveform generation, confirming that speech units have become a dominant building block of the field.

Table 1 summarizes the reviewed frameworks. Across the entire design space, two properties are constant: every framework is trained and evaluated on large English audio-visual corpora, and every competitive framework depends on self-supervised or ASR components that are themselves English-specific.

Table 1. Overview of the reviewed lip-to-speech synthesis frameworks.

Framework	Core approach	Training data	Key dependencies
Speech Units	ResNet/AV-HuBERT + Conformer; mel + HuBERT units; HiFi-GAN	LRS2, LRS3	HuBERT units (English)
In the Wild	ResNet + Conformer; multi-task (text + mel); Griffin-Lim	LRS2, LRS3, LRW	Lip-reading + ASR (English)
LipVoicer	Diffusion (MelGen) + ASR classifier guidance	LRS2 + LRS3 (~600 h)	Lip-reading + ASR; face image
Gabor	Lightweight Gabor visual features; multi-speaker	GRID	(not reproducible here)
V2SFlow	Speech decomposition + rectified flow Transformer	LRS3	AV front-end (English)
RESOUND	Source-filter; acoustic + semantic paths; speech units	LRS3	HuBERT units (English)

3. Available Serbian Audio-Visual Material

The audio-visual material available for this activity is the AI-SPEAK corpus, the bilingual (Serbian and English) audio-visual database produced earlier in the project and released publicly [7]. As the corpus is documented in prior deliverables, only the properties relevant to the present analysis are summarized here. It was recorded from 30 speakers and three synchronized camera angles (a frontal view together with left and right side views), in both Serbian and English, amounting to roughly five hours per language.

For anonymity, the released video preserves only the speaker's lip region as a region of interest, while the rest of the face is pixelated and the remainder of the frame is masked. This design suits the lip-reading studies for which the corpus was created, but, as Section 4 explains, it removes information that several lip-to-speech frameworks require.

4. Applicability to Serbian

This section sets out why none of the reviewed frameworks could be adapted to Serbian using the available resources. The obstacles are systematic and mutually reinforcing.

4.1 Insufficient Training Data

Lip-to-speech is exceptionally data-hungry because the visual-to-acoustic mapping is one-to-many and must be resolved statistically. The reviewed frameworks are trained on hundreds of hours of audio-visual speech: LRS2 provides 223 h, LRS3 provides 433 h, and LipVoicer combines both for roughly 600 h. The AI-SPEAK corpus, at roughly five hours per language, is well over an order of magnitude smaller and was recorded under controlled rather than in-the-wild conditions, with a limited vocabulary and only 30 speakers. Table 2 makes the gap explicit. No large in-the-wild Serbian audio-visual corpus exists to close it.

Table 2. Scale of common lip-to-speech datasets versus the Serbian portion of the AI-SPEAK corpus.

Dataset	Scope	Speakers	Conditions
LRW	~173 h, 500-word vocabulary	1000+	In the wild (BBC)
LRS2	223 h, sentence-level	1000+	In the wild (BBC)
LRS3	433 h, sentence-level (TED)	1000+	In the wild (TED/TEDx)
Lip2Wav	120 h, 20 h/speaker	5	In the wild (YouTube)

Dataset	Scope	Speakers	Conditions
AI-SPEAK (SR)	~5 h per language	30	Controlled, anechoic; lips-only ROI

4.2 English-Specific Pre-Trained Components

Even setting data volume aside, the reviewed frameworks are built on self-supervised and ASR components that are intrinsically English. The visual front-ends rely on AV-HuBERT pre-trained on LRS3 (English); the speech-unit approaches [1, 6] use HuBERT representations clustered on English audio; and the ASR-guided approaches use recognizers trained for English. These components encode the phonetic and acoustic structure of English and cannot be expected to transfer to Serbian, whose phoneme inventory and phonotactics differ. There is no Serbian equivalent of these components at the required scale, and producing one would itself require large Serbian (audio-visual) corpora that do not exist.

4.3 The Serbian ASR Does Not Address the Missing Component

Several of the strongest frameworks (the in-the-wild multi-task model [2], LipVoicer [3], and RESOUND [6]) rely on speech recognition, either as auxiliary supervision during training or as classifier guidance during generation. It is therefore reasonable to ask whether the Serbian automatic speech recognition system available within the project, a fine-tuned Whisper model, could supply this component. It cannot, for two reasons.

First, the recognition these frameworks depend on is visual, not acoustic. The linguistic content is provided by a lip-reading model that predicts text from the silent video; the audio recognizer, where it is used at all (for example as classifier guidance on the generated audio in LipVoicer, or as auxiliary supervision in the multi-task model), operates on an acoustic signal. At inference time a lip-to-speech system has no audio input, because the audio is precisely what is being generated, so an audio ASR has nothing to transcribe from a silent video. The component that is actually missing for Serbian is a visual speech recognition (lip-reading) model, which a fine-tuned audio Whisper model does not provide.

Second, building that visual component faces the same data barrier as the rest of the task. A Serbian lip-reading model would itself require a large Serbian audio-visual corpus to train, which does not exist; the AI-SPEAK corpus, at roughly five hours per language, is far too small. The contrast with the audio-visual speech recognition work in Deliverable D3.2 is instructive: a working Serbian recognition system was obtained there, but only by fine-tuning a large multilingual model pre-trained on thousands of hours across dozens of languages, and even then the task was transcription rather than waveform reconstruction. No comparable multilingual pre-

trained backbone exists for lip-to-speech synthesis, so the transfer-learning route that made Serbian recognition feasible is not available here.

4.4 Incompatibility with the Anonymized Data Format

Several frameworks require visual information that the AI-SPEAK corpus deliberately removes. LipVoicer, for example, conditions its mel-spectrogram generator on a still image of the speaker's face to recover speaker characteristics, and speaker-embedding approaches similarly rely on facial appearance. Because the corpus preserves only the lip region and masks the remainder of the face for anonymity, this class of methods cannot be applied to it without re-recording, which would defeat the corpus's privacy design.

4.5 Failure to Reproduce the Lightweight Candidate

The lightweight Gabor-feature framework [4] was identified as the only candidate whose low-data, low-compute orientation made it potentially viable for Serbian. It was therefore selected for practical evaluation. However, the reported results could not be reproduced within the project even for English: the obtained reconstructions did not approach the quality reported by the authors. As the most promising low-resource option already failed under favorable (English, matched-data) conditions, it offered no viable fallback for the markedly harder Serbian, low-data setting.

4.6 Summary

Taken together, these obstacles are decisive. The high-performing frameworks are blocked by data scale and by English-specific pre-trained components; the recognition-guided frameworks additionally require a Serbian visual lip-reading model that does not exist and that a Serbian audio ASR cannot replace; the diffusion and speaker-embedding frameworks are incompatible with the anonymized data format; and the one lightweight candidate could not be reproduced even under favorable conditions. No combination of the available resources yields a workable Serbian lip-to-speech system at this time.

5. Serbian Visemes as a Foundation for Future Work

Although direct adaptation is not currently feasible, one project outcome provides a concrete foundation for future Serbian lip-to-speech research. The core difficulty of the task, namely that distinct phonemes are visually indistinguishable, is precisely the phenomenon captured by visemes, the equivalence classes of phonemes that share a visual realization. A well-grounded phoneme-to-viseme mapping for Serbian is therefore directly relevant to the problem.

Such a mapping has been established within the project [8] (preprint, under review). It proposes a unified Serbian phoneme-to-viseme mapping derived from three complementary sources: linguistic knowledge, human visual perception of phonemes, and visual embeddings extracted with AV-HuBERT, all evaluated on the AI-SPEAK corpus. The study reports agreement among the three approaches and a statistically significant moderate correlation between perceptual

phoneme distances and AV-HuBERT embedding distances, indicating that the learned visual space partially reflects human perceptual structure. To the authors' knowledge, it is the first systematic phoneme-to-viseme study for Serbian.

This mapping characterizes which Serbian phonemes are visually confusable and is the kind of language-specific prior that a future viseme-aware Serbian lip-to-speech system could build on, for instance to constrain or interpret the visual-to-acoustic mapping where data is scarce. It thus represents the most concrete starting point currently available for the task.

6. Conclusion

This deliverable reviewed six representative lip-to-speech synthesis frameworks, spanning speech-unit reconstruction, multi-task ASR-supervised learning, diffusion-based generation, lightweight feature-based reconstruction, and recent speech-decomposition approaches, and assessed their applicability to Serbian. The review found that the field has advanced rapidly, but in a direction that increases rather than reduces its dependence on large English corpora and English-specific self-supervised representations.

None of the reviewed frameworks could be adapted to Serbian with the resources available within the project. The obstacles are systematic: the absence of a large in-the-wild Serbian audio-visual corpus, the reliance of every competitive framework on English-specific pre-trained components, the dependence of the strongest methods on a Serbian visual lip-reading model that does not exist and that the project's audio ASR cannot substitute for, the incompatibility of the anonymized AI-SPEAK corpus with methods that require facial appearance, and the failure to reproduce the only lightweight candidate even for English.

The most constructive outcome is the identification of the project's Serbian phoneme-to-viseme mapping as a concrete foundation for future work. A realistic path toward Serbian lip-to-speech would begin not from off-the-shelf adaptation but from the construction of language-specific resources (a larger Serbian audio-visual corpus and Serbian self-supervised representations) informed by the viseme structure established here.

References

[1] Choi, J., Kim, M., & Ro, Y. M. (2023). Intelligible lip-to-speech synthesis with speech units. *arXiv preprint arXiv:2305.19603*.

[2] Kim, M., Hong, J., & Ro, Y. M. (2023, June). Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[3] Yemini, Y., Shamsian, A., Bracha, L., Gannot, S., & Fetaya, E. (2024, May). Lipvoicer: Generating speech from silent videos guided by lip reading. In *International Conference on Learning Representations* (Vol. 2024, pp. 32147-32166).

[4] Dong, Z., Xu, Y., Abel, A., & Wang, D. (2024). Lip2Speech: lightweight multi-speaker speech reconstruction with gabor features. *Applied Sciences*, *14*(2), 798.

[5] Choi, J., Kim, J. H., Li, J., Chung, J. S., & Liu, S. (2025, April). V2sflow: Video-to-speech generation with speech decomposition and rectified flow. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[6] Pham, L. K., Tran, T. V., Pham, M. T., & Nguyen, V. (2025). RESOUND: Speech Reconstruction from Silent Videos via Acoustic-Semantic Decomposed Modeling. *arXiv preprint arXiv:2505.22024*.

[7] T. Nosek et al., "AI-SPEAK database – video & audio," Kaggle. https://www.kaggle.com/datasets/tijananosek/ai-speak-database-video-and-audio

[8] S. Suzić, T. Nosek, N. Simić, V. Stanojev, N. Jakovljević, V. Delić, and M. Sečujski, "A Unified Phoneme-to-Viseme Mapping for Serbian Based on Linguistic Knowledge, Human Perception, and AV-HuBERT Visual Embeddings," preprint, under review.