Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Deliverable index:	D4.2
Version:	1.0



Modules for synthesis of facial movements from speech

Executive Summary

This deliverable presents the research and development activities conducted within the AI-SPEAK project related to speech-driven facial animation for Serbian-speaking virtual avatars. The objective of this activity was to develop a framework capable of transforming speech signals into realistic facial movements represented through ARKit-compatible blendshape coefficients.

The work addresses several scientific and technical challenges. First, Serbian is a low-resource language in the context of speech-driven facial animation, with no publicly available datasets containing synchronized speech recordings and facial motion parameters suitable for training modern deep learning models. Second, realistic facial animation requires accurate modeling of articulation, coarticulation, and temporal speech dynamics. Finally, the resulting system must provide sufficiently accurate facial motion while remaining computationally efficient for practical deployment.

To address these challenges, a dedicated Serbian audio-blendshape corpus was created using the FaceCap application. The dataset contains synchronized speech recordings and facial motion represented through 52 ARKit blendshape coefficients. Based on this corpus, a speech-driven facial animation framework was developed using Serbian-adapted Wav2Vec 2.0 representations, phoneme-aware conditioning, and a Transformer-GRU architecture for blendshape prediction.

The developed system achieved a Pearson correlation coefficient of 0.866 between predicted and reference blendshape trajectories while maintaining a Real-Time Factor (RTF) of 0.020. Subjective evaluation further confirmed that the generated facial animations achieve high perceptual quality and generalize well to previously unseen speakers.

The activities described in this deliverable establish the first comprehensive Serbian speech-driven facial animation framework and provide a foundation for future AI-SPEAK research involving conversational avatars, digital humans, and multimodal human-computer interaction.

1. Introduction and Objectives

1.1 Motivation

Speech-driven facial animation aims to automatically generate realistic facial movements from speech signals, enabling natural communication through virtual avatars and digital humans. Recent advances in artificial intelligence have significantly increased interest in conversational avatars for applications such as virtual assistants, education, healthcare, customer support, and human-computer interaction.

Despite substantial progress in speech synthesis, realistic facial animation remains a challenging task due to the complex relationship between speech acoustics and facial movements. Accurate animation requires modeling of articulation, coarticulation, and temporal speech dynamics. Furthermore, most existing solutions have been developed for high-resource languages, while Serbian remains largely unexplored in this context.

Prior to the AI-SPEAK project, no Serbian dataset containing synchronized speech recordings and ARKit-compatible facial motion data was available. This activity therefore focused on establishing the resources and methodologies required for Serbian speech-driven avatar animation.

1.2 Objectives

The main objective of this activity was the development of a speech-driven facial animation framework for Serbian avatars.

Specific objectives included:

- creation of a Serbian audio-blendshape corpus,
- development of an ARKit-compatible avatar platform,
- investigation of speech representation learning approaches for blendshape prediction,
- integration of phonetic context information,
- objective and subjective evaluation of animation quality,

The developed framework provides a foundation for future AI-SPEAK activities involving conversational avatars, digital humans, and multimodal human-machine interaction.

2. Development of the Serbian Audio-Blendshape Corpus

2.1 Motivation for Corpus Development

The success of modern speech-driven facial animation systems strongly depends on the availability of synchronized speech and facial motion data. Existing datasets are predominantly focused on English and a limited number of other high-resource languages. Furthermore, many publicly available resources provide either video recordings without explicit facial motion parameters or facial tracking information that is not directly compatible with modern avatar animation platforms.

To support research activities within AI-SPEAK and enable development of Serbian-language facial animation systems, a dedicated audio-blendshape corpus was created. The corpus was specifically designed to provide synchronized speech recordings and ARKit-compatible facial motion parameters suitable for training deep neural networks.

2.2 Recording Procedure

Data collection was performed using the FaceCap application running on Apple iPhone devices. FaceCap provides real-time facial motion capture through Apple's ARKit framework and enables export of all 52 standard ARKit blendshape coefficients synchronized with recorded speech.

Five native Serbian speakers participated in the recording process, including three female and two male speakers. Participants were selected to provide diversity in speaking style, articulation patterns, and vocal characteristics while maintaining a manageable corpus size suitable for initial model development.

The recording protocol combined both shared and speaker-specific utterances. Shared utterances were recorded by all participants and provide consistent phonetic coverage across speakers. Additional speaker-specific sentences were selected from a carefully designed corpus intended to maximize coverage of Serbian phonemes and coarticulatory contexts.

Recordings were performed under controlled indoor conditions with stable illumination and fixed camera positioning. The audio signal was subsequently resampled to 16 kHz to match the requirements of the speech processing pipeline.

2.3 Dataset Characteristics

The final corpus contains 808 utterances with a total duration of approximately 76 minutes of synchronized speech and facial motion data.

For each recording, the dataset includes:

- Speech waveform.

- Frame-level ARKit blendshape coefficients.
- Speaker metadata.
- Automatically generated phoneme alignments.
- Train-validation partition information.

Facial motion is represented through the complete set of 52 ARKit blendshape coefficients. These coefficients provide anatomically meaningful descriptions of facial movement and are widely supported by commercial animation software, game engines, and avatar creation platforms.

The average utterance duration is approximately 5.7 seconds, providing sufficient temporal context for modeling both short-range articulatory events and longer speech dynamics.

The dataset is made publicly available at Zenodo platform and can be accessed via https://zenodo.org/records/20310208.

2.4 Phoneme Alignments

To facilitate contextual speech modeling, phoneme-level alignments were generated using WhisperX forced alignment tools.

The resulting alignments provide frame-level phonetic information and enable construction of phoneme-context representations used in subsequent experiments. Alignment information also supports detailed analysis of coarticulation effects and facilitates future research involving phoneme-aware or viseme-aware animation models.

The phoneme inventory used in the corpus consists of 30 Serbian phonemes supplemented by a silence symbol, yielding a total of 31 labels.

The availability of synchronized speech recordings, facial motion parameters, and phoneme alignments makes the developed corpus a valuable resource for future Serbian speech-driven facial animation research.

3. Avatar Creation and Animation Platform

3.1 Overview

In parallel with the development of blendshape prediction models, an avatar animation platform was created within the AI-SPEAK project. The goal was to provide a realistic digital avatar compatible with ARKit blendshape coefficients and suitable for evaluation and demonstration of speech-driven facial animation systems.

The platform enables visualization of predicted facial motion and provides a basis for future integration with machine learning models and multimodal conversational systems.

3.2 Avatar Development

A realistic 3D avatar was created using a photogrammetry-based workflow. High-resolution facial photographs were processed in RealityScan to reconstruct the initial 3D geometry, followed by refinement and retopology in ZBrush and Autodesk Maya.

The resulting mesh was optimized for facial animation and real-time rendering. The avatar was equipped with all 52 standard ARKit blendshape coefficients covering major facial regions including brows, eyes, jaw, and mouth.

The use of ARKit ensures compatibility with modern facial animation workflows and allows direct utilization of blendshape coefficients predicted by machine learning models.

3.3 Animation Environment

The final avatar was integrated into an Unreal Engine application developed within the project. The application enables loading and visualization of blendshape trajectories and serves as the primary environment for qualitative evaluation of generated facial animations.

During the current project phase, facial animations are generated by loading predicted blendshape trajectories exported from the machine learning models into the Unreal Engine application. The platform architecture was designed to support future integration with real-time data streams and external AI modules.

3.4 Platform Significance

The developed platform provides:

- visualization of predicted facial motion,
- qualitative evaluation of animation quality,
- demonstration of project results,
- a foundation for future conversational avatar systems.

Together with the blendshape prediction models developed within AI-SPEAK, it establishes the technological basis for future research on digital humans and multimodal human-machine interaction.

4. Audio-Based Blendshape Prediction

The central research activity of this deliverable focused on predicting facial motion directly from speech signals. The task was formulated as a sequence-to-sequence regression problem in which a speech waveform serves as input and a sequence of 52 ARKit blendshape coefficients sampled at 60 fps represents the output.

Unlike traditional rule-based approaches that rely on predefined phoneme-to-viseme mappings, the proposed framework learns the relationship between speech and facial motion directly from data.

The developed system is based on:

- Serbian-adapted Wav2Vec 2.0 speech representations,
- phoneme-context information,
- Transformer-based contextual modeling,
- GRU-based temporal decoding,
- region-specific blendshape prediction heads.

The following sections describe the individual components of the proposed architecture and their contribution to speech-driven facial animation performance.

4.2 Wav2Vec 2.0 Based Architecture

Speech Representation Learning

The quality of speech representations is one of the most important factors affecting speech-driven facial animation. Traditional handcrafted features such as MFCCs and filterbank energies capture acoustic information but often fail to represent higher-level phonetic structure relevant for articulation. To overcome these limitations, the proposed system employs a Serbian-adapted Wav2Vec 2.0 model.

Wav2Vec 2.0 is a self-supervised speech representation learning framework that learns contextualized speech embeddings from large amounts of unlabeled audio. Instead of relying on manually designed acoustic features, Wav2Vec learns representations directly from speech data. The model used in this activity was previously fine-tuned for Serbian speech recognition tasks and kept frozen during blendshape model training. Hidden representations from the last Transformer layers are combined to obtain contextual speech embeddings used as input to the animation model.

Temporal Alignment

The Wav2Vec model produces features at approximately 50 Hz, while blendshape coefficients are available at 60 fps. To ensure synchronization between input and output streams, extracted speech representations are interpolated to 60 fps. The resulting feature sequence forms the primary input to the facial animation model.

Transformer Encoder

The first trainable component of the architecture is a Transformer encoder. The encoder consists of multiple self-attention layers responsible for contextual modeling. The primary role of the Transformer is contextual modeling.

Unlike recurrent networks that process sequences sequentially, the Transformer allows every frame to attend to all other frames within an utterance. This capability is particularly important for speech-driven animation because facial movements are strongly influenced by coarticulation.

Experimental results presented later in this deliverable demonstrate that the Transformer contributes the largest performance improvement among all investigated architectural components.

GRU Decoder

The contextualized Transformer representations are subsequently processed by a two-layer GRU decoder, which models local temporal dynamics and converts contextual speech representations into blendshape trajectories.

While the Transformer captures global context, the GRU focuses on local motion generation. This combination proved more effective than either component individually.

Region-Specific Output Heads

The final architecture predicts blendshape coefficients through five independent prediction heads corresponding to different facial regions: brows, eyes, upper face, jaw, mouth.

Each head consists of a small multilayer perceptron followed by a sigmoid activation function. This design enables specialization for individual facial regions and improves overall prediction accuracy. Because articulation-related coefficients dominate speech-driven facial motion, the jaw and mouth heads receive particular attention during training and loss design.

The overall architecture combines the strong contextual modeling capabilities of Transformers with the temporal smoothing properties of recurrent networks, providing an effective framework for speech-driven facial animation.

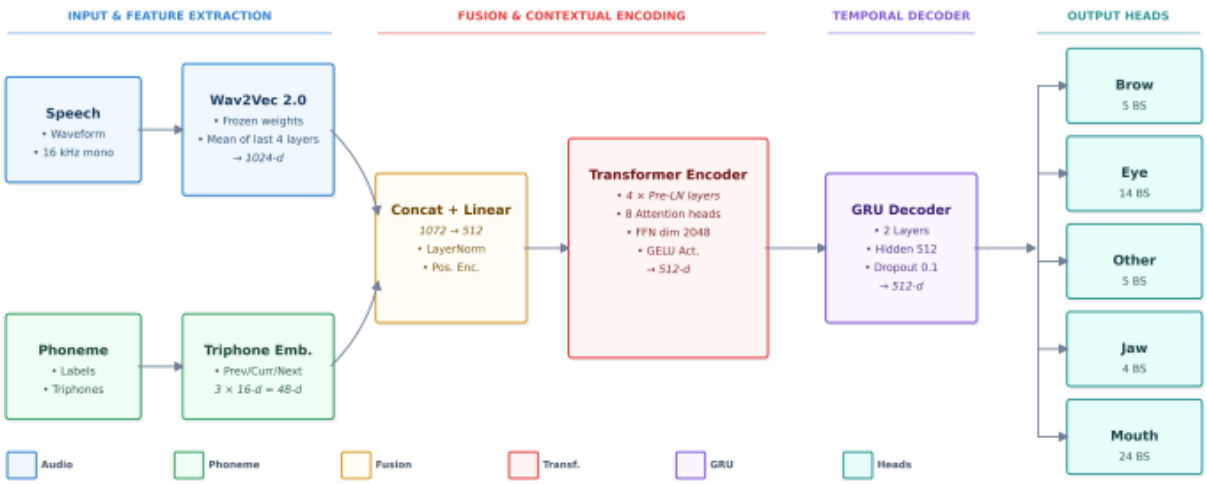


Figure 1 Overview of the propoced architecture.

5. Training Strategy

The quality of speech-driven facial animation depends not only on the network architecture but also on the design of the training objective. Standard regression losses often lead to over-smoothed motion and insufficient activation of important articulation-related facial regions. To address these limitations, several complementary loss functions were introduced.

5.1 Weighted Regression Loss

The primary optimization objective is based on a weighted Smooth L1 loss. Higher weights are assigned to articulation-related coefficients, particularly jaw, mouth, and tongue movements, as these contribute most strongly to perceived speech realism.

In addition, a dynamic frame-level weighting mechanism emphasizes active articulation frames, encouraging the model to better reproduce expressive speech events and articulation peaks.

5.2 Temporal Consistency Loss

To improve motion smoothness and reduce frame-to-frame jitter, a temporal consistency loss was incorporated. This objective encourages agreement between predicted and reference motion dynamics while preserving natural articulatory transitions.

5.3 Peak Preservation Loss

Preliminary experiments showed that important articulation events, such as strong jaw openings, lip closures, and vowel-related mouth shapes, were often underestimated by standard regression objectives.

To address this issue, a dedicated peak-preservation loss was introduced, encouraging the model to better reproduce articulation extremes and improving overall facial expressiveness.

5.4 Jaw-Specific Constraints

Because jaw motion plays a central role in speech articulation, additional jaw-oriented constraints were included during training. These constraints improve the timing and amplitude of jaw movements, resulting in more realistic speech-related facial motion.

The combination of weighted regression, temporal consistency, peak-preservation objectives, and jaw-specific constraints resulted in more accurate articulation, smoother motion trajectories, and improved perceptual quality of the generated animations.

6. Experimental Evaluation

6.1 Experimental Setup

The experimental evaluation was designed to assess both the predictive accuracy and practical usability of the proposed speech-driven facial animation framework. All experiments were conducted using the Serbian audio-blendshape corpus described in Section 2, with an 85/15 train-validation split.

Model performance was evaluated using objective and subjective criteria. Objective evaluation focused on Pearson correlation coefficient (Corr), mean absolute error (MAE), and real-time factor (RTF), while subjective evaluation was performed through a Mean Opinion Score (MOS) study involving human observers.

6.2 Audio-Only Speech-Driven Animation Performance

The proposed speech-driven facial animation framework combines Serbian-adapted Wav2Vec 2.0 representations, phoneme-context information, Transformer-based contextual modeling, and GRU-based temporal decoding. However, the results achieved exclusively on speech-derived information, i.e. audio-only input, are given in Table 1. The objective was to determine how accurately facial movements can be predicted directly from audio signals without relying on additional linguistic or visual modalities.

Table 1 Results of the system with audio-only input

Metric	Correlation	Mouth Correlation	Jaw Correlation	MAE	RTF
Value	0.843	0.940	0.890	0.025	0.025

These results demonstrate that high-quality facial animation can be generated directly from speech.

Particularly noteworthy is the mouth-region correlation of 0.940, indicating extremely strong agreement between predicted and reference articulation trajectories. Since the mouth region contains the most visually important speech-related movements, this result suggests that the developed framework successfully captures the majority of articulatory information present in the speech signal. The obtained jaw correlation of 0.890 further confirms the model's ability to reproduce major speech movements accurately. Perhaps most importantly, the system achieves an RTF of only 0.025, meaning that animation can be generated approximately fifty times faster than real time. Such computational efficiency is highly desirable for practical deployment scenarios involving interactive avatars.

6.3 Ablation Study

To better understand the contribution of individual components, several ablation experiments were conducted.

Transformer Encoder

Removing the Transformer encoder resulted in a noticeable degradation of performance, confirming the importance of contextual modeling for speech-driven facial animation. Models without Transformer layers exhibited weaker coarticulation modeling and less stable articulation trajectories.

Phoneme Context Information

Experiments performed without phoneme-context embeddings showed reduced prediction accuracy, indicating that explicit phonetic information provides useful complementary cues beyond acoustic representations alone. The importance of text as input will be further analyzed in D4.2.

Adversarial Training

A discriminator module was investigated to determine whether adversarial supervision could improve animation realism. However, the inclusion of adversarial training did not provide

measurable improvements and slightly reduced overall prediction accuracy. Consequently, discriminator-based training was not included in the final system.

Temporal Post-Processing

Savitzky-Golay filtering was evaluated as a post-processing technique for smoothing predicted trajectories. Although it reduced high-frequency fluctuations, objective evaluation showed no consistent improvement in performance. The final system therefore relies on motion dynamics learned directly by the neural architecture.

Overall, the ablation study confirmed that contextual modeling through the Transformer encoder and the inclusion of phoneme-context information provide the largest contribution to animation quality.

In Table2 all results from ablation study are summarized. The baseline model is the model that combines Serbian-adapted Wav2Vec 2.0 representations, phoneme-context information, Transformer-based contextual modeling, and GRU-based temporal decoding.

Table 2 Results of ablation study

Model	Correlation	MAE	RTF
Baseline	0.866	0.024	0.020
Baseline w/o phonemes (audio only)	0.843	0.025	0.025
Baseline w/o Transformer (GRU only)	0.810	0.029	0.019
Baseline with post-processing	0.858	0.024	0.029
Baseline with discriminator	0.844	0.026	0.022

6.8 Subjective Evaluation

While objective metrics provide important quantitative information, the perceptual quality of generated facial animation must ultimately be assessed by human observers. For this purpose, a Mean Opinion Score (MOS) study was conducted involving twenty-one native Serbian speakers.

Participants evaluated animations generated under three conditions:

1. Seen speakers,
2. Unseen speakers,
3. Synthetic speech generated using a text-to-speech (TTS) system.

The obtained MOS scores are summarized in Table 3.

Table 3 Results of subjective test

Condition	MOS
Seen Speakers	4.20
Unseen Speakers	3.95
TTS Speech	3.56

The highest score was obtained for speakers included in the training set. More importantly, the unseen-speaker condition achieved a MOS of 3.95, indicating good generalization to previously unseen speakers. As expected, animations generated from synthetic speech received lower ratings, suggesting that errors introduced by the TTS system propagate to the facial animation process.

Nevertheless, all evaluated conditions achieved scores well above the midpoint of the scale, confirming the overall perceptual quality of the proposed framework. Statistical analysis further confirmed significant differences between the evaluated conditions.

7. Conclusions

This deliverable presented the research activities conducted within the AI-SPEAK project related to speech-driven facial animation for Serbian avatars. The work addressed the complete development pipeline, including dataset creation, avatar platform development, neural architecture design, training methodology, and experimental evaluation.

A dedicated Serbian audio-blendshape corpus was created, containing synchronized speech recordings, ARKit-compatible facial motion parameters, and phoneme alignments. The dataset represents a valuable resource for future research in Serbian speech-driven facial animation.

An ARKit-compatible avatar platform was developed to support visualization and evaluation of generated facial animations. Together with the developed machine learning models, it provides a foundation for future AI-SPEAK activities involving conversational avatars and digital humans.

The proposed framework combines Serbian-adapted Wav2Vec 2.0 representations, phoneme-context information, Transformer-based contextual modeling, and GRU-based temporal decoding. Experimental results demonstrated that the proposed architecture can achieve high-quality facial animation using speech as the primary input modality. The conducted research further indicated that additional linguistic information derived from text can provide complementary cues that improve both objective performance measures and the perceived quality of generated animations. These findings motivate further investigation of multimodal approaches that combine speech and text inputs, which are addressed in the subsequent AI-

SPEAK activities described in Deliverable D4.2. Subjective evaluation further confirmed high perceptual quality and successful generalization to previously unseen speakers.

Overall, the conducted research demonstrates that realistic speech-driven facial animation is feasible for Serbian despite limited training resources. The developed corpus, avatar platform, and prediction framework represent important outcomes of the AI-SPEAK project and establish a foundation for future research on multimodal avatars, audiovisual speech technologies, and human-computer interaction