Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Deliverable index:	D4.3
Version:	2.0



Modules for synthesis of facial movements from text

Executive Summary

Building upon the speech-driven facial animation framework presented in Deliverable D4.2, this activity investigated how additional linguistic information can further improve avatar animation quality.

The first part of the research focused on extending the audio-based architecture with textual information. Experiments demonstrated that text provides complementary cues that are not fully captured by speech alone, leading to improvements in both objective performance measures and subjective animation quality. Several multimodal architectures combining speech and text representations were evaluated, including both real-time and offline approaches.

The second part of the research focused on visual speech representations. A dedicated study was conducted to establish a Serbian phoneme-to-viseme inventory using linguistic analysis, human perception experiments, and AV-HuBERT embedding analysis. The resulting viseme inventory is intended to support future research on visually grounded facial animation models.

Together, these activities provide a foundation for future AI-SPEAK developments involving multimodal avatar animation, visual speech modeling, and conversational digital humans.

1. Introduction

1.1 Motivation

Deliverable D4.2 demonstrated that realistic facial animation can be generated directly from speech using Wav2Vec-based speech representations and Transformer-GRU modeling. Although the obtained results were highly encouraging addition of phoneme-context information further improve results.

Text provides additional linguistic context, including phoneme identity, neighboring phonetic context, and sentence-level information. Such information may help resolve ambiguities present in the acoustic signal and improve modeling of coarticulation effects.

Furthermore, facial animation is ultimately a visual task. While phonemes are linguistic units, multiple phonemes often share very similar visual appearance. This motivates the investigation of visemes as visually grounded speech representations that may provide a more appropriate conditioning signal for facial animation systems.

1.2 Objectives

The objective of this activity was to investigate how linguistic information can enhance the speech-driven facial animation framework developed in Deliverable D4.2.

The main goals were:

- evaluation of textual conditioning for blendshape prediction,
- development of multimodal speech-text animation architectures,
- comparison of real-time and offline modeling approaches,
- analysis of phoneme and viseme representations,
- establishment of a Serbian phoneme-to-viseme inventory.

The following sections first examine the contribution of textual information to avatar animation and subsequently present the development of visually motivated speech representations for future AI-SPEAK activities.

2. Role of Textual Information in Speech-Driven Animation

2.1 Why Text Matters

Modern speech representations such as Wav2Vec 2.0 capture substantial phonetic information directly from speech. Nevertheless, some aspects of speech production remain difficult to infer exclusively from acoustic signals.

Speech articulation is inherently context-dependent. The facial realization of a phoneme is influenced not only by its own identity but also by neighboring sounds and longer linguistic context. As a result, acoustically similar speech segments may correspond to different facial configurations.

Text provides explicit information about:

- phoneme identity,

- phoneme order,
- word boundaries,
- sentence structure,
- contextual relationships between sounds.

Such information can complement acoustic representations and help resolve ambiguities that are difficult to infer from speech alone.

2.2 Coarticulation and Phonetic Context

One of the main challenges in speech-driven animation is coarticulation. Facial movements associated with a particular phoneme frequently begin before the phoneme becomes acoustically dominant and continue after this particular phoneme has ended. For example, lip rounding may start before a rounded vowel is produced, while jaw movements often depend on both preceding and following phonemes. Consequently, accurate animation requires information about local phonetic context rather than isolated speech sounds.

These observations motivate the incorporation of explicit linguistic information into speech-driven facial animation systems.

3. Extending the D4.2 Framework with Phonetic Context

3.1 Motivation

The speech-driven animation framework presented in Deliverable D4.2 relies primarily on speech representations extracted using a Serbian-adapted Wav2Vec 2.0 model. While these representations capture a substantial amount of phonetic information, additional linguistic context can further improve articulation modeling.

To investigate this hypothesis, explicit phonetic information was incorporated into the architecture developed in D4.2.

3.2 Triphone Embeddings

Phoneme alignments generated using WhisperX were used to construct contextual phonetic representations. Rather than considering only the current phoneme, local context was represented using triphones consisting of the current phoneme, the preceding phoneme and the following phoneme. Each triphone was mapped to a trainable embedding vector and combined with the acoustic representations extracted from Wav2Vec 2.0.

Triphone embeddings provide information about local coarticulation patterns and allow the model to distinguish between different realizations of the same phoneme occurring in different contexts.

3.3 Integration into the Speech-Driven Animation Pipeline

The triphone embeddings were incorporated into the Transformer-GRU architecture described in Deliverable D4.2. Acoustic and phonetic representations were jointly processed to predict sequences of ARKit blendshape coefficients. This extension required only minor modifications to the original architecture while providing the model with explicit information about articulation context.

Experimental results demonstrated that phonetic context provides useful complementary information beyond acoustic features alone. The inclusion of triphone embeddings improved blendshape prediction accuracy and produced more stable articulation trajectories, particularly in regions strongly affected by coarticulation. These findings indicate that linguistic information can improve speech-driven facial animation even when powerful self-supervised speech representations are available. The observed improvements motivated further research on richer text-based representations and multimodal speech-text architectures, which are presented in the following section.

Table 1 Objective measures on system with and without phoneme embedding

Model	Correlation	MAE	RTF
System with phoneme embeddings	0.866	0.024	0.020
System w/o phonemes (audio only)	0.843	0.025	0.025

4. Multimodal Speech-Text Architectures

4.1 Motivation

The results presented in the previous section demonstrated that explicit phonetic context improves speech-driven facial animation. However, triphone embeddings provide only local phonetic information and do not fully exploit the rich linguistic information available in text.

This observation motivated the development of multimodal architectures capable of jointly processing speech and textual information. The goal was twofold. First, textual representations were expected to provide complementary information that is difficult to recover from acoustics alone. Second, the study aimed to investigate temporal modeling architectures suitable for both high-quality offline animation and future low-latency deployment scenarios.

4.2 Speech-Text Fusion

The proposed framework combines acoustic and textual information to improve speech-driven facial animation. Unlike the D4.2 architecture, which incorporated linguistic information through triphone embeddings, the new framework utilizes complete textual sequences.

Acoustic information is represented using MFCC features, while the corresponding transcript is encoded at the character level using a dedicated bidirectional GRU text encoder. The resulting contextualized character representations capture both local phonetic structure and longer linguistic dependencies.

Fusion between speech and text is performed through multi-head attention. Acoustic features act as queries, while textual representations serve as keys and values. This mechanism allows the model to dynamically select linguistically relevant information for each animation frame. The attention output is subsequently projected into a shared latent space and forwarded to the temporal encoder.

In contrast to the triphone-based extension of the D4.2 framework, this approach enables the model to exploit complete textual context rather than only local phonetic neighborhoods. Although the temporal encoder may operate causally, the text branch relies on the complete transcript. Consequently, the investigated framework should be viewed as a multimodal low-latency architecture rather than a fully streaming speech-driven animation system.

The final prediction stage consists of two output heads. The primary regression head predicts the 52 ARKit blendshape coefficients, while an auxiliary classification head performs viseme prediction. The auxiliary task encourages the model to organize internal representations around visually meaningful speech categories and serves as an additional regularization mechanism during training.

4.3 Causal Temporal Convolutional Network

One objective of this activity was to investigate whether multimodal speech-text animation can be performed under low-latency constraints. For this purpose, a causal architecture based on Temporal Convolutional Networks (TCNs) was developed.

The TCN consists of twelve residual dilated convolutional blocks with exponentially increasing dilation factors of 1, 2, 4, 8, 16, and 32, repeated twice throughout the network. This design enables efficient modeling of long temporal dependencies while preserving causality.

Because each prediction depends only on past and current acoustic information, the TCN avoids the need for future speech context and provides a practical foundation for future interactive avatar applications. Residual connections improve optimization stability, while dilated

convolutions allow the receptive field to grow without a proportional increase in computational complexity.

4.4 Offline Bidirectional GRU

In parallel with the causal TCN, a fully offline architecture based on Bidirectional GRUs (BiGRUs) was investigated.

Unlike the TCN, the BiGRU has access to both past and future context, enabling more accurate modeling of coarticulation effects, phoneme transitions, and anticipatory articulation. The architecture consists of two bidirectional GRU layers followed by a residual feed-forward projection block and the dual prediction heads described previously.

By exploiting future information unavailable to causal models, the BiGRU serves as a high-quality offline reference architecture and consistently produces smoother facial trajectories and more accurate articulation peaks.

4.5 Training Objective

The multimodal architectures were trained using a composite objective designed to simultaneously optimize prediction accuracy, temporal smoothness, visual speech structure, and articulation expressiveness.

The primary component is a weighted Smooth L1 regression loss applied to the 52 blendshape coefficients. Larger weights are assigned to articulation-related coefficients such as jaw, mouth, and tongue movements, while an additional frame-level weighting mechanism emphasizes frames containing active articulation.

A temporal consistency loss penalizes excessive frame-to-frame fluctuations and reduces animation jitter. An auxiliary cross-entropy loss over viseme classes encourages the learned representations to capture visually meaningful speech structure. Finally, a dedicated peak-preservation loss improves reproduction of articulation extremes such as strong jaw openings, lip closures, and vowel articulations.

The final objective combines regression, temporal, auxiliary viseme, and peak-preservation losses, resulting in more stable and perceptually realistic facial animation.

4.6 Face Refiner and Ensemble Modeling

During experimental evaluation, it was observed that the primary prediction models occasionally underestimate the activation of certain blendshape coefficients, particularly in upper-face regions such as brows, blinks, and gaze-related movements.

To address this limitation, a dedicated Face Refiner module was introduced as a second-stage prediction model. Rather than operating directly on speech or text, the refiner receives the

blendshape trajectories predicted by the primary animation model and learns to estimate residual corrections.

The final Face Refiner is implemented as a Ridge-regression model trained on residual prediction errors. For each frame, the input consists of the predicted blendshape vector, its temporal derivative, and quadratic coefficient terms. This design allows the model to learn how prediction errors depend on both the current facial configuration and its local dynamics.

Hyperparameters are optimized on the validation set, while the final correction is interpolated with the original prediction to prevent overcorrection. Qualitative analysis showed that the refiner particularly improves eyebrow motion, blinking behavior, gaze-related coefficients, and other subtle upper-face dynamics that are often underrepresented in speech-driven animation systems.

The best-performing configuration combines predictions generated by both the TCN and BiGRU architectures through weighted averaging, followed by the Ridge-based Face Refiner. This ensemble approach achieved the lowest prediction errors.

Table 2 summarizes the influence of the refinement stage on objective performance.

Table 2 Comparison of architectures with and without Refiner

Model	MAE	Mouth-regio MAE	jaw-open MAE
TCN	0.0189	0.0230	0.0259
BiGRU	0.0195	0.0230	0.0250
TCN + Refiner	0.0188	/	/
BiGRU + Refiner	0.0193	/	/
Ensamble + Refiner	0.0181	0.0217	0.0240

4. Improvement Achieved Through Linguistic Conditioning

4.1 From Acoustic to Multimodal Modeling

The research activities described in Deliverables D4.2 and D4.2 investigated several levels of linguistic conditioning for speech-driven facial animation.

The initial framework relied primarily on acoustic information extracted from speech. Subsequently, explicit phonetic context was introduced through triphone embeddings, providing the model with information about neighboring phonemes and local coarticulation effects. Finally, a multimodal architecture incorporating complete textual information was developed, enabling the model to exploit broader linguistic context through attention-based speech-text fusion.

This progression allowed systematic analysis of how different forms of linguistic information contribute to facial animation quality.

4.2 Contribution of Phonetic Context

The first step beyond purely acoustic modeling involved the incorporation of phonetic context through triphone embeddings.

Experimental results demonstrated that explicit phonetic information provides useful complementary cues that improve blendshape prediction accuracy. Compared to the audio-only baseline, the inclusion of triphone embeddings resulted in a measurable improvement in trajectory correlation. In addition to numerical improvements, qualitative analysis revealed more stable articulation trajectories, improved anticipation of upcoming speech events, and more reliable modeling of lip closures and other speech-related facial movements.

These findings confirmed that local phonetic context provides information that is not fully captured by acoustic representations alone.

4.3 Contribution of Textual Information

Building upon the positive results obtained with triphone embeddings, the next stage of research investigated the use of complete textual information.

Unlike local phonetic context representations, textual conditioning provides access to broader linguistic information, including sentence-level structure and longer contextual dependencies. Experimental evaluation demonstrated that speech-text fusion gave prediction consistency and produces stable facial motion trajectories.

Furthermore, the attention-based fusion mechanism enables dynamic interaction between speech and text representations, allowing the model to utilize different parts of the textual sequence depending on the current speech segment.

4.4 Key Findings

Several important conclusions emerge from the conducted studies. First, speech signals alone contain sufficient information to generate realistic facial animation, as demonstrated by the results presented in Deliverable D4.2. Second, linguistic information consistently improves animation quality. Even relatively simple phonetic context representations provide measurable gains over purely acoustic models. Third, richer textual representations are beneficial for facial animation because of broader linguistic information and should not be viewed solely as an acoustic prediction problem.

Overall, the obtained results support a multimodal view of speech-driven facial animation in which acoustic and linguistic information provide complementary cues for facial motion

generation. These findings directly motivated the next stage of research focused on visually grounded speech representations, presented in the following section on Serbian phoneme-to-viseme mapping.

5. Serbian Phoneme-to-Viseme Mapping

5.1 Motivation

The results presented in the previous sections demonstrated that linguistic information significantly improves speech-driven facial animation. However, phonemes are primarily linguistic units, while facial animation is fundamentally a visual task.

Many acoustically distinct phonemes produce very similar facial appearances. For example, the phonemes /p/, /b/, and /m/ all involve bilabial closure and are often visually indistinguishable. Similar relationships exist among several other consonant groups.

These observations motivated the development of a Serbian phoneme-to-viseme inventory. The objective was to identify visually similar groups of phonemes that could serve as a more appropriate conditioning signal for future facial animation systems.

5.2 Methodology

The viseme inventory was developed by combining evidence from three complementary sources:

- linguistic analysis of Serbian phoneme articulation,
- human visual speech perception experiments,
- data-driven analysis using AV-HuBERT visual embeddings.

The linguistic analysis provided an initial grouping based on articulatory similarity. Human perception experiments were then conducted with nineteen native Serbian speakers who attempted to recognize isolated phonemes using visual information only. Finally, visual embeddings extracted from AV-HuBERT were used to investigate whether machine-learned visual speech representations exhibit similar clustering behavior. Together, these analyses provide complementary perspectives on visual speech organization.

5.3 Human Perception and AV-HuBERT Analysis

The perception experiment revealed several stable confusion patterns among Serbian phonemes. High confusion rates were observed for bilabial consonants, labiodental consonants, and several palatal groups, indicating that human observers frequently perceive groups of visually similar phonemes rather than individual phonemes.

A complementary analysis based on AV-HuBERT embeddings produced similar structures in the learned visual representation space. Hierarchical clustering revealed groups corresponding to

bilabial, labiodental, palatal, and vowel categories. Furthermore, a statistically significant correlation was observed between human perceptual similarities and AV-HuBERT embedding similarities, r = 0.543. This agreement suggests that both human observers and machine-learned audiovisual representations capture common aspects of visual speech production.

5.4 Unified Serbian Viseme Inventory

The final Serbian viseme inventory was obtained through consensus analysis combining linguistic knowledge, perceptual evidence, and AV-HuBERT representations. The resulting inventory contains eleven viseme classes. Vowels largely remain separate categories due to their distinctive mouth configurations, while several consonant groups are merged according to visual similarity. Final phoneme-viseme map is presented in Table 3.

Table 3 Final phoneme-viseme map

Viseme	Representative Phonemes
V0	/p/, /b/, /m/
V1	/f/, /v/
V2	/k/, /g/, /x/
V3	/i/
V4	/l/, /r/, /n/
V5	/t/, /d/, /s/, /z/, /t͡s/
V6	/t͡ɕ/, /d͡ʑ/, /ʂ/, /ʐ/, /t͡ʂ/, /d͡ʐ/
V7	/j/, /ʌ/, /ɲ/
V8	/e/
V9	/a/
V10	/o/, /u/

The final inventory exhibited strong agreement with the initial linguistic grouping, achieving an Adjusted Rand Index (ARI) of 0.78 and a Normalized Mutual Information (NMI) score of 0.94.

5.5 Relevance for Facial Animation

The developed viseme inventory provides a visually grounded alternative to phoneme-based conditioning. Compared to the full Serbian phoneme inventory, the proposed viseme representation is more compact and directly related to facial appearance. This may improve model generalization, reduce data sparsity, and provide a more intuitive connection between linguistic information and facial motion. The inventory therefore represents an important

resource for future AI-SPEAK research on multimodal avatar animation and visually informed speech-driven facial animation systems.

6. Planned Integration of Viseme Information

The speech-driven animation frameworks presented in Deliverables D4.2 and D4.2 currently rely on acoustic, phonetic, and textual representations. The Serbian viseme inventory developed within this activity provides a new visually grounded representation that may further improve facial animation quality.

Future work will investigate replacing or complementing phoneme-level conditioning with viseme-level representations. Because visemes are defined according to visual appearance rather than acoustic distinctions, they may provide a more compact and perceptually relevant conditioning signal for blendshape prediction.

Several potential benefits are expected:

- improved modeling of visually dominant articulatory events,
- reduced representation complexity,
- improved robustness in limited-data scenarios,
- better generalization across speakers,
- stronger correspondence between linguistic representations and facial appearance.

Future experiments will evaluate whether viseme-aware conditioning can improve blendshape prediction quality compared to existing phoneme-based approaches. Particular attention will be devoted to mouth shape prediction, jaw trajectory modeling, articulation peaks, and anticipation of upcoming speech events.

Beyond facial animation, the developed viseme inventory may also support future AI-SPEAK activities involving audiovisual speech recognition, lip reading, talking-face generation, and multimodal conversational avatars.

7. Conclusions

This deliverable presented research activities focused on enhancing speech-driven facial animation through the incorporation of linguistic information and visually motivated speech representations.

Building upon the audio-based framework presented in Deliverable D4.2, the conducted studies demonstrated that linguistic information provides complementary cues that improve facial

animation quality. Explicit phonetic context and text-based conditioning both contributed to more accurate articulation modeling and more stable blendshape trajectories.

To further investigate visually grounded representations, a Serbian phoneme-to-viseme inventory was developed using linguistic analysis, human perception experiments, and AV-HuBERT embedding analysis. The resulting inventory provides the first comprehensive mapping of Serbian visual speech categories and establishes a foundation for future viseme-aware animation systems.

The reported results support the view that speech-driven facial animation should be treated as a multimodal problem that combines acoustic, linguistic, and visual speech information. Together, the developed multimodal architectures and the Serbian viseme inventory represent important outcomes of the AI-SPEAK project and provide a basis for future research on conversational avatars, audiovisual speech technologies, and digital humans.