Project title:	"Multimodal multilingual human-machine speech communication"
Project Acronym:	AI-SPEAK
Deliverable index:	D1.2a
Version:	2.2

IMPLEMENTATION PLAN

of the Project "Multimodal multilingual human-machine speech communication" (AI-SPEAK), as initially established during the series of meetings that took place in January 2024 including the Project kick-off meeting, and finalized by the end of WP1 (Preparatory activities) in M4.

This Implementation plan results from a detailed analysis of the state of the art in the field of AI based speech technology and image processing, most notably:

- lip reading algorithms, which involve understanding lip movements and correlating them with spoken words; having in mind that deep learning methods, especially sequence-to-sequence models like Long Short-Term Memory (LSTM) networks and Transformer-based architectures, have shown promise in this area;
- methodologies for integrating audio and visual information including multimodal fusion techniques, such as early fusion (combining features at the input level) and late fusion (combining features at decision level);
- algorithms in the field of audio-driven facial animation i.e. facial animation generated directly from speech signals, involving mapping speech features to facial expressions and movements using deep learning approaches (Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs)).

1. Description of project activities

As described in Project proposal, the implementation is divided into the following 7 work packages:

- **WP1**: *Preparatory activities*
- **WP2**: *Data collection and processing*
- **WP3**: *ASR Implementation and experiments*
- **WP4:** *TTS Implementation and experiments*
- **WP5**: *Integration and evaluation*
- **WP6**: *Dissemination*
- **WP7**: *Management*.

The basic data related to the workpackages is listed in the following table:

WP No	WP title	WP Lead - team member's ID	Start month	End month	Duration
1	Preparatory activities	PI	1	4	4
2	Data collection and processing	TM3	5	18	14
3	ASR Implementation and experiments	TM2	5	30	26
4	TTS Implementation and experiments	TM5	19	30	12
5	Integration and evaluation	TM2	31	36	6
6	Dissemination	TM1	1	24	24
7	Management	PI	1	24	24

The analysis of the state of the art which was conducted in WP1, has resulted in detailed implementation, quality control and dissemination plans. The aim of WP2 is to collect two multimodal and multilingual corpora, necessary in order to implement all of the activities described in WP3 and WP4, including the development of modules for speech recognition from video, modules for speech recognition from audio and video, modules for speech re-synthesis from video, adaptation of the 3-D face model to pronounce phonemes of the Serbian language, development of modules for synthesis of facial movements from speech and synthesis of facial movements from text. Dissemination activities (WP6) will be conducted according to the Project budget, as described in the **Dissemination plan** (deliverable D1.2c). Project management (WP7) will be established in order to facilitate the coordination of research and monitoring of its results, and to ensure quality control during the Project, according to the **Quality control plan** (deliverable D1.2b).

Work package description

The following is a description of work packages as described in the Project proposal. Where relevant, a more detailed description has been added, resulting from the analysis of the state-of-the-art performed by team members in the first trimester of the Project. This principally concerns WP2, WP3, and WP4, where specific research directions and implementation details have been established.

Work package number	1	Work package title	Preparatory activities
Lead SRO's acronym	FTNUNS		
WP Coordinator - team member's ID	PI		
Team members' IDs	PI, TM1-TM8		
Objectives To organize a kick-off meeting, to implement the Project website, to provide a detailed analysis of the state of the art and produce detailed implementation, quality control and dissemination plans.			
Description of work (where appropriate, broken down into sub-activities), and role of the team members			

Kick-off meeting (Sub-activity 1.1; month 1) will be organized to discuss the Project implementation and to establish procedures. Implementation of Project website (Sub-activity 1.2; month 1) will allow public access to Project related information (e.g. publications and datasets) and thus increase Project visibility and facilitate Project management. Team members will analyse the state of the art (Sub-activity 1.3; months 1-4) and produce detailed implementation, quality control and dissemination plans (Sub-activity 1.4; months 3-4). The WP will be led by Milan Sečujski (PI).
Deliverables of the work package (brief description and month of delivery) D1.1. Project website (1) D1.2. Detailed implementation, quality control and dissemination plans (4)

Work package number	2	Work package title	Data collection and processing
Lead SRO's acronym	FTNUNS		
WP Coordinator - team member's ID	TM3		
Team members' IDs	PI, TM1-TM8		
Objectives To collect two audio-visual speech corpora, to establish the framework for their processing and prepare them for the implementation of WP3, WP4 and WP5.			
Description of work (where appropriate, broken down into sub-activities), and role of the team members Production of the AI-SPEAK multimodal speech corpus (Sub-activity 2.1; PI, TM5, TM7; months 5-12), development of tools for its processing and its preparation for a successful implementation of following WPs (Sub-activity 2.2; PI, TM5, TM7; months 9-18); collection and processing of the Internet speech corpus (Sub-activities 2.3 and 2.4; PI; months 5-18). Both multimodal speech corpora will be developed with support of a sub-contractor charged with data annotation. The WP will be led by Nikša Jakovljević (TM3), bearing in mind his references in the development of multimodal corpora and AI-based processing tools. The first of the two corpora, AI-SPEAK speech corpus, will contain recordings of speech in both Serbian and English from 25 adult speakers of both genders, together with video recordings of the movements of their lips. The intended quantity of speech data per speaker is 10 minutes, although it is possible that eventually more data will be obtained. The corpus will be recorded in strictly controlled conditions, in the IAC Mini anechoic chamber obtained through the Erasmus+ project SENVIBE. The script for the corpus will contain recordings of a fixed number of utterances in both Serbian and English, including spoken digits, names of letters, simple commands ("up", "down", "back"...) as well as a number of short sentences. The second corpus, termed Internet speech corpus, will be based on existing videos published on the Internet, and will include a far greater quantity of data, but with little or no control over textual content or recording conditions, and thus, with far greater diversity as regards speaker characteristics, sound quality, acoustic ambience, recording equipment, lighting or back-ground, and it will be used for unsupervised training in uncontrolled settings. Clearly, the focus will be on Serbian audio-video recordings, having in mind that the quantity of multimodal data available online for English is already quite large.			
Deliverables of the work package (brief description and month of delivery) D2.1. AI-SPEAK speech corpus (18) D2.2. Internet speech corpus (18)			

Work package number	3	Work package title	ASR implementation and experiments
Lead SRO’s acronym	FTNUNS		
WP Coordinator - team member’s ID	TM2		
Team members’ IDs	PI, TM1-TM8		

Objectives

To implement novel machine learning algorithms for speech recognition from video and from audio+video, as well as algorithms for re-synthesis of speech from video.

Description of work (where appropriate, broken down into sub-activities), and role of team members

Novel machine learning algorithms for speech recognition from video (lip-reading) will be implemented (Sub-activity 3.1, TM2, TM3, TM4, TM6; months 5-24) as well as for audio-visual (multimodal) speech recognition i.e. recognition of speech from audio and video (Sub-activity 3.2, TM2, TM3, TM4, TM6; months 13-30) as well as video-to-speech synthesis, i.e. re-synthesis of speech from video only (Sub-activity 3.3; PI, TM2-TM7; months 19-30). The WP will be led by Branislav Popović (TM2), head of ASR team.

Addendum to the original work package description: The Project aims to incorporate visual cues from facial movements through audio-visual speech recognition and lip-reading. Existing ASR models achieve exceptional accuracy in controlled settings. However, their performance deteriorates in noisy environments or with damaged speech signals. This limitation impedes the applicability of speech technology in different scenarios, including intelligent conversational agents and voice assistants. Visual speech recognition, or lip-reading, utilizes facial movements for speech content recognition without audio input. Integrating visual cues from facial movements enhances ASR systems, fostering robust, multilingual, multimodal, and flexible speech recognition. Multimodal speech recognition, blending audio and visual cues, notably enhances accuracy, particularly in noisy environments. Lip-reading has shown significant progress with deep learning, surpassing traditional methods. Deep learning-based automated lip-reading systems, particularly focusing on CNNs, have shown promising results both in terms of feature extraction and classification. Multiple surveys highlight the shift toward deep learning architectures and the importance of temporal context modelling. By leveraging deep learning and unsupervised learning methods, including 3D convolutional mechanisms, the Project seeks to improve feature extraction and phoneme representation. On the other hand, multimodal speech synthesis, particularly video-to-speech synthesis (VTS), has advanced due to self-supervised learning and the availability of large audio-visual datasets. VTS presumes direct conversion of video-only recordings into audio based on soft speech recognition, i.e. without relying on strict phonetic transcriptions. The Project aims to develop systems for speech re-synthesis from visual information, primarily for the Serbian language. Despite advancements, existing systems face challenges such as low accuracy, limited vocabulary, and synchronization issues between audio and video. The Project will conduct objective and subjective evaluations using standard metrics such as Word Error Rate (WER) and Phoneme Error Rate (PER). Furthermore, the Project will enhance existing applications, such as real-time medical transcription, by integrating lip movement detection.

Deliverables of the work package (brief description and month of delivery)

D3.1. Modules for speech recognition from video (24)

D3.2. Modules for speech recognition from audio and video (30)

D3.3. Modules for speech re-synthesis from video (30)

Work package number	4	Work package title	TTS implementation and experiments
Lead SRO’s acronym	FTNUNS		
WP Coordinator - team member’s ID	TM5		
Team members’ IDs	PI, TM1-TM8		

Objectives

To implement novel machine learning algorithms for synthesis of facial movements from speech and from text.

Description of work (where appropriate, broken down into sub-activities), and role of the team members

After the adaptation of the selected and procured avatar (3D face model) to support the phonemic inventory of the Serbian language (Sub-activity 4.1; PI, TM2, TM5-TM7; months 19-21), novel machine learning algorithms for synthesis of facial movements from speech (speech-to-lip) will be implemented (Sub-activity 4.2; PI, TM2, TM5-TM7; months 22-30) as well as for their synthesis from text (text-to-lip) (Sub-activity 4.3; PI, TM5, TM7; months 25-30). The WP will be led by Siniša Suzić (TM5), head of TTS team.

Addendum to the original work package description: Talking face generation, also known as speech-to-lip generation, reconstructs facial motions concerning lips given speech input. There are two principal approaches to visual speech synthesis: two-stage frameworks, in which the source is first converted to facial parameters and they are then converted to video, and one-stage frameworks, where the source is directly converted to video. State-of-the art approaches in speech-to-lip such as Wav2Lip achieve very good results regarding lip synchronization but lack the ability to control emotional facial expressions. We propose to extend the research on synthesis of lip movement from audio, i.e. speech-to-lip, by taking into account emotional content of input speech and translating it to the output video. This research direction will rely on existing audio-visual emotional speech corpora, such as MEAD. It will be possible to pursue this direction for Serbian as well, owing to the existence of the SEAC corpus of emotional speech data, collected and published within our latest project with the Science Fund (AI S-ADAPT)

The talking face will also be generated directly from text. This is mostly achieved by combining speech-to-lip systems with text to speech systems. Although we already did some work towards this direction in Serbian, we plan to extend the proposed approach with cross-linguality. Tacotron based models can be easily adapted to different languages since only audio with transcription is needed to finetune the model. However, in order to achieve intelligible and natural sounding voice approximately 20h of consistent-quality speech from single speaker is needed. We plan to use our previous cross-lingual TTS to produce synthetic Tacotron training data by exploiting phonetically and prosodically annotated data from different speakers.

Deliverables of the work package (brief description and month of delivery)

D4.1. Avatar adapted to Serbian (21)

D4.2. Modules for synthesis of facial movements from speech (30)

D4.3. Modules for synthesis of facial movements from text (30)

Work package number	5	Work package title	Integration and evaluation
Lead SRO’s acronym	FTNUNS		
WP Coordinator - team member’s ID	TM2		
Team members’ IDs	PI, TM1-TM8		

Objectives

To integrate developed algorithms into existing ASR/TTS systems and provide objective and subjective evaluation of their performance (or performance improvement, where applicable)

Description of work (where appropriate, broken down into sub-activities), and role of the team members

Novel algorithms will be integrated into a digital voice assistant for Serbian as well as system for real-time transcription of dictated medical findings (Sub-activity 5.1, TM2, TM5 and TM7; month 31-35), and extensive objective and subjective evaluation will be carried out (Sub-activity 5.2, PI, TM1-TM8; month 34-36), according to performance measures described in *1.1 Objectives*. TM5 and TM7 will develop user interfaces for subjective evaluation. The WP will be led by Branislav Popović (TM2), bearing in mind his long-standing experience in system integration.

The project aims to advance human-machine speech communication by integrating state-of-the-art deep-learning-based algorithms that incorporate visual cues from facial movements. As to speech recognition, the project aims to exploit visual cues to enhance existing ASR systems, resulting in robust multilingual, multimodal, and multiscale speech recognition capabilities. Concurrently, models will be developed to generate visual cues for avatars with natural lip movements, enhancing the expressiveness and intelligibility of synthesized speech. These avatars will support the full phonetic inventory of both Serbian and English languages. We will also implement the so-called soft recognition, using the direct end-to-end conversion of video-only recordings into audio, without any rigid phonetic transcriptions (speech re-synthesis from visual cues). Evaluation of system effectiveness will be conducted through a combination of standard objective measures and subjective evaluations, including real-world scenario testing. The methods will be integrated into existing speech technology applications for Serbian, namely, a digital voice assistant and a real-time transcription system for dictating medical findings. However, the applicability of these methods will extend to other speech technology products, including speech recognition for intensive care patients, automatic video captioning, and dictation aids for legal or medical documents. To evaluate the quality of the ASR system, Word Error Rate (WER) and Phoneme Error Rate (PER) will be used as standard objective measures (Serbian is a highly inflective language, which is why PER may represent a more realistic measure). The evaluation of TTS and speech re-synthesis (i.e., soft recognition) will be accomplished by using standard subjective quality assessment methods, e.g., Mean Opinion Score (MOS) or MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor). As to the evaluation of speech recognition from video only, objective evaluation measures will include Perceptual Evaluation of Speech Quality (PESQ), Spectro-Temporal Modulation Index (STMI), Short-Time Objective Intelligibility (STOI), Extended STOI (ESTOI) and word error rate (WER). Finally, the intelligibility and naturalness of re-synthesized speech, as well as the naturalness of accompanying avatars, will be evaluated through listening tests, using the Mean Opinion Score (MOS). For subjective evaluations, a specialized user interface will be defined and developed. The final incorporation into existing applications will provide us with the possibility to evaluate the effectiveness of the proposed methods in a real-world scenario.

Deliverables of the work package (brief description and month of delivery)

D5.1. Improved ASR/TTS systems for Serbian (35) D5.2. Objective and subjective evaluation reports (36)

Work package number	6	Work package title	Dissemination
Lead SRO's acronym	FTNUNS		
WP Coordinator - team member's ID	TM1		
Team members' IDs	PI, TM1-TM8		
Objectives To increase the visibility of the Project and provide open access to its results; to strengthen international scientific collaboration, ensuring impact beyond the Project timeframe.			
Description of work (where appropriate, broken down into sub-activities), and role of the team members Most dissemination activities, with the exception of the Project website creation (D1.1), will be based on a detailed dissemination plan (D1.2). Besides the organization of major dissemination events (Sub-activity 6.1; PI and TM1) – the 25th edition of the international conference SPECOM in the first year (months 1-7) and a workshop for regional IT companies (months 24-25) – dissemination activities will also include the preparation of scientific publications for open access journals, participation at scientific conferences, appearances in electronic and printed media and on social networks, as well as preparation of appropriate marketing material, including brochures, leaflets and roll-up banners (Sub-activity 6.2). Work package will be led by Prof. Vlado Delić (TM1).			
Deliverables of the work package (brief description and month of delivery) * *Papers published in journals and conference proceedings during the Project (the open datasets that will result from the Project will also contribute to Project visibility, but they are listed as deliverables of WP2).*			

Work package number	7	Work package title	Management
Lead SRO's acronym	FTNUNS		
WP Coordinator - team member's ID	PI		
Team members' IDs	PI, TM1, TM2		
Objectives To coordinate project activities and organize meetings, monitor progress, supervise communication and dissemination, establish procedures and quality control, produce administrative/financial reports and contingency plans.			
Description of work (where appropriate, broken down into sub-activities), and role of the team members			

PI will overview overall project activities of the 3 teams (ASR, TTS, and image processing team, led by TM2, TM5 and TM4 respectively), delivery of results, quality control, contingency plans and dissemination. He will particularly overview the communication between teams and leaders of sub-activities or specific tasks within them, who will report to the PI on a weekly basis. Regular project meetings will be organized every six months. The private section of the Project website will support communication between team members as well as clear planning of activities.
Deliverables of the work package (brief description and month of delivery) D7.1.-D7.12. Quarterly project reports (3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36) D7.13. Final project report (36)

2. Equipment

The computational resources needed for the implementation of the Project will be provided partly by FTS-UNS, with the access to cloud-based computational resources provided by “AlfaNum” during the project timeframe. The resources provided by FTS-UNS and “AlfaNum” will be complemented by computational equipment acquired within this project as well as to computational resources at the National Data Centre. The total of abovementioned computational resources is expected to suffice for the implementation of all Project activities, most notably the training of machine learning systems, which can, as it is well known, be extremely computationally expensive. The amount of around 4,000 EUR defined in the Project budget will be used for acquisition of 3 GPU laptops to accelerate training experiments conducted by WP leaders, as regards ASR and TTS.

3. Risk management

The risks initially identified at the time of the submission of Project proposal are briefly listed in the following table. For each risk category, potential risks are listed and planned risk mitigation measures are indicated. The table is followed by the list of additional risks that were identified or have actually ocurred in the meantime, and the measures for their mitigation.

Risk assessment	Description of the risk	Risk mitigation measure to be undertaken by members of the Project team or SRO	Risk level
Methodology risk	Description of the risk	There is a possibility that some of the implemented methods are not as efficient/accurate as expected.	medium
	Actions to be undertaken	The scientific backgrounds of project team members are sound and sufficiently diverse, and most of them have a lot of experience in similar projects, which is why it can be expected that they will be able to find creative solutions.	
Work packages, deliverables and milestones	Actions to be undertaken	Delays in data collection and processing can postpone subsequent activities.	low
	Actions to be undertaken	Implementation phases (Work Packages 3 and 4) span over a very long period. However, some research can be carried out on public speech corpora or either of the two corpora provided by the EC1, so it can start even earlier.	
Members of the project team and SROs	Description of the risk	The risk of some researchers leaving for another post or some other reasons is always present.	medium
	Actions to be undertaken	We will assign at least three researchers to each task and insist on research and implementation being well documented, so that there is always a team member who will be able to take over a task from someone who leaves.	
Procurement	Description of the risk	There is a risk that the public procurement of equipment can introduce delays	low
	Actions to be undertaken	We expect practically negligible consequences, since most project activities rely on equipment already at our disposal (e.g. anechoic chamber). Delays will be mitigated by temporary use of existing storage at FTNUNS until new storage is procured, temporary use of a slightly inferior privately owned Hero 8 camera and microphone etc.	
Budgetary issues	Description of the risk	Delays in the payments from the responsible authority or institution.	low
	Actions to be undertaken	Such delays can cause delays in the remuneration of researchers and (less likely) modifications of the dissemination plan, but no delay in research activities.	
Other risks	Description of the risk	Some project team members may become overburdened at some points during the Project.	medium
	Actions to be undertaken	None of the team members are engaged more than 30%, a limit in accordance with their engagement on other projects and curricular activities, but owing to the research being well documented, tasks may be re-assigned to others if needed.	