Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu

D3.8 Final Release of Speech and Natural Language Processing Tools

Work Package	3
Responsible Partner	FhG
Author(s)	Tugtekin Turan, Salima Mdhaffar
Contributors	Raul Monteiro, Afonso Mendes
Reviewer	Yannick Estève
Version	1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	19/02/2024	Initial Table of Contents (ToC)
0.2	19/03/2024	First Draft
0.3	25/03/2024	Input from Technical Partners
0.4	27/03/2024	Internal Review
1.0	28/03/2024	Ready for Submission

Executive Summary

The SELMA's final release is a significant accomplishment in Natural Language Processing (NLP), focusing on speech-related tasks. This version includes advanced neural architectures designed to minimize errors and improve performance.

SELMA's strength lies in its ability to automate the processing of large amounts of data in various languages. This release significantly improves the accuracy of the outputs. A notable feature is the incorporation of user-feedback updating systems where these systems continuously refine and enhance the performance of speech tasks and setting a benchmark for innovation in language technology tools.

The final release introduces enhancements to improve user experience and efficiency. The user interface has been comprehensively redesigned to create a more intuitive layout that simplifies navigation and utilization of the software's diverse features. Extensive documentation and support resources are now provided, equipping users with the tools to use the software and effectively comprehend its functionalities.

The final release underscores our commitment to open-source development. This encourages continuous improvement and adaptation to meet the dynamic needs of NLP field. The SELMA software is accessible and enables researchers and developers to contribute to continuous development.

In conclusion, this release showcases the result of our work in creating a powerful and effective tool for speech and natural language processing.

Table of Contents

Table of Figures

1. Introduction

The final report offers a comprehensive analysis of SELMA models and software, spanning various topics, including:

- Automatic Speech Recognition (ASR): The ASR module is a crucial component of our software that accurately transcribes spoken language into written text. With its expanded range of supported languages, it has become even more valuable in extracting important information from speech content.
- Speech-to-Speech Translation (S2ST): The S2ST module is responsible for translating spoken language from one language to another. It now supports a broader spectrum of languages and offers superior translation quality, achieved through multilingual self-supervised pre-training with unlabeled speech data. This module is important in breaking down language barriers and making content accessible to a global audience.
- Named Entity Recognition from Speech (NER-S): The NER-S module identifies and classifies named entities in spoken language into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. It has been refined to recognize an extended range of named entities from speech, offering invaluable insights and context.
- Automatic Post Editing (APE): The APE module is designed to correct errors automatically. It uses advanced machine learning algorithms to learn from human corrections and applies these learnings to new translations or transcriptions. This results in higher-quality output with user feedback and reduces the need for manual post-editing.
- Text-to-Speech Synthesis (TTS): The TTS module converts written text into spoken words. It has been updated to generate more natural and emotionally resonant speech. This module is particularly useful for creating audio content from text-based sources and enhancing accessibility.
- Speaker Clustering and Identification (SCI): The SCI module groups speech segments based on the speaker's identity. It can identify when the same speaker is speaking and separate different speakers in a news bulletin. This module is crucial for understanding dialogues and conversations in multi-speaker environments.

- Punctuation and Capitalization Recovery (PCR): The PCR module adds punctuation and capitalization to transcribed text, making it easier to read and understand. It has been enhanced to recover punctuation and capitalization in transcribed text across multiple languages, leveraging the power of ASR or S2ST models.

These enhancements yield invaluable tools for media monitoring staff and journalists by improving performance, rendering the software more stable and efficient. All components are deployed as containers and are readily available at our Docker hub, or are part of source code released in open source project.

2. Released Components

2.1 Automatic Speech Recognition (ASR)

End-to-end automatic speech recognition (ASR) models have been built in the framework of the SELMA project. Some members of the LIA partner are strongly involved in the development of the SpeechBrain project (https://speechbrain.github.io), and LIA uses this toolkit to develop its new ASR model for the SELMA framework. These ASR models are mainly built on pre-trained wav2vec 2.0 models.

For now, these programs are research tools; therefore, integration work is still necessary to make most of them accessible to a non-specialist public. Some of this software has been released on the SELMA GitHub: https://github.com/SELMA-project/LIA_speech/tree/main/asr. The ASR models were built in the framework of SELMA 2021 until March 2022 and are available in this repository, targeting the following languages: (1) Brazilian Portuguese, (2) French, (3) Modern Standard Arabic, and (4) Tunisian Dialect.

During the last months, the LIA partners developed a federated learning (FL) ASR system, which is also based on using the wav2vec2.0 model. FL is a distributed machine learning paradigm that aims to train a machine learning model without data sharing collaboratively. It consists of a network of multiple clients and one server. SpeechBrain and Flower toolkits have been used. Flower is an open-source framework that allows us to build FL experiments and considers the highly varied FL facility scenarios. This system has been released on the official GitHub of Flower: https://github.com/adap/flower/tree/main/baselines/fedwav2vec2.

2.2 Speech-to-Speech Translation (S2ST)

There are two speech translation tasks addressed by the SELMA project: speech-to-text translation and speech-to-speech translation. For speech-to-text translation, we focused on assessing the capability of speech translation models in extremely low-resource settings. With this goal, we have used the Tamasheq language as a use case. While this language is not part of the collection of languages initially targeted by the SELMA project, it allows us to assess the state-of-the-art performance in similar settings to many low-resource languages targeted by our project. During the SELMA project, LIA participated in two Speech Translation Challenges,

IWSLT 2022 and IWSLT 2023, translating Tamasheq to French and Tunisian Arabic Dialect to English, respectively. Additionally, for IWSLT 2023, LIA organized a new translation task from Pashto to French.

A recipe (including data preparation, training, and evaluation scripts) for the Tamasheq to French translation has been built and integrated into the SpeechBrain toolkit: https://github.com/speechbrain/speechbrain/tree/develop/recipes/IWSLT22_lowresource/AST/transformer.

LIA continues to work in Speech Translation and starts to participate in the IWSLT 2024 for three tasks Tamasheq to French, Tunisian Arabic Dialect to English and Levantin Arabic Dialects to English. We will release the systems trained during IWSLT 2024 later.

For the second task, speech-to-speech translation, we implemented an end-to-end approach and open-source released it within the SpeechBrain project as a recipe: https://github.com/speechbrain/speechbrain/tree/develop/recipes/CVSS/S2ST.

2.3 Named Entity Recognition from Speech (NER-S)

As for the previous tasks, the SpeechBrain toolkit was used to build a system for the MEDIA French corpus. This corpus is a dataset of phone audio recordings with manual annotations dedicated to semantic concepts extraction (SCE) from the speech in the context of human/machine dialogues. The corpus contains manual transcriptions and semantic annotations of dialogues from 250 speakers and totals less than 25 hours of speech. The semantic concepts extraction task is close to the named entity recognition from speech (NER-S) task, both slot filling tasks. The main difference comes from the semantic annotation, which is more generic for the NER-S task and more specific for the SCE task (a named entity is defined as a snippet of the global information contained in a document, while a semantic concept is defined for a specific task).

A recipe (including data preparation, training, and evaluation scripts) for the MEDIA corpus (ASR and SLU tasks) has been built and integrated into the SpeechBrain toolkit: https://github.com/speechbrain/speechbrain/tree/develop/recipes/MEDIA.

LIA has created a new dataset during the SELMA project for SLU for the Tunisian dialect. This corpus is a dataset of train reservation audio recordings with manual annotations dedicated to semantic concepts extraction from the speech in the context of human/human dialogues. The corpus contains manual transcriptions and semantic annotations of dialogues from 108 speakers and totals less than 8 hours of speech. A recipe (including data preparation, training, and evaluation scripts) for the TARIC-SLU corpus (ASR and SLU tasks) has been built and tested, which will be later integrated into the SpeechBrain toolkit. Its integration is not yet finalized due to a submission with anonymous information of the authors. The paper has been accepted, and we are working to put the recipe online.

2.4 Automatic Post Editing (APE)

Automatic Post Editing (APE) systems are designed to refine raw output from automated processes like transcription or translation. Therefore, the inaccurate output is subjected to APE, where errors can be corrected with a post-processing scheme. This iterative process ensures that the system evolves and becomes more reliable by incorporating user feedback and corrections into its knowledge base, enabling more precise future transcriptions.

SELMA's final tool, called M-PHANTOM, is designed to enhance the accuracy and relevance of transcription outputs. This system intelligently employs user feedback to construct a growing database of named entities and domain-specific terminology. The primary objective is to continually refine and improve the transcription process, reducing the frequency of errors in future outputs. Our Gradio-based application is the front end of M-PHANTOM, showcasing its workflow, which demonstrates the following capabilities and features:

- *Transcription Generation:* The Gradio application first converts audio input into a transcript over the SELMA's speech-to-text models. Users can upload or record audio directly on the “Transcribe” tab. Acceptable formats are “.wav” or “.mp3”. After selecting the audio's language, users initiate transcription with the press of a button, receiving text output promptly.
- *User Feedback Incorporation:* Users can contribute to the system's learning by correcting any inaccuracies in the transcript, which M-PHANTOM then refines its future performance. Submitting these corrections through the “Submit Feedback” button for system learning.

- *Keyword Visualization and Management:* The system also allows users to manage the evolving list of keywords, ensuring transparency and control over the correction process. The “Keyword Manager” tab is a dashboard where users review, audit, and manage the database's contents. It allows for the listening of audio associated with keywords and the removal of any inaccuracies.
- *Database Optimization:* Through user interaction and feedback, new and relevant entities are continually incorporated into the knowledge database, improving M-PHANTOM's lexicon and contextual understanding. The system synthesizes and incorporates the new entry into the database upon providing the text, choosing a locale, and selecting a synthetic voice.

The following figure shows the interface for the initial phase of audio transcription. Users can easily upload an audio file by dragging and dropping it into the marked area or clicking to navigate to their file. After hitting the "Transcribe" button will process the audio and display the transcribed text in the adjacent text box, ready for review and feedback.

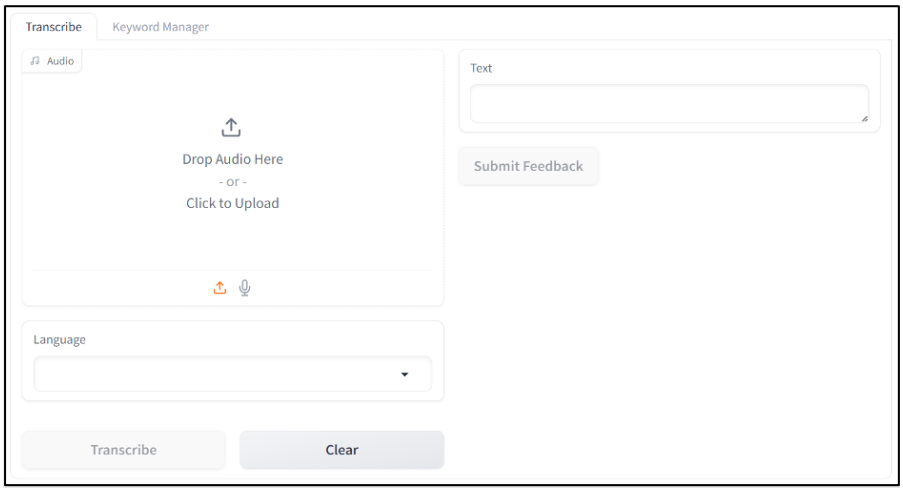

Figure 1 *M-PHANTOM Transcription Interface*

Keyword management features are also presented below where users engage with the keyword database actively. Here, one can review the existing keywords, listen to the associated audio clips, and remove any incorrect keywords.

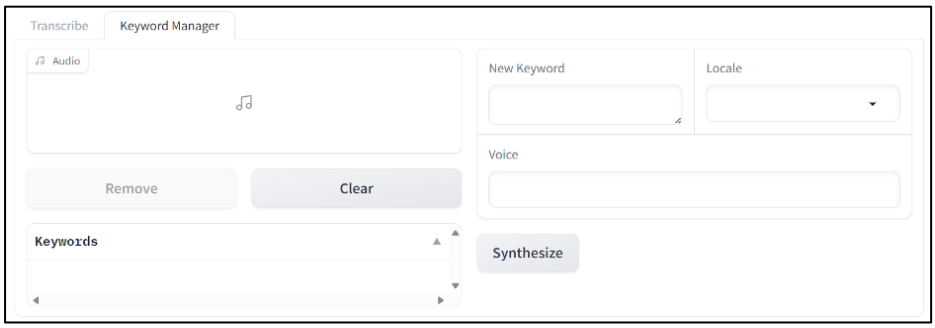

Figure 2 *M-PHANTOM Keyword Manager Interface*

Incorporating the principles from the keyword-spotting, M-PHANTOM enhances the capabilities of end-to-end automatic speech recognition outputs. Our final APE tool improves the recognition of rare named entities and adapts to user input by learning from interactions. The fusion of these methodologies ensures that both rare and common entities are captured with higher precision for more intelligent and responsive speech recognition systems.

2.5 Text-to-Speech Synthesis (TTS)

As soon as the first semester of the SELMA project, we released the first version of our TTS engine. It was a two-part system composed of an acoustic model and a vocoder. The acoustic model generates acoustic features from linguistic features (text in our case), and the vocoder synthesizes waveform from the acoustic features. For the acoustic model, we used Tacotron 2 with WaveRNN vocoder.

Then, we mainly worked on our baseline system regarding robustness and inference time. We considered moving from our two-part system to an end-to-end model to do this. This has the advantage of reducing error propagation due to the cascading system. On the other hand, using an end-to-end model gives us a faster inference time, which is very important since the model is deployed in production.

We found that variational autoencoder-based topology matches perfectly with our requirements. We have conducted several experiments that have shown that we can replicate the performance of Tacontron 2 + WaveRNN while decreasing the inference time by at least 150 times. The TTS API is accessible through the Plain X platform, and the docker image can be downloaded from our https://hub.docker.com/layers/selmaproject/selma-tts-avignon/pt_br-v2/.

To train the speech synthesis engine, we use the audio news bulletins that are produced by DW's Brazil department. The audio files have been downloaded from YouTube, and the scripts were retrieved from GitHub in a repository with all the text scripts that DW uses to produce its weekday news podcasts. The dataset contains approximately 32 hours of speech from 8 speakers.

2.6 Speaker Clustering and Identification (SCI)

Speaker Clustering and Identification (SCI) is a transformative feature within the field of media processing, offering a sophisticated approach to discern and distinguish between multiple speakers in an audio stream. SCI is integral for tasks such as transcribing interviews or news bulletins where multiple people speak. SELMA's final SCI tool performs concurrent speaker recognition, diarization, and speech-to-text conversion, vastly improving the clarity and utility of transcribed multi-speaker content.

SCI implements speaker identification with its multi-purpose capabilities. At its core, the interface performs simultaneous speaker recognition, diarization, and converting speech to text. This is achieved through algorithms that analyze audio streams, segregating them into discrete, speaker-specific tracks. The integration with SELMA's Whisper models further enhances the system's transcription accuracy. By streamlining these complex processes into one unified interface, SELMA offers an advanced SCI tool for tackling the challenges of processing audio with multiple speakers.

The final SCI tool stands out with its dynamic features that facilitate precise audio processing from the input stream:

- *Speaker Diarization:* The interface adeptly partitions the audio into distinct speaker-specific segments, enabling clear differentiation between participants in a conversation.
- *Customizable Models:* Leveraging the adaptability of the Whisper models, SELMA allows for incorporating specialized speaker models, enhancing recognition accuracy for specific use cases.
- *Feedback-Driven Learning:* A key innovation of SELMA is its ability to learn from user input. Corrections to speaker identifications are fed back into the system, continuously refining its accuracy.

- *Export Flexibility:* After processing, users can export the enriched data and updated models in various formats, tailoring the output to their requirements.

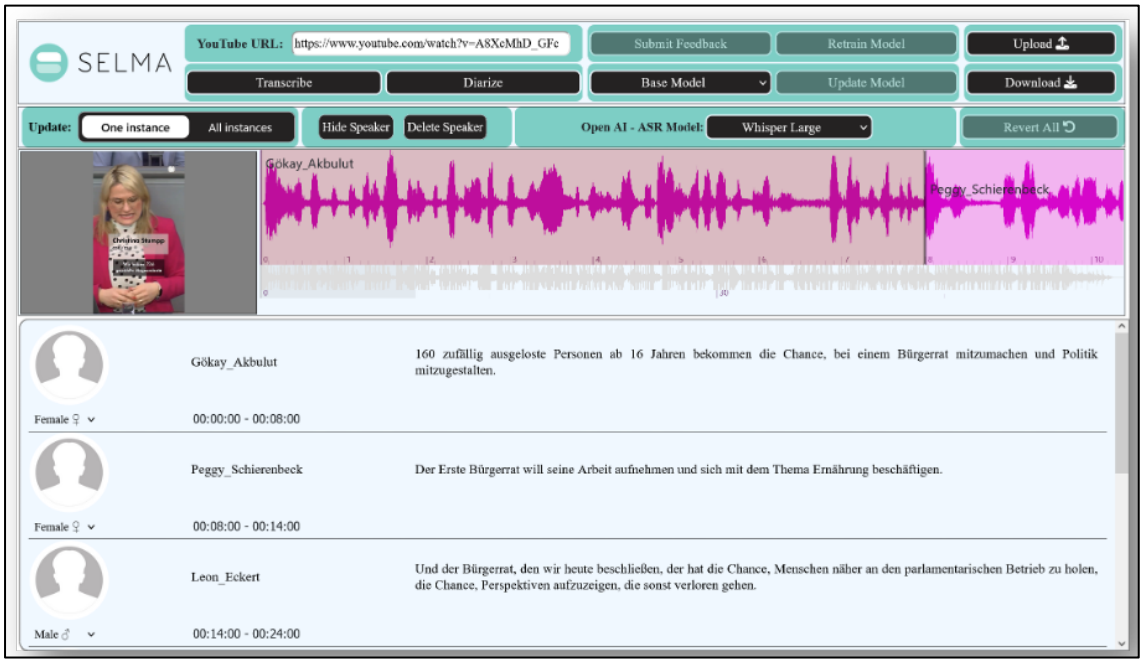

Figure 3 *SCI Interface Showing Speaker Diarization and Transcript Synchronization*

Each feature is crafted to ensure that SELMA meets and anticipates the user's needs with an effortless experience from start to finish. The SCI interface above enhances media processing with interactive elements that assist user engagement and ease. The waveform display visually represents the audio, enabling precise navigation and editing. Unique speaker identification tags simplify the speaker visualization throughout the audio. To streamline the user experience, functional buttons for actions like "Transcribe", "Diarize", and "Submit Feedback" are readily accessible, making the system intuitive even for those without technical expertise.

The technical workflow is based on a robust, Docker-based architecture, ensuring each component runs in an isolated environment, https://hub.docker.com/r/selmaproject/media-interface. This design allows for easy updates and customization of processing containers to specific recognition and transcription needs. Upon initiating an audio file upload, the Dockerized workflow orchestrates various services: it executes speaker diarization to distinguish between different speakers, aligns the transcription with audio segments, and accepts user edits for speaker updates. Feedback provided by users is a crucial input for the system's

adaptive learning, allowing for iterative refinement of the speech models within their respective containers. This Dockerized system presents a scalable and flexible approach to audio processing for users who demand precision and adaptability.

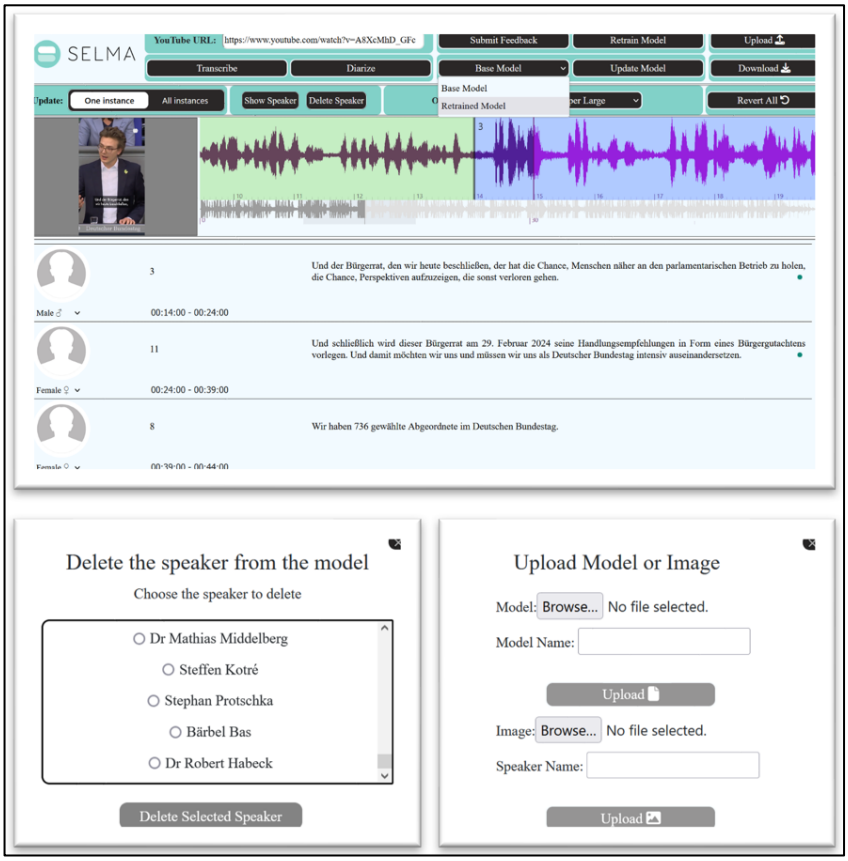

Figure 4 *Managing and Refining Speaker Recognition Data over the SCI Interface*

The Figure above shows the capabilities for editing speaker-specific segments: users can select a speaker segment, view the associated transcription, and make necessary corrections. The lower part reveals two crucial functionalities: the option to "Delete" the speaker from the model, which allows for the removal of unwanted data, and the "Upload" part can enhance the system's database by adding new speaker profiles. This process ensures the accuracy of the current transcription and continually optimizes the model for future analyses.

SELMA's final SCI system provides a useful tool in media processing capabilities. By fusing speaker recognition and diarization with speech-to-text transcription and user feedback, SELMA enriches the immediate user experience. It demonstrates a model where user interaction

can lead to continuous system improvement, promising more accurate and reliable speaker tasks in media services. Through Dockerized containers, the interface utilizes a modular design. By ensuring that speaker information is stored locally, SELMA aligns with data privacy standards, providing users with a powerful tool that protects data privacy.

2.7 Punctuation and Capitalization Recovery (PCR)

The Punctuation and Capitalization Recovery (PCR) tool is a state-of-the-art language processing application designed as part of our NLP tools. Hosted on the Hugging Face, https://hf.co/spaces/H2020SELMA/punctcap, and SELMA's Docker Hub, https://hub.docker.com/r/selmaproject/punctuation-casing, this tool addresses a fundamental challenge in text post-processing by restoring proper punctuation and capitalization to unstructured or improperly formatted text data. Such a tool is necessary for numerous real-world applications, such as processing transcribed speech or translated content to increase readability efforts.

Our tool's interface, built using Gradio, allows users to input raw text without punctuation and capitalization. The sophisticated models within the PCR tool then process this text to accurately predict and insert punctuation marks and adjust the casing of the letters. PCR leverages the XLM-RoBERTa machine-learning model that has been finely tuned for the nuances of each supported language. The result is a well-structured text with enhanced readability. The PCR tool functions as a web service encapsulated in a Docker container, which ensures ease of deployment and scalability.

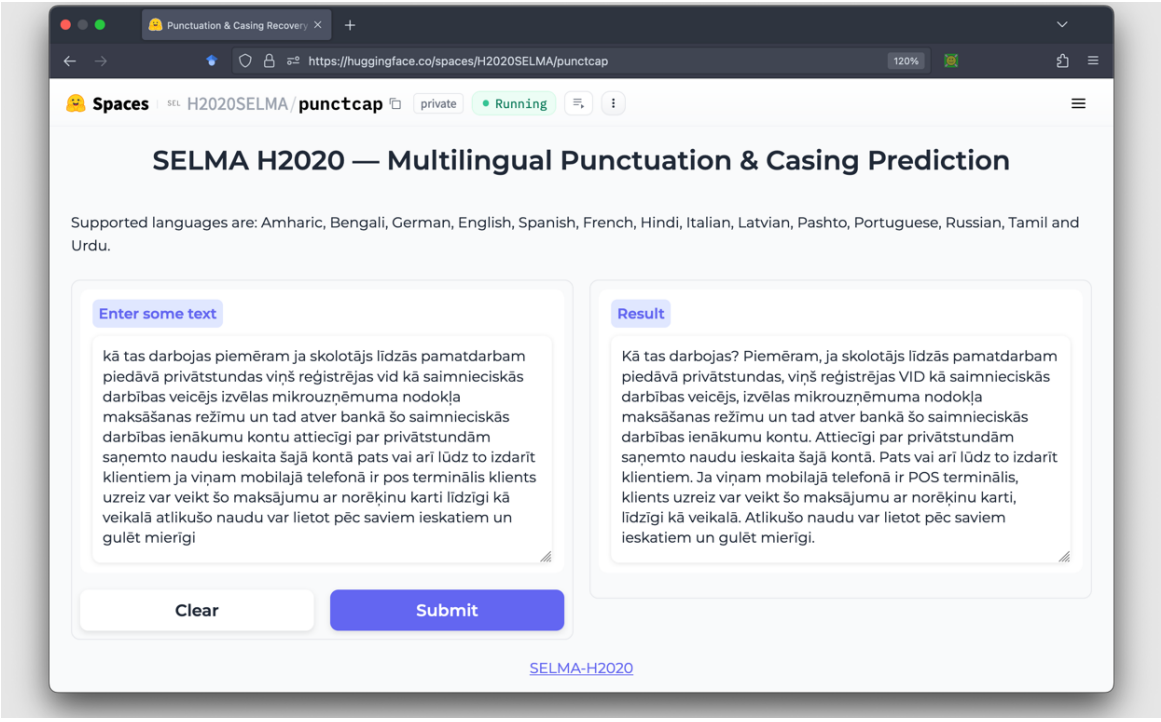

Figure 5 *Multilingual Punctuation and Capitalization Enhancement Interface Featuring Example Text from LSM.lv, which is a Latvian Public Media Website*

PCR employs a multi-lingual model trained to understand the syntax of the supported languages. A dedicated Urdu model is implemented to provide its specific linguistic features. At the same time, the multi-lingual class handles the punctuation and capitalization tasks for many other languages. Moreover, the underlying multi-lingual model operates on ONNX (Open Neural Network Exchange) format, offering high performance and interoperability. The punctuation and capitalization model utilizes a SentencePiece tokenizer for text segmentation, where the model's inference mechanism strips the original text of any existing punctuation, tokenizes the text, and applies the punctuation and capitalization predictions. The details and codebase are also provided at SELMA's GitHub page: https://github.com/SELMA-project/punctuation-capitalization-recovery.

The PCR tool exemplifies the practical application of advanced NLP techniques to address the challenges of text normalization across languages. It is multilingual by design, and covers languages widely spoken such as English, Spanish, and German to less commonly supported

languages like Amharic, Bengali, and Pashto, bridging the gap in other publicly available offerings like punctuator2[1] or fastPunct[2].

3. Conclusions

This final release signifies a pivotal advancement, specifically in speech-related tasks. By incorporating state-of-the-art neural architectures, our tools significantly reduce error rates and enhance performance across a diverse array of languages and functionalities. These solutions under WP3 illustrate a comprehensive effort to address and improve upon the multifaceted challenges in speech and text processing. They provide critical tools for media monitoring and journalism professionals, facilitating a more efficient and accurate dissemination of information.

Adding a user-feedback updating system illustrates a significant innovation, allowing real-time refinement of models based on direct input. This mechanism ensures that SELMA's performance, like in speaker identification or post-editing tasks, remains state-of-the-art by continuously learning from user interactions to enhance accuracy and reduce biases. The deployment of SELMA components in containerized environments, accessible via our Docker hub, underscores the project's dedication to open-source principles and collaborative development. This approach accelerates innovation and ensures that SELMA remains adaptable to the evolving needs of the natural language processing community.

[1] https://github.com/ottokart/punctuator2

[2] https://github.com/notAI-tech/fastPunct