SHORT TERM SCIENTIFIC MISSION (STSM) SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: CA19102
STSM title: Language In The Human-Machine Era
STSM start and end date: 2021-08-15 - 2021-08-21 15/08/2021 – 21/08/2021
Grantee name: Dr Alina Secară

PURPOSE OF THE STSM:

(max.200 words)
Under the title Understanding how Machine Learning Could Be Applied for Romanian Theatre Captioning for Deaf and Hard of Hearing Audiences, the mission aimed to enable Dr Secară to understand basic machine learning principles and explore ways in which it could be applied to accessibility services particularly for Romanian. The mission took place at RomSoft, Iasi, Romania where the grantee was able to talk to and shadow data scientists and machine learning specialists to investigate the topic above and aquire a basic understanding of the technical opportunities, as well as the resources needed to develop an automatic speech-based solution to support the delivery of captioning for Deaf and Hard of Hearing for Romanian.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS – max 500words

The work carried out during the knowledge exchange week fell within one of the two objectives set in the application:
1. understand the potential of technology to shape access services in Romanian;
This was achived through discussions, shadowing and demos from IT specialists.
The main interest was to investigate to what extent speech technologies, and particularly automatic speech transcription, could support the creation of an (semi)automatic theatre captioning system for Romanian. In practice, this would take an existing theatre play script in Romanian and automatically display on a chosen device (LED screen, mobile phone, tablet), and in synch, the lines/ sentences delivered by the actor on the stage.

In order to develop such a system, a deep learning process needs to be implemented for automatic speech recognition (ASR), so that the conversion of speech to text is done quickly and accurately. Such systems are routinely used for the creation of meeting captions, thus improving accessibility. One example is Amazon Transcribe. However, such systems are usually developed in English only, or are available for a very limited number of languages. For Romanian such systems are not widely available, even if some work was carried out – usually in research centres. For example, the domain-adaptable, large-vocabulary automatic speech recognition system for the Romanian language (LVCSR-ROM) developed in 2011 is currently not accessible https://speed.pub.ro/LVCSR-ROM .

Funded by the Horizon 2020 Framework Programme
of the European Union

One issue therefore is the scarcity of speech systems and the underlying resources for Romanian – we explored this issue and present some results of our research in the next section.

The desired technology needs to be able to recognise a variety of voices delivering lines from the script with a variety of pitches and speeds. Identifying different speakers even when their speech overlapped – technique called *speaker diarization* – was discussed and open-source technologies looked at. Methods for capturing training data for the fine-tuning of the ASR system were also identified and discussed. However, until the fine-tuning process is implemented, the basic ASR system needs to be implemented – in this respect available options were considered (Amazon, Microsoft, Google), yet the challenge of adapting and fine-tuning such closed systems kept coming up. Further investigations started during the mobility and need to continue into the cost and conditions of fine-tuning existing systems.

In parallel, additional aspects regarding the implementation of such a solution were discussed: latency for an optimal user experience; delivery medium (stage screen, tablet, mobile device); possibility of integrating multiple languages which users can choose from; the technology & workflow needed for producing these alternative language versions (machine translation can be explored, although in the creative sector the results are not expected to be overwhelming due to the pattern-based training of current MT systems).

2. set up a dialogue for long-term collaboration, such as a joint research proposal, with the IT company.

The company outlined that they have experience in participating in EU-funded projects and reiterated their interest in joining a potential consortium to apply for funding to research and develop an automatic solution for automatic theatre captioning for Romanian and potentially other EU languages, too. (see also final section)

DESCRIPTION OF THE MAIN RESULTS OBTAINED – max 500 words

An overview of the resources and techniques needed were explained, with an introduction to machine learning, speech synthesis, voice embeddings, speaker diarization and machine translation. The grantee was able to discuss and explore hands-on some of the topics explained. In particular, in order to create a system for automatic speech transcription in the context of theatre captioning, the following elements are needed, as a minimum:

1. a corpus of Romanian speech training data and language models

We investigated the CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, available for English only, to get an idea of the data needed to potentially create such a resource for Romanian. The EN corpus contains speech data uttered by 110 English speakers with various accents, 400 sentences for each speaker https://datashare.ed.ac.uk/handle/10283/3443
Another resource available for English is the LibriSpeech corpus https://www.openslr.org/12

We looked into the use of DeepSpeech Mozilla https://mycroft.ai/contribute/, an open speech to text technology for training from scratch approach with fine tuning the language model for correcting the output. The DeepSpeech architecture was used previously with Romanian, for example using the SWARA speech corpus: Romanian speech datasets freely available with speech data recorded from 17 speakers, manually segmented at utterance-level and semi-automatically labelled at phoneme-level; approximately 21 hours of high-quality read speech data.

We also looked into using The Montreal Forced Aligner https://montreal-forced-aligner.readthedocs.io/en/latest/introduction.html which performs orthographic transcription of an audio file and generates time-aligned output. It allows for training on any data and can be used with languages other than English. However, the pretrained acoustic models are not available for Romanian.

Future work can involve potential improvement of the Word Error Rate by adaping the parameters of the model and limit constraints such as the current small number of speakers and domains. We have also identified additional leads to contact for additional information on acoustic and language models for Romanian.

2. speaker diarization

Once a corpus is collected and cleaned, speaker diarization can be implemented. This is a task to label audio recordings with classes that correspond to speaker identity. In our situation this would mean that the system would be trained to automatically identify which actor speaks at a certain time and then correctly automatically assign their respective lines in the respective transcription.

For Romanian, speaker diarization was carried out on audio files of parliamentary speeches with different challenges (interruptions, noise, overlapping of speaker discourses, large variety of voices etc.) as part of the LIUM project. More than 21 hours of speech belonging to a few hundred speakers were used to evaluate the automatic process of segments assignation. In our project we would need to fine-tune such a system with data gathered in theatres during a number of rehearsals for the same play.

FUTURE COLLABORATIONS (if applicable)

Both parties will start investigating opportunities in recent Horizon Europe calls for a possible consortium creation to work on a automatic captioning system: https://tinyurl.com/27yyu2k4