UNIVERSITÀ DI TRENTO
Department of
Information Engineering and Computer Science

Università degli Studi di Trento
Dipartimento di Ingegneria e Scienza dell'Informazione

Scholarship Reference	**E-Gruppo-GPI**
Company (name and address)	**GPI, Via Ragazzi del '99, 13, 38123, Trento (Italy)**
Type of Scholarship	● **Professional Training**
Title of Scholarship	**Creating version 2 regarding a Speech Emotion Recognition AI Agent**
Industrial Tutor (full name + email address)	**Paolo Ranzi (paolo.ranzi@gpi.it)**
Academic Supervisor (full name + email address)	To be defined

Short Description of Internship and Thesis Activities, and Expected Outcome:

TITLE:
Creating version 2 regarding a Speech Emotion Recognition AI Agent

DESCRIPTION:
In GPI version 1 of a Speech Emotion Recognition AI agent is already running. In simple words, such and AI agent is able to detect 5 human emotions from human audio. The data-set consists in 9 Hrs of clean audio in Italian. The data-set is proprietary since it is owned by GPI. This AI agent is already implemented during the telehealth sessions offered by GPI. A telehealth session is a videocall between clinician and patient. The clinician can see on his screen in real-time those 5 human emotions in real-time. Every 4 secs an emotion is printed on the clinician screen. Whenever the patient stops talking, the AI agent stops itself automatically. The goal is to push such AI agent to a more advanced version (version 2).

METHOD:
Version 1 is based on the Deep Learning model developed in Patel et al. 2021, with the difference that a Variational AutoEncoder (VAE) has not been implemented. In a nutshell, version 1 takes an average of the
Mel-Frequency Cepstral Coefficient (MFCC) spectrogram. This is particularly useful since the deployment of the AI agent shows a 300 ms delay from the end of the 4 secs and the print of the emotion on the screen. Thanks to such an average, the AI agent performs almost real-time. The accuracy (computed against 30 % test-set) of version 1 is already around 84 %.
The challenge is to see whether the AI agent may be pushed towards version 2. In other words a VAE must be implemented. Further, the analysis of the whole spectrogram by using computer vision techniques and transfer learning (see Luna-Jiménez et al. 2021) should be implemented, as well. These improvements should increase overall accuracy.

HURDLES TO OVERCOME:
- the version 2 (namely, the version computing the whole spectrogram) may be slower, thus attracting complaints from end users (i.e. clinicians);

GOAL:
Push the Speech Emotion Recognition AI Agent from version 1 (average of the spectrogram) towards version 2 (whole spectrogram).

SCHEDULE (ROUGH ESTIMATE):
- month 1-2: read literature about anonymization and pseudo anonymization (e.g. Qian et al. 2017); read relevant literature about Speech Emotion Recognition;
- month 3: pre-processing data-set; build the Deep Learning model (by transfer learning/fine-tuning) with VAE included; train it;
- month 4: further iterations of the model and further training; test and validate the whole pipeline with unseen data;
- month 5: once the pipeline is stable, provide some help to developers with the deployment;
- month 6: writing thesis, documentation and polish code;

REFERENCES:
- Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., & Fernández-Martínez, F. (2021). Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. Sensors, 21(22), 7665.
- Patel, N., Patel, S., & Mankad, S. H. (2021). Impact of autoencoder based compact representation on emotion detection from audio. Journal of Ambient Intelligence and Humanized Computing, 1–19.
- Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X. Y., ... & Deng, Y. (2017). Voicemask: Anonymize and sanitize voice input on mobile devices. arXiv preprint arXiv:1711.11460.

Required Candidate Skills and Prerequisites:

SKILLS:
- experience with Python programming (e.g. TensorFlow, PyTorch), specifically Convolutional Neural Networks and VAEs;
- (optional) previous knowledge about scientific literature regarding Speech Emotion Recognition;
- some understanding of Linux systems (for getting computational power) and version control (e.g. GitHub, GitLab etc.);
- solid understanding of statistics;
- good English skills.