Abstract: “Humanitude” is a healing technique generalized by Gineste and Marescotti [5, 4] which claims to provide emotion recognition, multimodal fusion, audiovisual analysis, deep learning keywords: semantic data. This methodology based on affective speech is currently investigated in the MSH-HUMAVOX project which aims to better understand why this approach can significantly improve the life quality and reduce behavioral disorders associated with senile state. Hence, this work focuses on the analysis of audio speech signal possibly combined with video which showed its capability to convey relevant information about emotion and socio-cultural codes independently from the semantic content [11]. The goal of this internship is to investigate and propose innovative analysis multimodal methods allowing to recognize emotions from audio speech recordings using additional available information such as video and/or semantic data.

keywords: emotion recognition, multimodal fusion, audiovisual analysis, deep learning

Goals
— Bibliographical study for identifying the best state-of-the-art methods for multimodal emotion recognition
— Implementation/Proposal of new techniques for audio-visual emotion recognition
— Analysis and interpretation of the more relevant emotion audio-visual features

Methodology
The starting point of this research is our previous works on speech emotion recognition [10] and our work on prosody analysis of social attitudes which showed the relevance of several acoustic parameters such as the fundamental frequency ($F_0$) curve shape, the loudness and the duration of the estimated phonemes [2, 3]. The present study will consider more recent works for speech and video emotion recognition [6, 7] based on convolutional neural networks to discover additional features (or hidden units) present in recordings which convey relevant information about the emotion information. We also expect to investigate fusion strategies to efficiently combine all the available information present in each modality. We will define the best architecture (i.e. recurrent convolutional neural networks, Res-U-net, or wavenet [8]) in terms of accuracy and adaptability through a comparative evaluation with the state of the art [1].

Our study, will have a particular consideration of attention-based approaches which are promising in comparison to classical methods by their capability to focus on regions of interest of the input in a large number of prediction tasks [11]. Finally, we will apply the future new developed methods on real data collected in the MSH-Project “Humavox” using the emotion “Humanitude” taxonomy and we will develop a software prototype allowing to predict the emotional content from a speech signal.

Required profile
— good machine learning and signal processing knowledge
— mathematical understanding of the formal background
— excellent programming skills (Python, Matlab, C/C++, keras, tensorflow, pytorch, etc.)
— good motivation, high productivity and methodical works

Salary and perspectives
According to background and experience (a minimum of 577.50 euros/month). Possibility to pursue with a 3-year-funded PhD contract with French or international research partners.

Références