Master 2 research internship - 2022

# Multimodal Speech-Video Emotion Recognition applied to "Humanitude" Geriatric Medicine

Feb. 2023 - Aug. 2023 (6 months)


| | | |
|---|---|---|
| Supervisor(s) | : | Dominique Fourer, Desire Sidibié |
| Team / Laboratory | : | SIAM / IBISC (EA 4526) - Univ. Évry/Paris-Sacay |
| Collaborators | : | LaBRI, IMS (Univ. Bordeaux) |
| Contact | : | dominique.fourer@univ-evry.fr and drodesire.sidibie@univ-evry.fr |



**Abstract :** "Humanitude" is a healing technique generalized by Gineste and Marescotti [5, 4] which claims to provide optimal communication skills in elder care facilities. This approach based on affective communication between health professionals and elderly patients has succeeded to improve the cognitive capabilities of fragile people hospitalized in the context of health care in EHPAD. 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 [9, 3]. 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 to 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 an 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


[1] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92 :60–68, 2017.
[2] D. Fourer, T. Shochi, J-L. Rouas, and M. Guerry. On going bananas : Prosodic analysis of spoken japanese attitudes. In *Proc. Speech Prosody 2014 (SP'14)*, Dublin, Irland, May 2014.
[3] D. Fourer, T. Shochi, J-L. Rouas, and A. Rilliard. Perception of prosodic transformation for japanese social affects. In *Proc. Speech Prosody 2016 (SP'16)*, Boston, USA, June 2016.
[4] Y Gineste and R Marescotti. Interest of the philosophy of humanitude in caring for patients with alzheimer's disease. *Soins. Gerontologie*, (85) :26–27, 2010.
[5] Yves Gineste and Jérôme Pellissier. Humanitude. *Paris : Armand Collin*, 2007.
[6] Chen Guanghui and Zeng Xiaoping. Multi-modal emotion recognition by fusing correlation features of speech-visual. *IEEE Signal Processing Letters*, 28 :533–537, 2021.
[7] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques : A review. *IEEE Access*, 7 :117327–117345, 2019.
[8] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet : A generative model for raw audio. *arXiv preprint arXiv :1609.03499*, 2016.
[9] T. Shochi, D.Fourer, J-L. Rouas, G. Marine, and A. Rilliard. Perceptual evaluation of spoken japanese attitudes. In *Proc. International workshop on audio-visual affective prosody in social interaction and second language learning (AVAP'15)*, Bordeaux, France, March 2015.
[10] Sylvain Xia, Dominique Fourer, Liliana Audin, Jean-Luc Rouas, and Takaaki Shochi. Speech emotion recognition using time-frequency random circular shift and deep neural networks. In *Speech Prosody 2022*, 2022.
[11] Wenpeng Yin, Sebastian Ebert, and Hinrich Schütze. Attention-based convolutional neural network for machine comprehension. *arXiv preprint arXiv :1602.04341*, 2016.