Development and Evaluation of an AI-based pipeline to extract Emotion from Audio Recordings

Terry David Vuignier

Supervisor(s): Prof. Dr. Tobias Nef, Matilde Castelli
Institution(s): University of Bern, ARTORG Center for Biomedical Engineering Research
Examiners: Prof. Dr. Tobias Nef, Matilde Castelli

Introduction

Parkinson's disease (PD) is the second most common degenerative disease that significantly impairs patients and their relatives [1]. Treatments such as Levodopa and Deep Brain Stimulation can relieve patients' symptoms. However, their intake or parameters must be adjusted to the disease stage [2]. Since PD affects the emotional state of patients, speech sentiment analysis could give insight of the illness phase.

Materials and Methods

Machine Learning (ML) algorithms, such as Random Forest (RF) and Support Vector Machine (SVM), as well as a Convolution Neural Network (CNN) model were developed to classify audio recordings into three classes: negative, neutral and positive emotion. The models were trained with French public datasets, such as Oreau2 and Cafe, but also with English datasets, such as Crema and Tess. Mel-Frequency Cepstral Coefficients (MFCC), Chroma, Zero-Crossing Rate (ZCR), pitch and two measures of the intensity are the selected acoustic features which were extracted from the audio recordings to feed the models. Silence removal and addition of noise were used as data augmentation. At the same time, a newly French dataset was collected. One half was added in the training set and the other half was used to assess the models. The whole ML process is depicted on Fig.1. Finally, a part of the collected recordings was given to ChatGPT-4o to classify them based on acoustic features only or combined with the text.

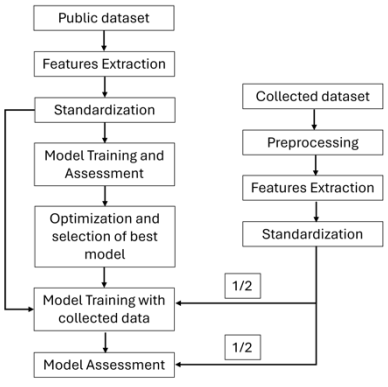

Fig. 1 Proposed pipeline for Machine Learning training and assessment.

Results

The models trained and tested with public and augmented datasets obtained an accuracy and F1-score of 80% for RF classifier and 84% for SVM classifier. The scores of the best ML models trained with public datasets and assessed with collected data were between 64% and 68% of accuracy, and between 35% and 47% of F1-score. Regarding the CNN model, the accuracy ranged from 51% to 65%, and the F1-score was between 31% and 42%. Detailed results are shown on Table 1. The accuracy and F1-score of ChatGPT-4o was 26.76% and 23.3% using only acoustic features, respectively. With the addition of text, theses scores were at 84.51% and 77.21%.

Assessment	Acc ML	F1 ML	Acc CNN	F1 CNN
Chroma+MFCC(13)	68,33	47,88	62,5	38,89
Chroma+MFCC(13)+ZRC	66,66	41,86	51,66	31,27
Pitch+RMS+LUFS Tess	64,16	35,92	65,83	42,23

Table 1 Results of Machine Learning (ML) and CNN models. Acc stands for accuracy while F1 represents the macro F1-score.

Discussion

The underwhelming results got during the assessment of the different models are due to the difference between acting speech in the training set and spontaneous speech in the collected dataset, the background noise and the quality of the smartphones microphones. The results can be enhanced with better microphone quality during data collection, combination of other extracted features, addition of text and spontaneous speech dataset into the training set.

References

[1] D. Hemmerling et al., "Vision transformer for parkinson's disease classification using multilingual sustained vowel recordings", pages 1–4, 2023.
[2] Katherine Marshall and Deborah Hale, "Parkinson disease. Home Healthcare Now", 38:48–49, 01 2020.

Acknowledgements

I would like to express my thanks to Matilde Castelli and Prof. Dr. Tobias Nef for their support and guidance throughout this interesting project. The encouragements and help from the members of the Gerontechnology and Rehabilitation group were sincerely appreciated.