Investigating Pre-Trained Self-Supervised Deep Learning Models for Disease Recognition

Julius J. Bijkerk
12219967

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Prof. dr. P.P.G. (Paul) Boersma

Institute for Logic, Language and Computation
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

June 4, 2023

Contents

1 Introduction **4**
- 1.1 Research Question . . . 5
- 1.2 Hypothesis . . . 6

2 Theoretical Background **6**
- 2.1 Diseases Affecting Speech . . . 6
- 2.2 Traditional Speech Features . . . 7
  - 2.2.1 MFCC . . . 7
- 2.3 Pre-trained Self-Supervised Deep-Learning . . . 7
  - 2.3.1 Deep-Learning . . . 7
  - 2.3.2 Self-Supervision . . . 7
  - 2.3.3 Pre-Training . . . 7
  - 2.3.4 Fine-Tuning . . . 7
  - 2.3.5 Classifier . . . 7
  - 2.3.6 Wav2Vec . . . 7
  - 2.3.7 huBERT . . . 7
- 2.4 Literature Review . . . 7
- 2.5 Research Gap or Gap Analysis . . . 8

3 Method and approach **8**
- 3.1 Data . . . 8
  - 3.1.1 The TORGO database . . . 9
- 3.2 Replication . . . 9
  - 3.2.1 Exploratory Data Analysis (EDA) . . . 9
  - 3.2.2 Dataset preprocessing . . . 10
  - 3.2.3 Dataset Splitting . . . 10
  - 3.2.4 Fine-tuning Wav2Vec . . . 10
  - 3.2.5 Feature Extraction . . . 10
  - 3.2.6 Training the Classifier (yet to be chosen) . . . 10
  - 3.2.7 Model Validation (necessary?) . . . 10
  - 3.2.8 Model Evaluation . . . 10
  - 3.2.9 Model Deployment . . . 10
- 3.3 MFCC . . . 10
- 3.4 huBERT . . . 10

4 Results and Evaluation **10**

5 Discussion **11**
- 5.1 Real-World Application . . . 11
- 5.2 Ethics . . . 11
  - 5.2.1 Bias . . . 11
  - 5.2.2 Data privacy . . . 11
- 5.3 Future Work . . . 11

6 Conclusion **11**

1 Introduction

DOCTOR: "Can you please say 'Aah' for me?"
PATIENT: "Aaaah"
DOCTOR: "Thank you. I see your pitch and loudness have stabilized more since the last telephone appointment.[1] Since these could be the first signs of Parkinson's disease, I would like to schedule an appointment for further investigation."

The above dialogue could be the first step in the diagnosis of Parkinson's disease, and possibly many other diseases, in the near future.

This will have been made possible by the enormous advancements in artificial intelligence (AI) in recent years. These achievements are often imagined in terms of self-driving cars or a robot beating the top-human players of games such as Chess and Go, but there are many more fields in which AI could be of significant value. One of these major promising areas is healthcare. The diagnosis of diseases often requires time-intensive appointments or procedures, for doctors and patients. Additionally, it may be hard for specialists as well to remember and recognize subtle changes in disease biomarkers such as voice pitch or other changes in speech over time.

Together with the increasing interest in AI, significant attention is now being given to data and its collection. Over the years, a vast amount of data has been collected within the health industry, such as images and speech audio of people with certain conditions. This creates the possibility of recognizing these conditions by machine. Together with the fact that most people on the planet now own a (smart)phone, this could optimize healthcare in terms of completely non-invasive, low-cost, and pseudo-real-time diagnosis of different diseases or disorders [Costantini et al., 2023]. Therefore, further research in this area would be very valuable.

A distinction could be made between the two main kinds of conditions that have been proven to be identifiable by AI. Firstly, these are psychiatric disorders, including bipolar disorder, depression, and stress [Ma et al., 2020]. Secondly, also neurodegenerative diseases, like Alzheimer's, and Parkinson's disease, and speech impairments (aphasia, dysarthria, and dysphonia) [Hecker et al., 2022] seem to be recognizable with the help of machine learning.

During my initial stages of research, I found many different AI methods in a wide variety of papers. The following are some samples of these methods: Naive Bayes, Logistic Regression, K-Nearest Neighbours, SVM, LSTM, and CNN. In this research, I will lay the focus on the most promising methods until now in the field of audio recognition: pre-trained self-supervised deep-learning models. Examples of these are Wav2Vec and huBert, both models created by Facebook's AI research department. The most promising of these methods will be compared and tested against each other on the dataset of one specific disease, like Alzheimer's or Parkinson's.

[1] [Ma et al., 2020]

This thesis is structured as follows: hereafter you will find the research question and my hypothesis as the final part of the introduction. Section 2 will cover the theoretical background, which includes the current state of research and relevant technical explanations. In section 3, I will elaborate on my methodology for replicating the research paper: *'On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches'*, by Schu, Guilherme and Janbakhshi, Parvaneh and Kodrasi, Ina.[2] This includes an explanation of the TORGO dataset. In section 4, the results of my method will be presented and evaluated. Section 5 covers a sketch of how this could be implemented in the real world, the ethical considerations it brings, and possible further research directions within this area. Additionally, in sections 4 and 5 the sub-questions will be answered. Lastly, in section 6, I will present the conclusion, containing the answer to the main research question of this research.

1.1 Research Question

The topic of this research will be the following: "Investigating Pre-Trained Self-Supervised Deep Learning Models for Disease Recognition". Which will be narrowed down to the following main research question:

How effective are pre-trained self-supervised deep-learning models in recognizing diseases?

This main research question is broken down into separate sub-questions whose answers will be fundamental to answering the main question. These sub-questions are the following:

1. What types of diseases and disorders can be recognized through speech, and how do these conditions alter speech patterns?

2. What is the traditional approach to diagnosing these diseases, and how does it compare to the intended use of pre-trained self-supervised models, or artificial intelligence in general?

3. What are the key principles and capabilities of self-supervised deep-learning models, and how are they pre-trained?

4. How well do pre-trained self-supervised models perform in disease recognition tasks compared to traditional methods?

[2] `https://arxiv.org/abs/2211.08833`

5. What are the ethical considerations when using AI for disease recognition through speech?

6. What are the overall current limitations and the potential future improvements for using pre-trained self-supervised deep-learning models in disease recognition?

1.2 Hypothesis

The hypothesis for the main research question is that pre-trained self-supervised deep-learning models could be very effective in recognizing diseases in an early stage and therefore could be a major contribution to healthcare, but at this moment there are still some important issues to be solved before it can work effectively in practice. For example, data shortage, privacy issues and lack of explanation.

2 Theoretical Background

In this chapter, I will give an overview of the concepts that will be necessary for understanding the origin of the research question, its relevance, and how to develop an answer to it. I will discuss and refer to earlier research which will show the current state of the problem and the research area it is in. Therefore, this part functions as well as a technical background. Finally, section 2.5 provides a gap analysis.

2.1 Diseases Affecting Speech

Human speech could be affected by a vast amount of different diseases. Human voice production occurs through complex movements of different physical systems in combination with neurological control systems. Both of these systems can be affected by a speaker's health condition [Gómez-García et al., 2019]. When I mention physical systems, think of: vocal cords, glottis and oral cavity for example. And when I mention neurological control systems, think of: Broca's and Wernicke's areas.

Their combination encodes both linguistic information (speech content, like usage of words and length of sentences) and acoustic information (internal speech quality, like pitch and tone). Within this research there will be a focus on the latter: the quality of speech, and not the contents.

- Distinction between disease and disorder (say something about neurological and stress)

- Traditional assessment by a doctor

- Speech biomarkers, which can be extracted by machines, as described in the next section/

2.2 Traditional Speech Features

Handcrafted speech features (INBETWEEN): Pitch, tone, f1, f2, MFCC, jitter, shimmer

- PRAAT

2.2.1 MFCC

- Focus on MFCC since this one will be used as a comparison to wav2vec and huBERT.

2.3 Pre-trained Self-Supervised Deep-Learning

2.3.1 Deep-Learning

2.3.2 Self-Supervision

2.3.3 Pre-Training

Transfer learning

2.3.4 Fine-Tuning

2.3.5 Classifier

2.3.6 Wav2Vec

- explain the architecture and training methods of the specific models I'm using
  - Wav2vec, Wav2Vec2.0, XLSR-wav2vec
  - Facebook's Wav2Vec2 XLS-R model is a large multilingual model trained on 128 languages and with 436K hours of speech
  - Transformer
  - Encoder-Decoder

2.3.7 huBERT

- explain the architecture and training methods of the specific models I'm using
  - Facebook's AI modification of Google's BERT model.

2.4 Literature Review

Incorporate Literature review within the Theoretical Background

- substantiate claims and explanations in previous subsections with found literature

2.5 Research Gap or Gap Analysis

3 Method and approach

The critical start of this research consisted of finding a relevant research paper that performed disease detection in the way I intended, i.e., using a Pre-Trained Self-Supervised Deep Learning Model (preferably Wav2vec). I found a considerable amount of research published within this area, but not something that directly suited the intention of this current paper. Most obstructions were in terms of datasets that they used (hardly accessible or for example also containing the interviewer's voice) or in terms of the model (using deep learning on spectrograms, not on raw audio or the use deep learning within a multi-model approach, which are out of the scope of this thesis).

At some point, I discovered the article: *'On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches'*, by Schu, Guilherme and Janbakhshi, Parvaneh and Kodrasi, Ina.[3] This is about exploring the difference in results between disease recognition with and without noise filtering and on noise only. One of the aspects of that research is performing Wav2vec on speech audio of the TORGO dataset.

While researching the topic, I discovered that most medical data sets are only available on request. Besides that, as a thesis student, you also need your supervisor to request the data, for example, the DemantiaBank dataset [Lanzi et al., 2023], but also other valuable data resources, for example the ones that are described in [Hecker et al., 2022]. This does not make the research impossible but it is a real obstacle in this field.

When handling new data, some ideas are important to keep in mind. First, how is the data structured: patient only or audio of a conversation between patient and doctor? Should the speech audio be extracted from the whole, to filter out silences, or filter out the background noise [Schu et al., 2023]? Lastly, wav2vec expects a sample rate of 16000 kHz, resampling might be necessary.

Therefore I started by setting up a working environment that makes use of Google Colab for coding, Google Drive for easy storage, and GitHub for version control. The programming for the replication and the follow-up versions will be done in Python accompanied by different machine-learning libraries.

3.1 Data

Explain Torgo and possibility to apply algorithm to similar datasets.

[3] `https://arxiv.org/abs/2211.08833`

3.1.1 The TORGO database

The TORGO database, as exuberantly described in [Rudzicz et al., 2012], consist of male and female speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). As a consequence of their disease, these individuals all suffer from dyarthria. Dysarhtria is caused by disruptions in the neuro-motor interface. These disruptions bend the motor commands to the vocal articulators, resulting in atypical and relatively incomprehensible speech in most cases [Kent, 2000]. For each of the individuals with dysarthria, there is a matched individual in the control group (except for one dysarthric male). This comes down to:

```
TORGO
├── 3 Dysarthric females
├── 3 Non-dysarthric females
├── 5 Dysarthric males
└── 4 Non-dysarthric males
```

All subjects are recorded with two separate microphones, while performing some sort of speech test

3.2 Replication

During the research phase, I discovered a relevant article that I want to use for my replication: *Comparing Acoustic-based Approaches for Alzheimer's Disease Detection* [Balagopalan and Novikova, 2021]. Since the replication of their method, only on the basis of their paper, did not work out easily, I reached out to the authors: Jekaterina and Aparna. Aparna responded shortly after with some very helpful information about their implementation. Not entirely coincidental, they based their approach on the work of Mehrdad Farahani [4] [5] which I also discovered during my research.

Mehrdad created an emotion recognition model for Greek speech using wav2vec2-xlsr. The approach of Jekaterina and Aparna was to modify this notebook into their Alzheimer's recognition instead of emotions. My approach is thus the same.

This approach consists of the following steps:

3.2.1 Exploratory Data Analysis (EDA)

First impression of what the data looks or sounds like.

[4] https://github.com/m3hrdadfi/soxan/blob/main/notebooks/Emotion_recognition_in_Greek_speech_using_Wav2Vec2.ipynb

[5] https://scholar.google.com/citations?user=0raqKZEAAAAJ&hl=en

3.2.2 Dataset preprocessing

May consists of resampling, wav2vec expects 16000 kHz or redistributing files over folders.

3.2.3 Dataset Splitting

Split the original dataset into a: training set, an evaluation/validation set, and a test set.

3.2.4 Fine-tuning Wav2Vec

Fine-tune Wav2Vec on training set.

3.2.5 Feature Extraction

Extract features from raw audio samples (of all subsets), by converting them into a set of feature vectors (embeddings) with the use of the now pre-trained Wav2Vec.

3.2.6 Training the Classifier (yet to be chosen)

Train the desired classifier on the set of feature vectors obtained from the previous step.

3.2.7 Model Validation (necessary?)

Use the evaluation/validation set to tweak the model's hyperparameters to avoid overfitting.

3.2.8 Model Evaluation

Evaluate the model's performance on the test set.

3.2.9 Model Deployment

Test model on new data, for example: own speech recording.

3.3 MFCC

3.4 huBERT

4 Results and Evaluation

Findings

Show the performance results of the different algorithms against each other.

- If possible, also consider discussing the computational efficiency of each method (e.g., training time, prediction time), as this can be an important factor in real-world applications.

Precision, recall, F1-score, and AUC-ROC (accuracy)

There are important differences to consider here about which has the greatest negative impact: False Positives or False Negatives. E.g. is it preferred to classify someone as 'having a certain disease', and potentially treating like it, while the person does not have the disease, compared to the other way around? If both types of errors are equally important, the F1 score or AUC-ROC might be the preferred evaluation measurement.

5 Discussion

5.1 Real-World Application

5.2 Ethics

5.2.1 Bias

5.2.2 Data privacy

5.3 Future Work

- Something with noise (over phone)

6 Conclusion

- Summarize and repeat all of the previous.
  - Don't mention complete new things.
  - Answer sub-questions and combine them to give an answer to the main research question, which is again the following: "What are the possibilities and constraints of using pre-trained self-supervised deep-learning models for disease recognition through speech?".

References

[Balagopalan and Novikova, 2021] Balagopalan, A. and Novikova, J. (2021). Comparing acoustic-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2106.01555*.

[Costantini et al., 2023] Costantini, G., Cesarini, V., Di Leo, P., Amato, F., Suppa, A., Asci, F., Pisani, A., Calculli, A., and Saggio, G. (2023). Artificial intelligence-based voice assessment of patients with parkinson's disease off and on treatment: Machine vs. deep-learning comparison. *Sensors*, 23(4):2293.

[Gómez-García et al., 2019] Gómez-García, J. A., Moro-Velázquez, L., and Godino-Llorente, J. I. (2019). On the design of automatic voice condition analysis systems. part i: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51:181–199.

[Hecker et al., 2022] Hecker, P., Steckhan, N., Eyben, F., Schuller, B. W., and Arnrich, B. (2022). Voice analysis for neurological disorder recognition–a systematic review and perspective on emerging trends. *Frontiers in Digital Health*, 4.

[Kent, 2000] Kent, R. D. (2000). Research on speech motor control and its disorders: A review and prospective. *Journal of Communication disorders*, 33(5):391–428.

[Lanzi et al., 2023] Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., and Cohen, M. L. (2023). Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438.

[Ma et al., 2020] Ma, A., Lau, K. K., and Thyagarajan, D. (2020). Voice changes in parkinson's disease: What are they telling us? *Journal of Clinical Neuroscience*, 72:1–7.

[Rudzicz et al., 2012] Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2012). The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:523–541.

[Schu et al., 2023] Schu, G., Janbakhshi, P., and Kodrasi, I. (2023). On using the ua-speech and torgo databases to validate automatic dysarthric speech classification approaches. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.