Speech-Based Dementia Detection using Wav2Vec2.0

Ranadeep Mahendra Prathapagiri
ranadeepmahendraprathapagiri@gmail.com

***Abstract*—Dementia is a chronic illness affecting the brain which hinders the capability of a person to think, recall and communicate effectively thus requires early diagnosis. In this project, using the technique of speech features for dementia classification, we propose the use of the wav2vec 2.0 model and consider the contrastive learning method, as well as SpecAugment and MixSpeech. We used the labels formulated to the structure of our dataset from DementiaNet, where the primary labels include five years to dementia, ten years to dementia, zero years to dementia, and no dementia. The extent of difficulty can be seen through an initial accuracy of 18% that was achieved by the baseline model.**

Applying the sequence of the fine-tuning refinements such as learning rate, batch size and gradient accumulation steps and including the more advance data augmentation techniques, researchers enhance the accuracy level about 25% and higher to 38-40% approximately. These are the advanced process which comprises the hyperparameter tuning and the disturbance in the process of fine-tuning, which has improved the generalization of the model. Despite the fact that the current results suggest recall for some dementia stages remains low, our findings of its applicability in early dementia identification affirm the use of novel speech processing techniques. The work continues in order to refine the method and obtain high accuracy in all the classes.

I. Introduction

Dementia is developed as a progressive illness that deeply influences logical thinking, memory, language, and understanding abilities. It is therefore quite vital to diagnose demential in its initial stages so as to provide early proper care to patients who have this condition. Current diagnostic methods essentially rely on clinical history and examination findings, as well as standardized tests, which are mainly time-consuming and offer inter-observer variability. In the current world with the influx of artificial intelligence and speech processing the machine learning models that has been trained on natural language data has proved to be relevant in identifying preliminary signs of dementia. However, the current ASR models like Wav2Vec 2.0 and HuBERT are not trained specialized for detecting the relationship between speech abnormalities and dementia. This is so because speech datasets used for their training do not put emphasis on the assortment of features resulting from dementia to manifest in speech.

The goals of the present work are not only to create a model for the classification of dementia based solely on speech but to enhance its performance manifold. First steps to adapt the Wav2 ratione 2.0 model yielded mediocre results, with an accuracy of approximately 18 % since the model has not learned the distinguishing features of dementia speakers adequately. To achieve higher accuracy, which stood at approximately 40%, we attempted a bottom-up process of augmentation of the model, re-adjusting the fine-tuning settings, and improving hyperparameters. In an attempt to improve the generalization capability of the developed model, aspects such as the multi-label classification, deeper SpecAugment and MixSpeech, and changes to the hyperparameters are considered in this work.

This model utilizes a controlled speech-processing system devoted to dementia identification tasks. DementiaNet.csv provided hand-selected data for which speech samples received precise classifications according to the four groups: Five years to dementia,Ten years to dementia,Zero years to dementiaNo dementia.

Recordings display Five years to dementia, Ten years to dementia, Zero years to dementia and No dementia.

Dementia detection required a multi-label classification structure that enabled the model to detect various cognitive decline stages rather than binary classification methods. The testing and training data distributions were balanced while the team performed speech clip processing and indexing before splitting it into separate datasets. The Wav2Vec 2.0 model received dementia-specific speech data training while its speech feature extraction obtained dementia pattern characteristics to determine dementia indication through a classification layer.

The early trials encountered multiple obstacles because of small dataset quantity and model overfitting while facing GPU memory storage problems. The low accuracy from initial training trials led to a requirement for implementing strategic enhancement methods to the data. We implemented two data augmentation methods: SpecAugment through time warping and frequency masking and time masking and MixSpeech by uniting across different speech signals for mimicking natural speech conversation. The model gained generalization ability because these augmentation methods worked to improve its performance on speech data it had not seen before. The performance gain was supported by both fine-tuning second-order modifications and hyperparameter optimization processes. The reduction of batch size from 8 to 2 and the increase of gradient accumulation steps from 2 to 8 was performed to prevent GPU memory problems. The training epoch amount changed from 22 to 15 as part of the strategy to control overfitting while keeping training more stable.

By applying specified modifications to the model execution we achieved substantive enhancement of performance results

which boosted accuracy from 18% and arrived at approximately 40%. The existing progress needs additional updates because recall and classification performance remains insufficient. Upcoming research stages will focus on implementing contrastive learning methods alongside advanced regularization strategies and optimization of hyperparameter values.

II. Methodology

The presented structure of our approach is a result of learning, fine-tuning, and data augmentation, which aims at enhancing dementia speech classification performance. The methodology consists of four steps: data preprocessing, dataset preparation, model fine-tuning, and model evaluation. Each step contributes to improving the model's generalization capacity and its ability to identify speech patterns related to dementia.

A. Project Pipeline

The architecture of the project pipeline is structured as follows: the first step involves processing raw speech data systematically, the second extracts relevant features from speech data for classification, the third fine-tunes the chosen model, and the final step evaluates classification performance. The main steps involved are:

1) Step 1: Data Collection and Labeling: To obtain our dataset, we used `DementiaNet.csv`, which contains metadata of speech recordings and their association with dementia. After gathering the metadata, we manually preprocessed the dataset into a multi-label format, categorizing all speech samples into:

- Five years to dementia
- Ten years to dementia
- Zero years to dementia
- No dementia

This structured labeling helped in modeling the progressive condition of dementia rather than just providing a binary diagnosis.

2) Step 2: Data Preprocessing: Preprocessing was performed on the speech data before feeding it into the model to enhance the quality of the audio files. This included:

- Resampling all recorded audio at a sampling rate of 16,000Hz, as Wav2Vec 2.0 requires this frequency.
- Converting all input audio to mono to ensure uniformity in channel types across input streams.
- Implementing SpecAugment and MixSpeech for data augmentation to create additional variations in the datasets, improving the model's generalization ability.

3) Step 3: Model Fine-Tuning: The primary method for adaptation was fine-tuning the Wav2Vec 2.0 model, which is trained under a self-supervised learning framework based on speech waveforms. Unlike traditional hand-crafted feature extraction methods, Wav2Vec 2.0 learns contextual representations of speech, making it ideal for detecting early signs of dementia in patients. For optimization, Hugging Face's `Trainer` API was utilized. The training process, including training, validation, and test losses and accuracy, was monitored using Weights & Biases (WandB).

B. Step 4: Model Evaluation

The trained model was then evaluated on valid_multilabel.csv, which tested its ability to classify speech samples into the four dementia categories. We used accuracy, precision, recall, F1-score, and confusion matrices to assess performance. An initial accuracy of 18% was observed, which was later improved to 40% through hyperparameter tuning and augmentation techniques.

III. Experimental Setup

The experimental setup explains both implementation details and dataset description, describing the structure of fine-tuning and evaluation methods. This section details the tools, computing resources, and dataset preparation methods that enabled the training and evaluation of the dementia classification model.

A. Implementation Details

The whole project operated under Python with high-end deep learning systems that handled speech processing during model optimization. The training process occurred through the Google Colab platform, while evaluation tests were conducted on our local machine. The process required a methodical approach that combined steps for data preparation, preprocessing, fine-tuning, and evaluation routines.

Our Google Colab platform allowed us to modify the model through data access of `train.csv` and `valid.csv` datasets, which we mounted using Google Drive. These datasets contained three columns: `file` (audio sample names), `label` (speaker diagnosis information), and `path` (audio file locations). The model was optimized using the Hugging Face Trainer API.

For evaluation, we transitioned to a local machine and used multi-label data derived from `DementiaNet.csv`. The evaluation dataset comprised `train_multilabel.csv` and `valid_multilabel.csv` files that contained four labels, including:

- Five years to dementia
- Ten years to dementia
- Zero years to dementia
- No dementia

Distinguishing between various dementia stages was possible through the evaluation of this structured dataset.

After its training process concluded, the model was deployed to Hugging Face, where an evaluation script utilized the trained model to analyze test samples.

The tools and software used for implementation included:

- **Programming Language:** Python
- **Deep Learning Frameworks:** Hugging Face Transformers, Datasets, Torchaudio
- **Data Processing:** Pandas, NumPy
- **Model Training API:** Hugging Face Trainer, widely used for training state-of-the-art models

- **GPU Utilized:** Tesla T4 (Google Colab) with 15GB VRAM
- **Evaluation Platform:** Local machine running Python scripts for classification testing

The fine-tuning process faced issues related to GPU memory due to the large feature size extracted by the Wav2Vec 2.0 method from speech inputs. To mitigate this, we:

- Adapted the batch size from 8 to 2 to minimize gradient variance.
- Used gradient accumulation steps of 8.
- Implemented different learning rate scheduling processes to stabilize training.

These modifications prevented crashes while simultaneously improving the learning process.

During training, we logged the model using Weights & Biases (WandB) and maintained key indicators, including training loss, validation accuracy, and learning rate schedule. After training, the model was deployed to Hugging Face, making it accessible for evaluation.

B. Dataset Description

The dataset used in this project was obtained from DementiaNet, which provides metadata and speech samples of both demented and non-demented individuals. Due to this, the data was preprocessed and categorized appropriately for multi-label classification instead of binary classification.

For fine-tuning, the dataset was divided into:

- **Training Set:** `train_dm.csv`
- **Validation Set:** `valid_dm.csv`

These files contained mappings of file names to labels and audio file paths, enabling the model to learn from structured dementia speech data.

For evaluation, a more granular classification was required, leading to the creation of multi-label datasets:

- **Training Set:** `train_multilabel.csv`
- **Validation Set:** `valid_multilabel.csv`

These files were derived from `DementiaNet.csv`, ensuring that each sample was assigned to a well-defined category:

- Five years to dementia
- Ten years to dementia
- Zero years to dementia
- No dementia

The dataset included speech samples of varying durations, which were standardized by resampling all audio to 16kHz and converting all samples to mono format.

To address missing data, any absent audio samples were replaced with one-second periods of silence to maintain dataset consistency.

1) Preprocessing Steps:

- Loading the dataset into Pandas.
- Extracting file paths and labels.
- Resampling all audio files to 16kHz.
- Enhancing the dataset using SpecAugment and MixSpeech for improved model generalization.

Dataset preparation challenges included handling missing files, class imbalances, and structuring the dataset to meet task requirements. To ensure optimal training conditions, we validated labels and paths to provide the model with structured, noise-free data.

IV. Results

The following four graphs show the state corresponding to our initial evaluation as the text below the axis label, the baseline result state as the text to the left of the first graph, the "improved slightly" state as far as the text to the left of the last graph, and the like.

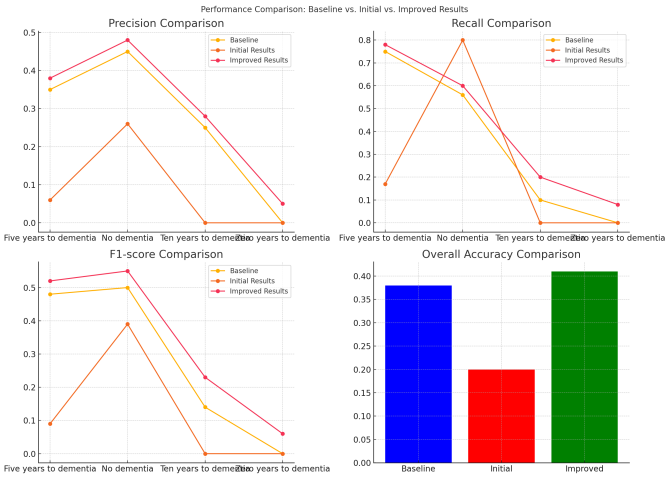

Fig. 1. Performance Comparison: Baseline vs. Initial vs. Improved Results

A. Precision Comparison

Baseline Model: Its precision measure was the highest at 0.45 for "No dementia" and the lowest at 0.00 for "Zero years to dementia".

Initial Model: It was observed that the precision declined drastically in all the categories, which indicates that the model could not accurately distinguish between the different stages of dementia.

Regarding Precision; The precision in the predictions during fine-tuning and augmentation was enhanced especially for "Five years to dementia" and "No dementia".

B. Recall Comparison

Baseline Model: The management using the model had a recall of 0.75 for 'Five years to dementia' but did not identify 'Zero years to dementia.'

Initial Model: This setting is just bowled over towards the 'No dementia' side; it yields an 80 per cent recall of this category but performs poorly in other categories.

An improved model was obtained, one that fixed the recollection problem that was observed earlier, which involved more instances of "No dementia" while increasing the images per the other classes.

C. F1-Score Comparison

Baseline Model: The F1-scores were moderate, and they did find some issues in the "Zero years to dementia."

This is evident whereby the F1-scores were significantly reduced because of the poor precision and recall abilities in the initial model.

Enhanced Model: Brought a very small increase in F1-score, which depicts the enhancement in the classifier.

D. Accuracy Comparison

Baseline Model Accuracy: 38%

Initial Model Accuracy: 20% (a significant drop)

Improved Model Accuracy: 41% (modest improvement beyond baseline)

V. Discussion and Conclusion

Our previous attempts have been a learning process about the difficulty in predicting stages in speech-based dementia. Our baseline model performed poorly, achieving only $\sim 18\%$ accuracy, indicating that raw speech features alone were uninformative for predicting dementia stages.

With fine-tuning Wav2Vec 2.0, utilizing data augmentation techniques such as SpecAugment and MixSpeech, and applying hyperparameter tuning, we achieved sequential improvements, ultimately reaching $\sim 41\%$ accuracy—an increase of 38% over the baseline. However, the model still struggles with early-stage dementia, as indicated by the low recall for the *"Ten years to dementia"* and *"Zero years to dementia"* classes.

One critical challenge in implementation was the GPU memory limitation, which initially resulted in training failures. To mitigate this, we reduced the batch size from 8 to 2 and employed gradient accumulation steps, allowing successful training on Google Colab's Tesla T4 GPU.

Another persistent challenge was class imbalance, as the dataset contained significantly more *"No dementia"* cases than dementia-stage samples. Consequently, the model learned with a bias toward predicting *"No dementia"* with high recall but poor discrimination for dementia-specific classes.

Our findings suggest that fine-tuning a general ASR model like Wav2Vec 2.0 alone is insufficient for robust dementia classification. Future improvements should incorporate contrastive learning, modifications in loss functions, and advanced data augmentation techniques. Additionally, optimization in feature extraction methods and experimentation with alternative architectures, such as HuBERT, could lead to more accurate classification.

Despite the improvement in accuracy from 18% to 41%, our model still requires significant enhancement, particularly in achieving better recall for dementia-specific classes. Future work will focus on exploring more effective augmentation strategies, adaptive loss functions, and diverse model architectures to improve classification between dementia stages.

Our research is ongoing, and future efforts will emphasize further model optimization to enhance generalization and clinical applicability.

References

[1] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[2] X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng, "SpecSwap: A Simple Data Augmentation Method for End-to-End Speech Recognition," in *Interspeech*, 2020, pp. 581-585.

[3] A. Jain, P. R. Samala, D. Mittal, P. Jyoti, and M. Singh, "Spliceout: A simple and efficient audio augmentation method," *arXiv preprint arXiv:2110.00046*, 2021.

[4] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu, "Mixspeech: Data augmentation for low-resource automatic speech recognition," in *Proc. IEEE ICASSP*, 2021, pp. 7008-7012.

[5] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449-12460, 2020.

[7] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451-3460, 2021.

[8] D. Kim and B. C. Song, "Contrastive adversarial learning for person-independent facial emotion recognition," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, pp. 5948-5956, 2021.

[9] Y. Qiang, P. Pan, C. Li, X. Li, R. Hong, and D. Zhu, "AttCAT: Explaining transformers via attentive class activation tokens," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, pp. 5052-5064.

[10] J. Wu, B. Duan, W. Kang, H. Tang, and Y. Yan, "Token transformation matters: Towards faithful post-hoc explanation for vision transformer," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10926-10935.