Indian Institute of Information Technology,
Design and Manufacturing, Kancheepuram

Meghana Venna
CS20B1060

**Internship Report**

# ACOUSTIC PATTERNS OF SPEECH AS A PREDICTOR OF ALZHEIMER'S DEMENTIA

## 1 Summary of Internship

### 1.1 Motivation

Alzheimer's disease (AD) is a neurodegenerative disease that is the most common form of dementia. It leads to a gradual decrease in cognitive functioning, including memory, language, and problem-solving skills. If left undiagnosed until the critical stage, Alzheimer's cannot be controlled and the severity of the disease increases. Every year, 50 million people are affected by dementia worldwide. Since there is no cure for Alzheimer's, early detection is critical to delay its progress.

Current methods for early detection of Alzheimer's are expensive and not scalable. For example, brain scans like MRIs can cost an average of 20,000 rupees per scan. This makes it difficult to screen large populations for Alzheimer's.

In the early stages, Alzheimer's affects the part of the brain that controls language and speech ability. This suggests that tasks that involve speech, such as picture description, question and answers, and word sequences, could be used to detect Alzheimer's. The patient's response to these tasks can help in Alzheimer's detection.

There is a famous test called the Boston aphasia examination that involves 5 primary such tasks. The goal of this project is to detect dementia based on the acoustic patterns in audio recordings of a set of patients, which involves both Alzheimer's and non-Alzheimer's patients. This is an affordable, non-invasive, and scalable solution with decent accuracy.

As part of my internship, I plan to use state-of-the-art deep learning techniques to classify these audio samples based on purely acoustic signal patterns. Deep learning techniques have been shown to be very effective in tasks such as image classification and natural language processing. It is hypothesized that these techniques inspired from image classification architectures can be used to develop a reliable and accurate method for detecting Alzheimer's.

### 1.2 Scope

#### 1.2.1 About the Dataset

The ADReSS challenge dataset consists of speech recordings and transcripts of spoken descriptions of the famous Cookie Theft picture. This test is used to assess the structural language and speech ability of patients. Some signs of Alzheimer's disease (AD) include requiring the interviewer's help, taking many pauses during description, forgetting the context in the middle of description, and mumbling while talking.

The dataset includes 86 AD subject samples and 78 normal subject samples. In some audio samples, the interviewer helps the patient continue the conversation. To focus only on the patient's audio, we manually cut out the interviewer's parts using Audacity software.

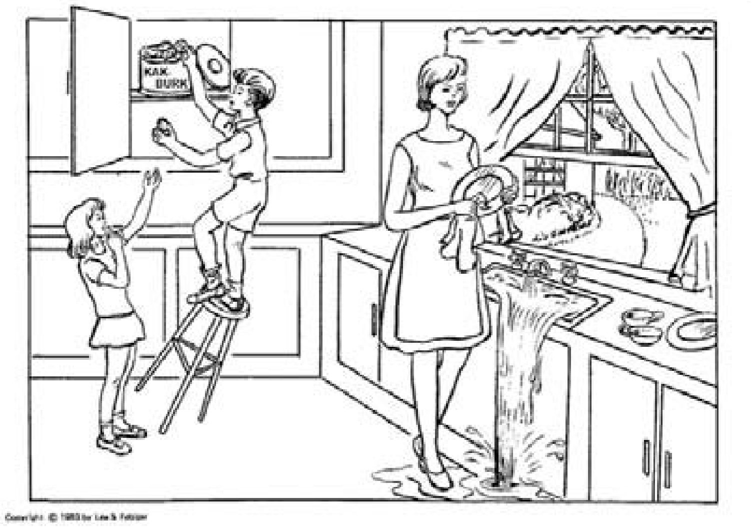


Figure 1: The Cookie Theft picture

The audio samples are loaded using the librosa library with a sampling frequency of 44 kHz. Some samples have a high amount of background noise, both stationary and non-stationary.The first level of noise reduction was done using open source libraries like noise reduce. This helped to remove the majority of stationary and moderate amounts of non-stationary noise. Further noise reduction was done manually using Audacity software.

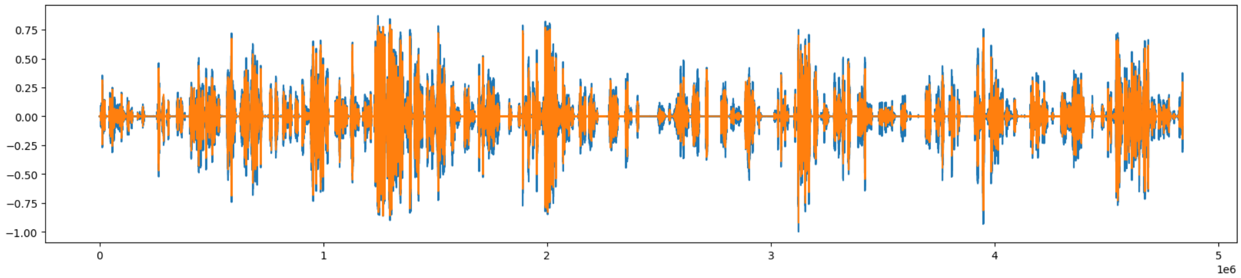


Figure 2: Removing Noise

To standardize the data, each sample was upsampled by breaking it down into one-second audio samples.

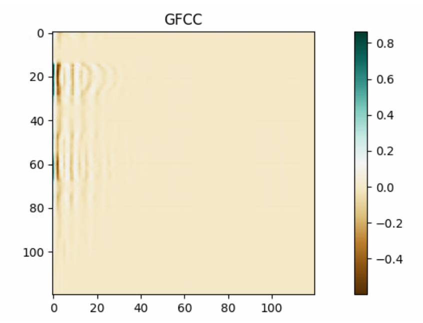
#### 1.2.2 Extraction of Features

To capture the distinctive acoustic patterns from audio signals, we can mimic the way the human ear perceives the audio input. This can be done using cepstral coefficients. The main types are MFCC (Mel frequency cepstral coefficients), GFCC (Gammatone frequency cepstral coefficients), and CQCC (Constant Q cepstral coefficients).

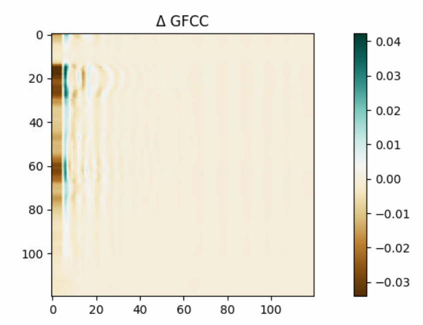
In previous research, MFCCs have been explored extensively. This time, I was interested in trying GFCCs. The idea is to convert audio signals into gammatone spectrograms. Gradient changes in spectrograms over time can indicate any useful patterns. Therefore, the first derivative (ΔGFCC) and second derivative (ΔΔGFCC) are extracted from GFCC.

GFCCs are extracted with a 167-milliseconds window length with 50% overlap(120x120 images). This window length is chosen to capture the short-term spectral characteristics of the audio signal. The 50% overlap between windows helps to ensure that the extracted features are smooth and continuous.

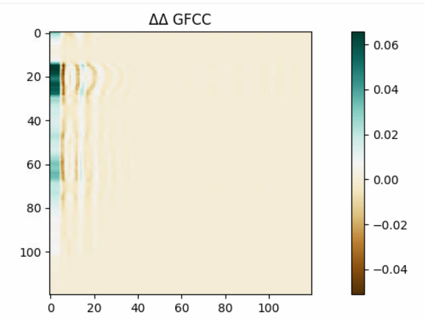
Three of the spectrograms stacked together can equivalently represent the RGB properties of an image. Through this similarity, we tried to exploit the state-of-the-art image classification architectures present in deep learning literature.



GFCC



ΔGFCC



ΔΔGFCC

### 1.2.3 Classification Models

After audio signals were converted into gammatone spectrograms, three types of features were extracted: GFCC, ΔGFCC, and ΔΔGFCC. These features were then stacked together to represent the RGB properties of an image.

The spectrograms were then fed into the custom-designed convolutional neural network, which was trained to classify the audio signals into two categories: Alzheimer's disease (AD) and normal.

The CNN was designed to achieve better accuracies with a relatively small number of parameters.

Focal Loss function along with Adam optimizer is used in CNN considering that there is imbalance in class samples.Focal loss is believed to perform better than general cross entropy loss. It helps models learn from difficult samples.The model is penalized more for making mistakes on difficult samples, which helps to improve the model's performance on these samples.

### 1.2.4 Results

The results of this study suggest that GFCCs are effective features for AD recognition problems when used with DNN models. A test accuracy of 67% was achieved. The train accuracy was 75%,suggesting that the model is well-trained and is not overfitting to the training data.

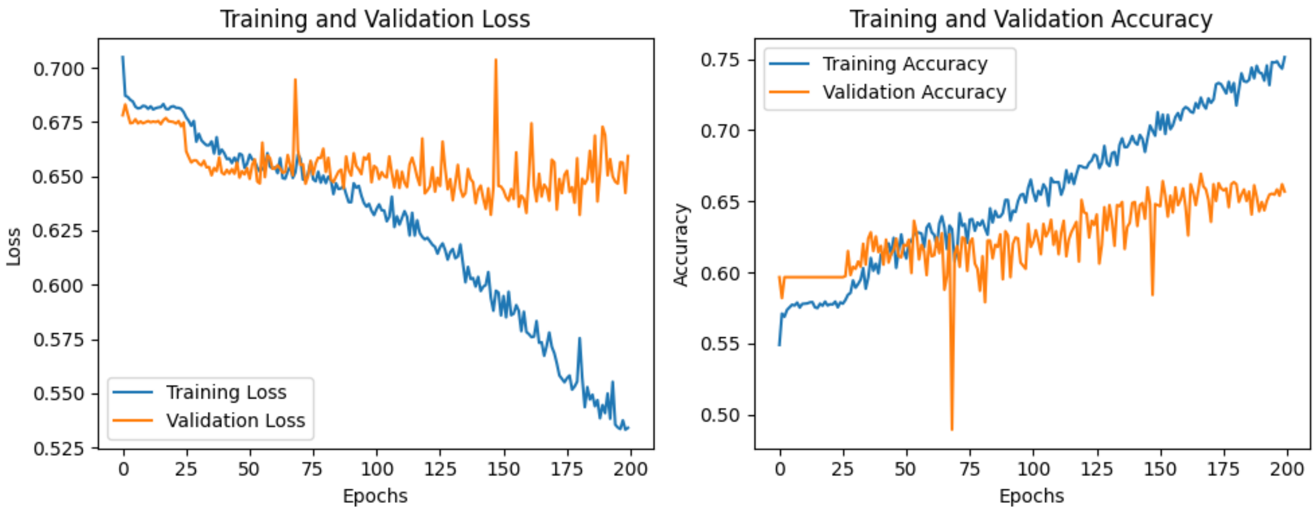


Figure 3: Training vs Validation

## 2 Conclusions

The results of this study suggest that GFCCs are effective features for AD recognition problems when used with DNN models. However, there is still room for improvement. The test accuracy of 67% is not yet high enough for clinical use. Future work will focus on improving the accuracy of the model, as well as developing a more efficient and scalable implementation.

## 3 Future scope

Some potential areas for future work include:

- Using Attention-based mechanisms: Attention mechanisms can be used to focus on specific parts of the audio signal that are more important for classification. This could be done by using a self-attention mechanism, which would allow the model to learn which parts of the audio signal are most relevant to the classification task.

- Building a custom-made loss function: The loss function is a critical part of any machine learning model, and it can have a significant impact on the model's performance. A custom-made loss function could be designed to better capture the nuances of the GFCC-based audio classification task.

- Employing EfficientNets: EfficientNets have shown state-of-the-art results on a variety of tasks, including image classification, object detection, and natural language processing. It would be interesting to explore the use of EfficientNets for GFCC-based audio classification. EfficientNets are known for their ability to scale effectively, and they could potentially be used to improve the accuracy and efficiency of audio classification models.

- Employing NFNet: NFNet is a recently proposed neural network architecture that has shown promising results on a variety of tasks. It could be interesting to explore the use of NFNet for GFCC-based audio classification.

## References

[1] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer's dementia recognition through spontaneous speech: The adress challenge,” 2020.

[2] A. Meghanani, A. C. S., and A. G. Ramakrishnan, “An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 670–677.

[3] A. Brock, S. De, S. L. Smith, and K. Simonyan, “High-performance large-scale image recognition without normalization,” *CoRR*, vol. abs/2102.06171, 2021. [Online]. Available: https://arxiv.org/abs/2102.06171

[4] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *CoRR*, vol. abs/1708.02002, 2017. [Online]. Available: http://arxiv.org/abs/1708.02002

[5] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, “Linguistic features identify alzheimer's disease in narrative speech,” *Journal of Alzheimer's Disease*, 2016.

[6] T. Sainburg, “timsainb/noisereduce: v1.0,” Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[7] A. Team, “Audacity,” https://www.audacityteam.org/.

[1] [2] [3] [4] [5] [6] [7]