Exploring Acoustic and Linguistic Features in Alzheimer's Dementia

Thesis submitted in partial
fulfillment of the requirements of the degree of

Master of Science
in
Electronics and Communication Engineering by Research

by

Nayan Anand Vats
2019702006
nayan.vats@research.iiit.ac.in

Advised by Dr. Anil Kumar Vuppala

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
April, 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Exploring Acoustic and Linguistic Features in Alzheimer's Dementia" by Nayan Anand Vats, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Prof. Anil Kumar Vuppala

To my parents, Uncle and Aunt

Acknowledgements

Firstly, I thank the Institution for providing the proper infrastructure and work culture for research. The massive responsibility that our Guides take to set up the pipeline for research work is exceptional. The flexibility to choose my problem statements with the feedback of the Guide and colleagues made the research work very exciting. I sincerely thank my Guide, Dr. Anil Kumar Vuppala, for supporting my research. His thoughtful criticism and suggestions helped me to set higher goals and achieve perfection at every step of my work. His strong work ethic and ideology to maintain originality in ideas and thoughts motivated me to approach the problem differently. He emphasized the importance of writing and reading research papers as a continuous process, which helped me learn the art of presenting my thoughts in writing very well. Even during the COVID pandemic, regular lab meetings and presentations arranged by him helped me brainstorm and transfer intensive knowledge to carry out my research. I thank my colleagues, Dr. Krishna Gurugubelli, Dr. Sai Ganesh Mirishkar, Purva Sharma, and Aditya Yadavalli. They gave time to share their expertise in speech signal processing, Natural Language Processing, Deep Learning, and Machine Learning. This helped me to address my research problem efficiently. I was also fortunate to attend the courses taught by Dr. Anil Kumar Vuppala, Dr. C. V. Jawahar, Dr. Santosh Nannuru, and Dr. Naresh Manwani. These courses helped strengthen my foundation in Speech Signal Processing and gave me the knowledge and confidence to apply Machine Learning and Deep Learning techniques to solve real-world problems. Last, but not least, I am forever grateful for conducting my research work in the IIITH Speech Processing Lab, which has a great legacy of work in Speech Signal Processing. I look forward to taking their legacy ahead by contributing my best in the field of Speech Signal Processing.

Abstract

Dementia is a chronic and progressive syndrome that affects the cognitive functioning of an individual. Alzheimer's, a neurodegenerative disorder, is the leading cause of dementia. The only way to control the progression of the disease is its early detection, followed by drug and non-drug interventions. The speech production chain is the presentation of cognitive abilities and is majorly affected in the early stages of Alzheimer's disease. Speech signals are ubiquitous and facilitate easy recording, storage, and transfer. For these reasons, researchers have long strived to develop complementary tools for Alzheimer's Dementia detection using acoustic and linguistic clues derived from speech utterances. This thesis is one more attempt at finding the differentiating auditory and linguistic clues in the utterance(audio and transcript) of an Alzheimer's Dementia patient. The ADReSS INTERSPEECH-2020 and ADReSSo INTERSPEECH-2021 challenges provide a balanced dataset for the Alzheimer's Dementia classification task. This thesis explores the efficacy of different acoustic features to capture distinct patterns in the speech utterance of AD patients. The features are obtained by evaluating Cepstral Coefficients over different acoustic algorithms and techniques. Mel-frequency and Linear Prediction methods are used to capture the Vocal tract characteristics; Residual Coefficient, Glottal Volume Velocity, and Zero Frequency Filtering approach for excitation source characteristics; Envelope Modulation Spectrum and Long Term Averaging Spectrum capture the prosody characteristics of speech utterance. The next part of this thesis explores the Single-frequency-filtering-based(SFF) high spectrotemporal resolution feature using the filter bank approach for Alzheimer's Dementia detection. Experiments are performed using different machine learning classifiers over the acoustic features extracted from the challenge datasets. The current study also demonstrates the performance of the BERT model for the dementia classification task. Finally, the performance of individual and combined acoustic features is reported. Also, the classification score is evaluated by score level fusion of the acoustic models and BERT model to observe the complementary characteristics of acoustic features to the BERT model. Acoustic models perform best when combined with linguistic features, suggesting the complementary nature of acoustic and linguistic features. Also, the high spectro-temporal resolution Single Frequency Filtering feature captures the characteristics of speech patterns better for Alzheimer's Dementia classification than traditional source-filter model-based features.

Keywords— Alzheimer's Dementia Classification, ADReSS INTERSPEECH-2020, ADReSSo INTERSPEECH-2021, Speech Production, Source-Filter Model, Mel-frequency, Linear Prediction, Residual Coefficient, Glottal Volume Velocity, Zero Frequency Filtering, Envelope Modulation Spectrum, Long Term Averaging Spectrum, BERT, Instantaneous frequency, Filter Ban Single Frequency Filtering

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Alzheimer's Dementia

Dementia is a term used to describe a group of symptoms affecting memory, thinking, and social abilities severely enough to interfere with daily life. Dementia is usually caused by damage to the brain cells by various diseases. Dementia can be of several types, namely Alzheimer's Dementia, Vascular Dementia, Lewy body dementia, Frontotemporal Dementia, and Mixed Dementia. Alzheimer's disease is the most common cause of Dementia and accounts for 60-80% of dementia cases. Alzheimer's Dementia is an irreversible, progressive, neurodegenerative disease primarily found in the elderly around 60 years and above[1]. Once diagnosed with the disease, the subject lives 4-5 years on average.

Alzheimer's causes a loss in brain neurons, particularly in the cortex. The changes are primarily in the part of the brain that affects learning. Microscopic changes start in the brain long before the symptoms appear. Although the cause of Alzheimer's disease is not entirely understood, two significant players in its progression are plaques and tangles. Amyloid plaques are chemically sticky substances and can accumulate and disrupt the signaling between the neurons. APP(Amyloid Precursor Protein) helps the cell repair and grow after injury. APP gets used, broken down, and recycled over time like other proteins. However, a set of chemical reactions creates an insoluble monomer Beta Amyloid. These monomers tend to be chemically sticky and bond together just outside the neuron and form beta-amyloid plaques(clumps of beta-amyloid); the plaques can get between neurons, disrupting the neuron signaling (memory impairment). These plaques can also cause an immune response that may cause inflammation and damage the surrounding neurons. Another major cause, the tangles, are found inside the cell as opposed to beta-amyloid plaques, as shown in Figure 1.1. A unique protein, namely Tau assures that the neurons are held together. The amyloid plaque causes the Tau protein to change shape, stop supporting the neurons, and clump up with other Tau proteins and get tangled, leading to apoptosis(programmed cell death). As neurons die, large-scale changes begin to take place in the brain, and the brain shrinks, as seen in Figure 1.2. This process is progressive and can occur over many years.

In the early stages of Alzheimer's, the symptoms might not be detectable. As the disease spreads through the brain, lack of orientation, attention, mood swings, and behavior changes follow. As the disease progresses, the patients might have short-term memory problems, loss of motor skills, language problems, cognitive skills, personality impairment, and decreased alertness and awareness of the surroundings. Finally, there is difficulty in speaking, swallowing, and even moving. Eventually, there is a loss of long-term memory and complete disorientation of physical and mental state. The subjects become suspicious about family and friends and lose their sense of time and location. Alzheimer's Dementia has a physical, social, and economic impact not only on people suffering from it but also on their caretakers, family, and society. Alzheimer's patients are often put on expensive,

inaccurate, and invasive medication facilities with frequent side effects. The day-to-day routine of Alzheimer's patients needs enormous support from caregivers, family, and society. The disease burdens caregivers, including social, psychological, physical, and economic aspects. As of 2020, 50 million people worldwide have Alzheimer's Dementia, and unlike the number one cause of death(heart disease), which has decreased by 7.3%, the number of deaths by Alzheimer's has more than doubled from 2000 to 2019. Study shows the staggering impact of the disease on the economy and healthcare system of the countries and the need to address it urgently[2].

Research Institutions are actively working towards prevention, early diagnosis, and disease progression monitoring[3]. Early diagnosis of the disease increases the effect of drug and non-drug interventions to delay or even prevent cognitive decline[1, 4]. Early detection of dementia is challenging due to the lack of reliable biomarkers, overlapping symptoms with normal aging, and low accuracy of existing cognitive screening tools. The primary requirement to curb Alzheimer's dementia growth rate is the availability of easy(non-invasive, low-cost, quick) diagnostic methods and tools that can detect the disease accurately at an early stage. Currently used MMSE(Mini-Mental State Examination) is a complex tool and does not have a substantial role as stand-alone evidence for Dementia Progression[5]. Brain scans such as CT scans, MRIs, and PET helps to observe the biological signs of the disease and are better predictors than MMSE. Still, they are invasive and expensive and cannot be used for frequent mass screening at a global scale[6]. The most common symptom of dementia is the disturbance in short-term and medium-term memory[7]; affected, among other things, are the patient's speech and language at the preclinical stage of the disease[8, 9, 10, 11, 12, 13, 14, 15, 16]. Dr. Alois Alzheimer's studied the first case of Alzheimer's Dementia[17] and gave clues about speech and language impairment. He provided details about memory disturbance, amnestic writing disorder, paraphrasic derailments(mispronunciation, slip of tongue, word substitution), and perseverations(repetition of words, ideas, or subjects) in spontaneous speech. Speech can serve as a complementary tool for Alzheimer's disease and can guide future interventions[18]. Speech production requires message formulation, language coding, neuro-muscular commands, and vocal-track response. The speech production chain is mapped to individual cognitive abilities. The urge or intention to speak requires the speaker's brain to form a sentence with the intended meaning and map the sequence of words into physiological movements needed to produce the corresponding sequence of speech sounds. Speech signals are time-variant signals and are a constant flux of information. It is ubiquitous, can be recorded seamlessly, and is easily transferable. These valuable characteristics of speech signals motivate us to look for acoustic clues for Alzheimer's Dementia in speech utterances/recordings of subjects. Also, speech recordings are accompanied by transcripts or can be transcribed using an Automatic Speech Recognition system(ASR) to capture the linguistic clues for Alzheimer's Dementia. It is noteworthy that the linguistic and acoustic clues are often complementary and can contribute to an excellent secondary diagnostic tool for Alzheimer's Dementia.

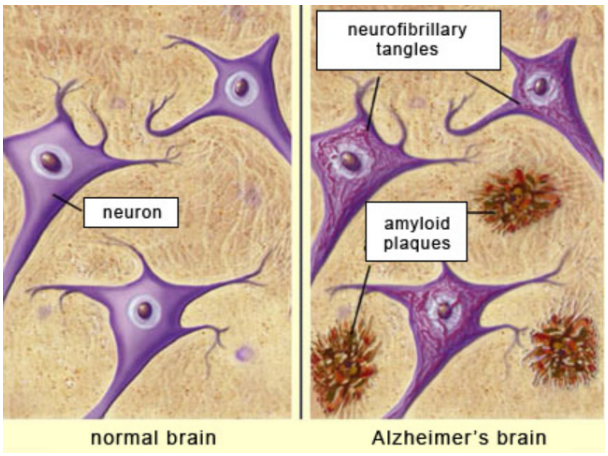

Figure 1.1: Plaque and tangles

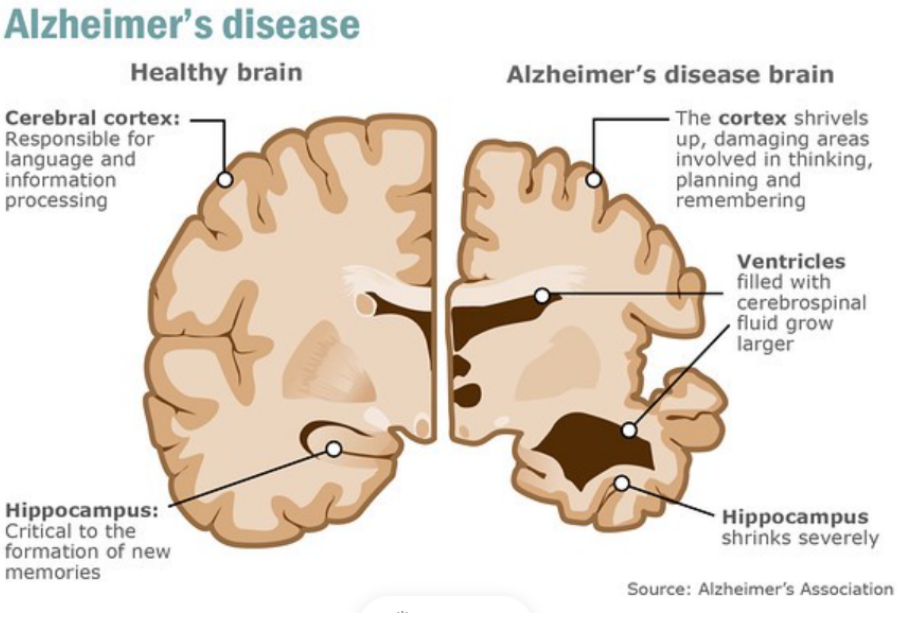

Figure 1.2: Brain Shrinkage Alzheimer's

1.2 Dataset

Studies for early Alzheimer's Dementia (AD) diagnosis using acoustic and linguistic approaches have been conducted for a long time. However, most of the studies lacked a standard dataset. The performance of different techniques for Alzheimer's Dementia detection is evaluated over various tasks like sentence repetition, picture description, conversation, dialogue, verbal fluency, backward counting, etc. In such a situation, comparing and setting state-of-the-art performance for Alzheimer's detection is tough. To take the research for Alzheimer's Dementia detection forward, the availability of standard datasets, well-defined tasks, and performance metrics is indispensable. For the above reasons, in this thesis, the experiments are conducted over two standard challenges for Alzheimer's Dementia detection(one follow-up of another) described as follows: -

- Alzheimer's Dementia Recognition through Spontaneous Speech,
  The **ADReSS** Challenge(INTERSPEECH-2020)[19]
- Alzheimer's Dementia Recognition through Spontaneous Speech Only,
  The **ADReSSo** Challenge(INTERSPEECH-2021)[20]

The ADReSS challenge dataset consists of speech recordings and transcripts of spoken picture descriptions elicited from the participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam. It consists of 54 AD (Alzheimer's Dementia) and 54 NON-AD subjects in the training dataset. The test dataset consists of 24 AD and 24 non-AD subjects. The dataset is balanced for age and gender to minimize the risk of bias in the prediction task.

The ADReSSo dataset is an improved form of the ADReSS dataset. However, the ADReSSo dataset is a speech only dataset and does not provide any transcriptions for the recording like the ADReSS dataset. It consists of a total of 237 audio files sampled at 44kHz. The audio files were split into the cross-validation dataset and test datasets. The cross-validation dataset contains 70% of the total dataset with 79 control(healthy) speakers and 87 speakers with AD. The test dataset contains 36 control speakers and 35 AD speakers.

The main objective of the ADReSS and ADReSSo challenge is to provide a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender. The challenge defines a shared task that can compare different approaches to AD recognition from spontaneous speech. The dataset is statistically balanced to mitigate biases such as repeated occurrences of speech from the same participant, variations in audio quality, and imbalances of gender and age distribution. The data is a set of audio recordings obtained from a picture description task, namely the "cookie theft picture description task." The endeavor of ADReSS and ADReSSo challenge is to give way to automatic assessment of Alzheimer's Dementia most accurately and cheaply. The data demography, along with the MMSE scores, is defined in table 1.1 and table 1.2.

Table 1.1: Train Data Demographic Distribution

	AD			non-AD		
Age-interval	M	F	MMSE(sd)	M	F	MMSE(sd)
[50,55)	1	0	30.0(n.a)	1	0	29.0(na)
[55,60)	5	4	16.3(4.9)	5	4	29.0(1.3)
[60,65)	3	6	18.3(6.1)	3	6	29.3(1.3)
[65,70)	6	10	16.9(5.8)	6	10	29.1(0.9)
[70,75)	6	8	15.8(4.5)	6	8	29.1(0.8)
[75,80)	3	2	17.2(5.4)	3	2	28.8(0.4)
Total	24	30	17.0(5.5)	24	30	29.1(1.0)

Table 1.2: Test Data Demographic Distribution

	AD			non-AD		
Age-interval	M	F	MMSE(sd)	M	F	MMSE(sd)
[50,55)	1	0	23.0(n.a)	1	0	28.0(n.a)
[55,60)	2	2	18.7(1.0)	2	2	28.5(1.2)
[60,65)	1	3	14.7(3.7)	1	3	28.7(0.9)
[65,70)	3	4	23.2(4.0)	3	4	29.4(0.7)
[70,75)	3	3	17.3(6.9)	3	3	28.0(2.4)
[75,80)	1	1	21.5(6.3)	1	1	30.0(0.0)
Total	11	13	19.5(5.3)	11	13	28.8(1.5)

1.3 Motivation

Alzheimer's Dementia is a neurodegenerative disease. Along with affecting an individual's cognitive abilities, it also affects the speech and language abilities of the subject. Speech can provide complementary clues for early disease detection, which otherwise requires a complex process(expensive and time-consuming process) to detect. Speech is a constant flux of information and provides tremendous information about an individual's cognitive state. Speech is relatively easy to elicit and has proven to be a valuable source of clinical data. It is also ubiquitous and can be seamlessly acquired. Though speech signal is enormous, extracting the correct information for a particular application is challenging. The manifestation of AD varies in different people depending on their age group and background. The process is progressive, and the symptom overlaps with normal aging. This makes it very difficult to point to the characteristics of the speech signal affected by the disease. Therefore, Alzheimer's dementia detection from spontaneous speech is a novel challenge with great potential for future research and development. This thesis contributes to AD detection by looking at various acoustic characteristics that can differentiate Alzheimer's Dementia patients from healthy subjects.

1.4 Objective and scope of thesis

Considerable research has shown the manifestation of Alzheimer's Dementia through speech and language. However, there is no firm evidence of any baseline feature in the speech signal symbolic of Alzheimer's Dementia. In such a situation, it is necessary to analyze the speech signal with different lenses to capture any possible markers in the utterances of Alzheimer's Dementia patients. Speech can be analyzed for segmental and supra-segmental characteristics, vocal track and vocal fold response, different time-frequency resolutions, etc. Fortunately, a strong legacy of speech-processing algorithms provides a variety of algorithms to look at speech signals with various lenses. It is useful to visualize speech signals as a decomposition of components of the speech production mechanism, as shown in figure3.1. The vocal tract response is explored with Mel frequency cepstral coefficients(MFCC), Linear prediction cepstral coefficients (LPCC), and Statistical features(stat feat). The excitation source is explored through the Glottal Volume Velocity Cepstral Coefficients(GVVCC), Zero Frequency Filtering Cepstral Coefficients(ZFFCC), and Residual Cepstral Coefficients. For prosody, Envelope Modulation Spectrum(EMS), Long Term Averaging Spectrum(LTAS), and the very prominent openSMILE feature comParE are used. The thesis also explores a high spectro-temporal resolution Single Frequency Filtering approach(SFF). SFF gets away with the windowing effect and captures the co-articulation characteristics in the speech signal more appropriately. Acoustic features are used to extract information regarding how things are uttered and the characteristics of the speech production system affected the most in Alzheimer's. A brief description of these acoustic features is given in table 1.3.

The other alternate way to capture distinctive characteristics of Alzheimer's Dementia patients is through transcripts. The ADReSS-2020[19] challenge provides the manual transcript for all recording instance. Transcripts consist of lexical, semantic, and syntactic meanings. Alzheimer's dementia patients are often characterized by a lack of semantic and syntactic clarity in their utterances. Using transcripts allows us to differentiate AD patients based on the content. Transfer Learning has been the key for classification tasks with a smaller dataset. A pre-trained model understanding natural language can be fine-tuned for AD classification tasks very well. This thesis uses the Bi-directional Encoder Representation from Transformer(BERT) to classify Alzheimer's Dementia. BERT outperforms the acoustic models and captures complementary characteristics of AD patients.

Table 1.3: Acoustic feature sets used for the detection of Alzheimer's Dementia from speech

Feature set label	Features set name	Source of Information	Dimension
S1	MFCC_Stat	vocal tract system	156
S2	LPCC_Stat	vocal tract system	144
S3	ResCC_Stat	Excitation source	156
S4	GVVCC_Stat	Excitation source	156
S5	ZFFCC_Stat	Excitation source	156
S6	EMS_Feat	Prosody	48
S7	LTAS_Feat	Prosody	99
S8	STAT_feat	vocal tract system	156
S9	comParE	Prosody	6373
S10	SFFCC	High-spectro-temporal	6373

The objective of this thesis is to contribute to the Nobel initiative of automatic Alzheimer's detection. This work uses the ADReSS-2020[19] and ADReSSo-2021[21] dataset for the experiments. The goal is to incorporate automatic tools into the diagnostic pathway in current memory services and provide low-cost, easy treatment.

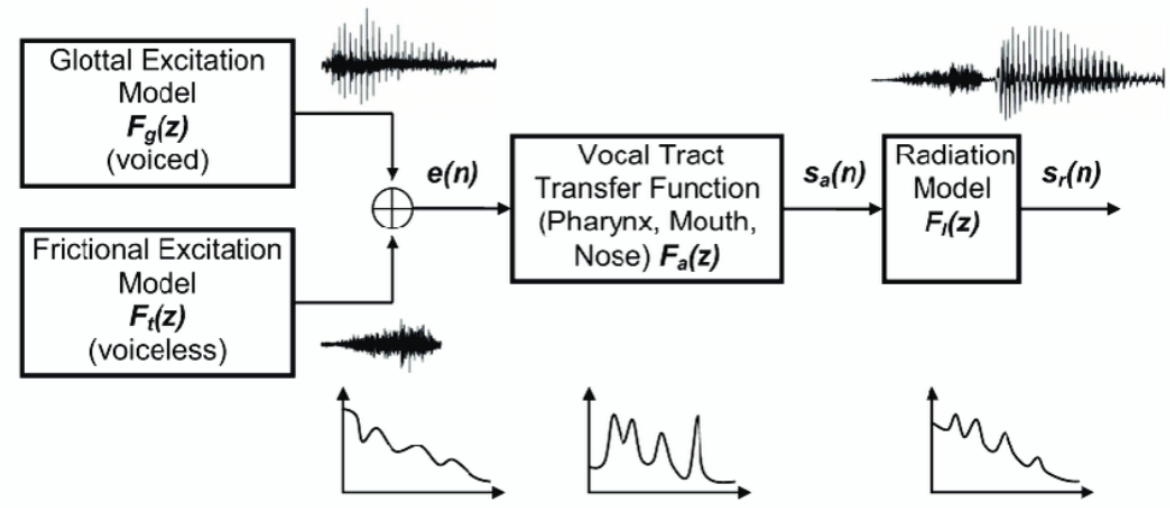

Figure 1.3: Source-filter model

1.5 Organization of thesis

The rest of the thesis is organized as follows: -

- In chapter 2, we do the literature survey giving an overview of all the effort put into classifying Alzheimer's Dementia from speech and text data. The literature survey gives a perspective of the work done and the way to conduct research in this direction.
- Chapter 3 evaluates the performance of acoustic algorithms and the BERT model on the ADReSS-2020[19] dataset for Alzheimer's Classification. It specifically assesses algorithms related to speech production and speech prosody. The complementary nature of acoustic features to linguistic features is discussed by fine-tuning the BERT model on the manual transcripts provided with the dataset.
- Chapter 4 discusses the importance of improved time-frequency resolution in differentiating Alzheimer's Dementia from healthy speech. Single-frequency Filtering(SFF) is used to obtain a better time-frequency resolution over the speech utterance.
- Chapter 5 concludes this thesis and puts light on the future work that can be done to improve the detection of Alzheimer's Dementia from spontaneous speech.
- Chapter 6 enumerates the publication the thesis is based on.

Chapter 2

Literature Survey

In this chapter, we look at all the development that has taken place in Alzheimer's dementia detection. In most cases, AD detection is performed either using audio recordings or with the help of transcripts generated from audio recordings. The latest development uses pre-trained models and deep embeddings for AD classification.

2.1 Speech and Alzheimer's

The most common symptom of dementia is the disturbance in short-term and medium-term memory[7]; along with few other symptoms, patient's speech and language are affected at the preclinical stage of the disease as discussed in [8, 9, 10, 11, 12, 13, 14]. In [22], authors use dialogue interaction data to look at speech silence patterns and other prosodic features. The paper evaluates content-free features such as speech rate, turn-taking pattern, and other speech parameters to identify Alzheimer's Dementia from spontaneous speech. The best performance is obtained for Additive Logistic Regression. In [23] state-of-the-art paralinguistic features, namely eGeMAPS, emobase (Emotion Baseline,openSMILE), ComParE 2013 feature set, and new Multi-Resolution Cochleagram (MRCG) features(openSMILE v2.1 toolkit) are evaluated over the voiced segments obtain from speech utterance from 82 AD and AD 82 non-AD subjects. The paper introduces a new Active Data Representation(ADR) technique to get utterance representation from segmental features. ADR is a set of statistical parameters obtained ahead of SOM(self-organised maps) clusters.

2.2 Linguistic and Alzheimer's

In [24], the author performs a detailed analysis of lexical features, namely N-rate, P-rate, A-rate, and V-rate (rate for part-of-speech (POS) category), TTR, Brunet's index(W) and Clause-like semantic unit(CSU) to evaluate the severity of DAT using machine learning techniques. The experiments found that closed-class words were particularly helpful in predicting the level of language deficit in patients. In [25], authors use conversational data to obtain word embeddings using 'w2vec' and 'GloVe.' 'GloVe' and 'w2vec' are known to capture the semantic information. Embedding for individual words is combined in four different techniques to classify dementia using conversational transcripts for each session with logistic regression. The classification performance is compared for manual and ASR-generated transcripts using the Kaldi toolkit. In [26], the author performs exploratory factor over the top 50 features selected out of 370 features using Pearson correlation coefficients. The feature vector mainly comprises linguistic features relating to POS, syntactic complexity, grammatical constituents, psycholinguistics, vocabulary richness, information content, and repetitiveness. Factor analysis with 4 factors reveals

that the relatively large set of linguistic measures can be mapped to four latent variables, broadly representing syntax/fluency, semantics, acoustic differences, and other information content.

2.3 Acoustic and Linguistic in Alzheimer's

In [27], authors evaluate low-level prosodic, voice quality, and spectral descriptors with the help of the openSMILE and COVAREP toolkit. Experiments were also performed with ComParE, IS10-Paraling, and VGGish features. Fisher vector embedding, BoAW(Bag of Acoustic Words), and statistical functionals were evaluated over the low-level acoustic descriptors to obtain global representation for each audio recording. Since the dataset [28] provided both audio recordings and corresponding transcripts. The paper explores text modality with BERT and its variants as well. Embeddings for each word in a transcript use min/max/Rang/StdDevNorm pooling to get the global representation of a transcript. Performances are evaluated for individual features and by combining the top-performing features.

2.4 Language Models and Alzheimer's Dementia

In [29] the author tries to capture the language characteristic in Alzheimer's using pre-trained Language Models like BERT and ERNIE. The Language Models are fine-tuned with pause encoded transcripts from the cookie theft dataset [28] for Alzheimer's classification. Ensemble method is used over multiple fine-tuned Language Models to avoid overfitting.

2.5 Latest Advancements

Study shows phonological and articulatory impairment in Alzheimer's at presentation or in the early stages of the disease [30]. In [31], hesitation ratio outperforms articulation rate, speech tempo, and rate of grammatical errors to capture the characteristic differences between AD and control speech. In[32] frequency and duration of voice-breaks, shimmer(amplitude perturbation quotient) and noise-to-harmonic ratio characterises people with dementia. In [33, 34] voiced, unvoiced and pause information is explored for AD detection. A total of 13 prosody features relating to Voice Activity, Articulation and speech rate are explored in[35] for AD detection. [36] applies features selection(Weka attribute selection function) over three groups of features relating to voice quality, speech and silence, and spectral attributes over the DementiaBank speech recordings. Feature selection improves the classification accuracy and prevents overfitting the data. [37] also introduces a two-stage wrapper feature selection method using a backward feature elimination performed if there is an increase in the total accuracy in the AD classification task. [38] measures alteration in rhythm of the utterance of subjects using syllabic variability for AD detection. A larger feature set is used in [36], that explores a total of 263 features relating to pitch, voice breaks, voice quality, jitter, shimmer, duration statistics, pause and MFCC for AD detection. Low-level paralinguistic features and functionals including timing and duration of vocalisations and pauses, speaking rate, and voice quality measures are evaluated in [39] over the Pitt dataset[40] for AD classification task. In [41] 22 metric were evaluated over speech and pause patterns for AD classification. Most of the above approaches used handcrafted features and traditional classification algorithms.

Latest studies see the use of PittDataset[19], balanced in terms of age and gender; the dataset provides speech recording and manual transcription for the cookie-theft picture description task. Most of the work presents the dominance of linguistic features and models over the acoustic features for the AD classification task. In[42]

a hierarchical neural network with an attention mechanism trained on linguistic features performs better than three acoustic-based systems, namely Bag-of-Audio-Words (BoAW) quantizing different low-level descriptors, Siamese Network trained on log-Mel spectrograms, and Convolutional Neural Network trained on raw waveforms. [43] performs grid search over the combination of Audio-feature(MFCC, eGeMAPS, Average duration), TF-IDF features, and several doc2vec embedding representations before passing them to the ML models. A multi-modal(audio, text) fusion approach using bi-LSTM is explored for the individual and combination of Acoustic(COVAREP), Lexical and disfluency features in [44]. [45] explores a multi-scale(word, phoneme over the text data) along with multi-modal approach for AD detection. Text data is analyzed both at word level and phoneme level and the best performance is achieved through the text classifier using phoneme representation with FastText phoneme embedding. The paper concludes that subword information, particularly phoneme representation, can be useful in cases of data scarcity. [46] makes the use of i-vector and x-vector as pre-trained acoustic features and pre-trained BERT model for textual feature embeddings. The performance of individual and combined features is evaluated using an SVM classifier. [47] trains three different MLPs from scratch with disfluency, acoustic(ComParE 2013), and intervention features as inputs for the AD classification task. The paper also uses transfer learning to leverage the features learnt in classification for the regression task. [48] demonstrates the superiority of features from pretrained model(VGGish) fol small number of audio recordings in the PittDataset[19]. The papers uses a modified Convolutional Recurrent Neural Network(CRNN) with three different acoustic features(ComParE, eGeMAPS, VGGish) and four different textual features that includes pretrained Language model features namely RoBERTa, Transformer-XL, GPT and a set of handcrafted features.

The latest PittDataset[21] is a speech-only dataset(no manual transcripts) and requires Alzheimer's Dementia detection straight from the speech signals. The dataset is well filtered to avoid repeated speech occurrences from the same participant (typical in longitudinal datasets), variations in audio quality, and gender and age imbalance. Most of the research using the PittDataset, revolves around using the transcripts generated from the ASRs or obtaining embedding from the speech signal directly. [49] uses three different pre-trained embeddings(trill, allosaurus, and wav2vec 2.0) and low-level descriptors of the eGeMAPS v2.0 as an input to a simple 1D Convolutional Neural Network based model. The highest dev and test accuracy of 75.3% and 78.9% respectively is obtained for wav2vec2 feature. [50] explore two state-of-the-art ASR paradigms, Wav2vec2.0 (for transcription and embedded acoustic feature extraction) and time delay neural networks (TDNN). The test results on best acoustic-only and best linguistic only are 74.65% and 84.51% respectively. [51] makes use of conventional acoustic features, pre-trained deep features(wav2vec 2.0) and their combination as input to LR, SVM, DT and NN. The highest precision is obtained with conventional feature whereas the highest accuracy of 67.61% is obtained for the SVM-combo. [52] makes use Deep textual embedding(DTE) and handcrafted features. For the purpose of DTE 9 different transformer based models including BERT and variants are used. For the handcrafted feature syntactic, readability, and lexical diversity features are explored. [53] explores several acoustic and linguistic models for AD detection. The acoustic models used are x-vector model, encoder-decoder ASR embeddings, prosodic features(total speech time, total pause time, percentage pause time, speech pause time, mean pause duration, and pause variability), VGGish and eGeMAPS feature. Linguistic models were developed with BERTs trained on different ASR transcripts. The best performance for acoustic model is obtained for The x-vector model and encoder-decoder automatic speech recognition embeddings. Bert with the commercial ASR transcripts provided the best result for linguistic models. [20] used Logistic regression over paralinguistic acoustic features namely MFCC, GeMAPS, eGeMAPS, ComParE-2016 and IS10-Paraling. The paper also extracts three different linguistic features over two different ASR generated transcripts, namely BERT-word, BERT-sentence and Linguistic Inquiry and Word Count (LIWC) feature. Different fusion modalities are explored using ensemble methods. The

best accuracy of 81.69% is obtained by fusion of linguistic and acoustic features. [54] uses Kaldi and OpenSMILE to obtain four different acoustic features that are Emobase, IS10, VGGish and X-Vector. CLAN and NLTK are used to obtain fifty linguistic features and sentence embeddings are obtained using Universal Sentence Embedding(USE) tool. The main contribution of the paper is to model from scratch four modular multi-modal architectures using Multi Headed Attention(MHA) and CNN. The unified model with Linguistic and X-Vector features achieved the best performance in respect to the accuracy(77.2%), precision(78.7%), and F1(76.3%) score.

As discussed above, a lot of active research has taken place for detecting Alzheimer's Dementia from speech and text. However, most techniques are either very basic or advanced techniques. Speech features related to speech production, prosody, and filter-bank features are not explored. This motivates us to use acoustic features related to speech production, prosody, and filter-bank techniques. Using these features can help us pinpoint the variation in speech patterns of Alzheimer's Dementia patients. The manual transcript provided with the dataset allows us to evaluate the BERT model and the complementary nature of the acoustic model to linguistic models for Alzheimer's classification. Finally, we demonstrate the superiority of high spectro-temporal feature based on Single Frequency Filtering for Alzheimer's Dementia classification.

Chapter 3

Acoustic features, BERT model, and their complementary nature for Alzheimer's detection

3.1 Motivation

Speech analysis involves extracting relevant features from a speech signal to suit a specific use case. Feature extraction methods in speech processing fall into categories like filter-bank techniques, source-filter model-based approaches, and statistical features. Raw speech signals have high redundancy and variability, making it essential to reduce both for efficient processing and interpretation. Different acoustic features extracted from speech data provide machine learning and deep learning models with distinctive and minimally redundant characteristics, improving classification accuracy. Our first work takes up the ADReSS INTERSPEECH-2020 challenge, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge[19]". The dataset provides both audio recordings and manual transcripts of the subjects. In many acoustic approaches for detecting dementia, state-of-the-art openSMILE and Mel Frequency Cepstral Coefficients (MFCC) features are explored [36, 19, 27]. However, speech features such as Linear Prediction Cepstral Coefficients (LPCC) and Filter Bank features still need to be explored. Furthermore, the excitation source features of speech that can capture intonation, speech rate, and pause information are not analyzed to detect dementia. Hence, this study aimed to look at the state-of-the-art excitation source performance and filter bank features to detect Alzheimer's. This work uses nine different acoustic features. These features can help us find the attributes in the human speech production system affected by Alzheimer's dementia. The work explores distinct vocal tract, excitation sources, and prosodic and statistical features.

Given the success of linguistic features in prior studies on Alzheimer's dementia classification, this study makes use of the manual transcript provided with the recordings to showcase the effectiveness of the BERT model in the dementia classification task. This work highlights the complementary nature of acoustic models to the BERT model in Alzheimer's detection.

The following sections briefly describe the acoustic features and the BERT model for Alzheimer's dementia detection.

3.2 Acoustic features for Alzheimer's detection

The features used in our first work revolve around the source-filter model of the speech production system. Speech is composed of a sequence of sounds (phonemes). The individual sound is produced by the combined action of the vocal tract and vocal cords, as shown in figure 3.1. Air thrust from the lungs goes through the closed

Table 3.1: Acoustic feature used for the detection of Alzheimer's dementia from speech

Feature	**Source of Information**	**Dimension**
MFCC_stat (S1)	Vocal Tract System	156
LPCC_stat (S2)	Vocal Tract System	144
ResCC_stat (S3)	Excitation Source	156
GVVCC_stat (S4)	Excitation Source	156
ZFFCC_stat (S5)	Excitation Source	156
EMS_feat (S6)	Prosody	48
LTAS_feat (S7)	Prosody	99
STAT_feat (S8)	Vocal Tract System	156
comParE (S9)	Prosody	6373

vocal folds (voiced speech). The vibration of the vocal folds creates impulsive air pressure (glottal pulses). The glottal pulses act as the source in the source-filter model. The fundamental frequency of the glottal pulses (impulse train) gives the perception of the pitch of the utterance and accounts for the source characteristic of the speech signal. Glottal vibrations from the vocal folds pass through the vocal tract. The vocal tract, composed of the larynx, pharynx, tongue, mouth cavity, and nasal cavity, can take up different shapes, generating different sounds. The effect caused by various vocal tract shapes can be mimicked with digital filters with varying coefficients, as seen by a simplified figure 3.2. The rate at which the shape and, hence, the filter coefficient change is around 25ms. Various acoustic algorithms try to capture the vocal track or source properties during this interval of 25ms. The assumption is that the speech signal is stationary, and the shape of the vocal tract doesn't change in this interval. A list of all the features used in our work is given in 3.1.

The vocal tract response is characterized by amplitude peaks (resonance) at different frequencies called the formats. The shape of the formats is characteristic of the sound produced, whereas the fundamental frequency of the glottal pulse gives the perception of pitch. The speech production system can be represented by a set of equations in 3.1 and 3.4 in the time and frequency domain, respectively. Here s(n), the speech signal is produced by a convolution of vocal tract response h[n] with the source u[n]. Equation 3.3 represents the vocal tract transfer function H(z) by an all-pole model. The raw speech waveform is a convolution of the vocal tract (filter) and excitation source (source) response, as seen above. In most use cases, it is helpful to suppress one of the characteristics, keeping the other. The classical technique of separating the vocal tract response and the source response is through cepstral analysis. The cepstral analysis, which is a fundamental tool for feature extraction and other acoustic methods used in this work, is discussed as follows:-

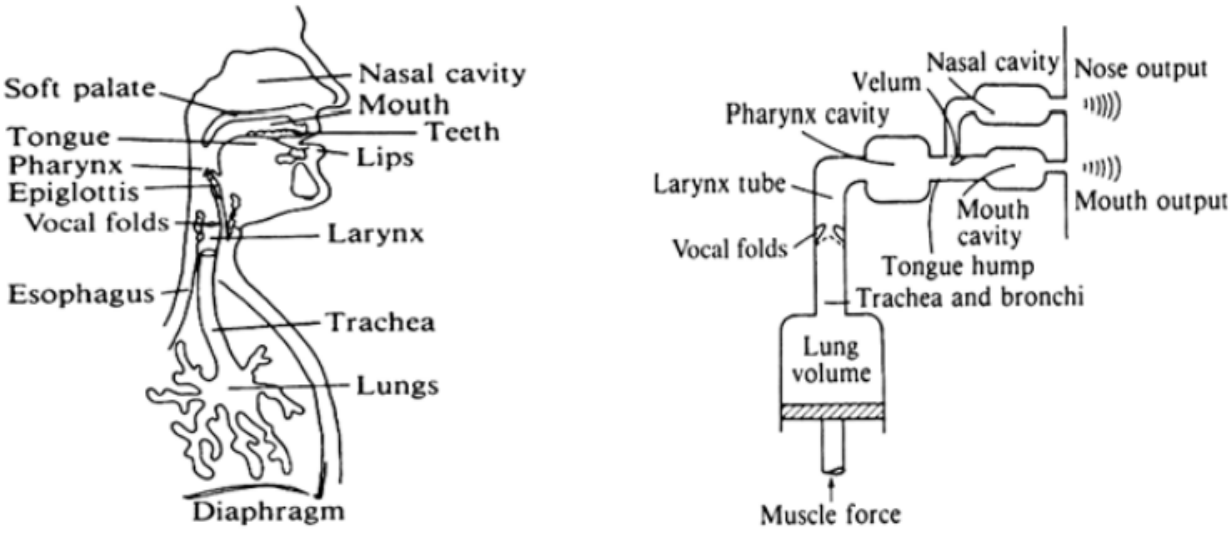

Figure 3.1: Speech Production Schematic

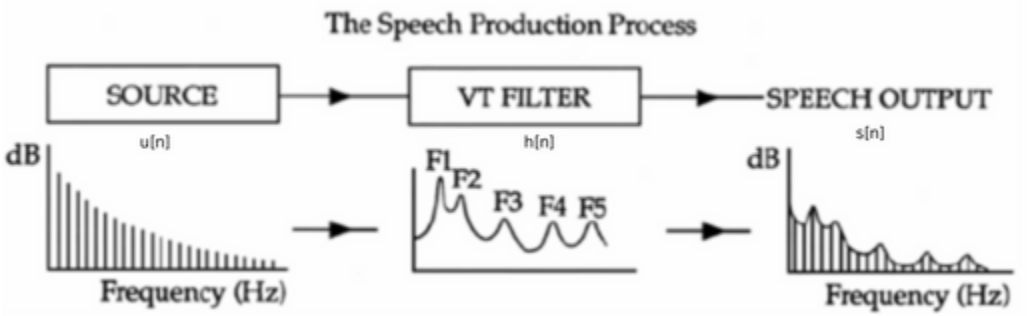

Figure 3.2: Source Filter Model

$$\hat{s}(n) = u[n] * h[n] \tag{3.1}$$

$$\hat{S}(f) = U(f).H(f) \tag{3.2}$$

$$H(z) = \frac{S(z)}{U(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_k.z^{-}k} \tag{3.3}$$

3.2.1 Cepstral Analysis

Cepstral Analysis is a mathematical tool to separate the vocal-tract properties and the source properties in the speech signal. Speech signals can be mathematically formulated as a convolution of the source and filter. To extract the filter characteristic from the speech signal directly is difficult. The focus of cepstral analysis is to obtain the coefficients representative of filter characteristics H[k] (vocal tract transfer function). The log spectrogram converts the convolution of the source and filter into summation as seen in equation 3.5. Taking logarithm over the spectrogram allows compressing the dynamic range of FFT and bringing out the slow varying vocal tract characteristics seen by the red curve in figure 3.4 . Assuming the log-spectrum to be a waveform, Fourier analysis over it would give a series of coefficients. The lower coefficients will represent the slow-varying part of the log spectrogram. These coefficients capture the vocal tract characteristics and can be used for further analysis and calculations. These coefficients are termed Cepstral Coefficient and can be obtained as shown in the block diagram 3.3. Cepstral analysis is performed after applying acoustic algorithms to the speech signal. This allows us to obtain a better representation of vocal tract characteristics. Cepstral coefficients are largely uncorrelated, allowing for efficient statistical modeling of features. Throughout this thesis, we have used the cepstral analysis for compact feature representation.

$$\hat{S}(k) = U(k).H(k) \tag{3.4}$$

$$log(\hat{S}(k)) = log(U(k)) + log(H(k)) \tag{3.5}$$

After Inverse FFT

$$C[n] = C_u[n] + C_v[n] \tag{3.6}$$

Figure 3.3: Cepstral Coefficient Extraction

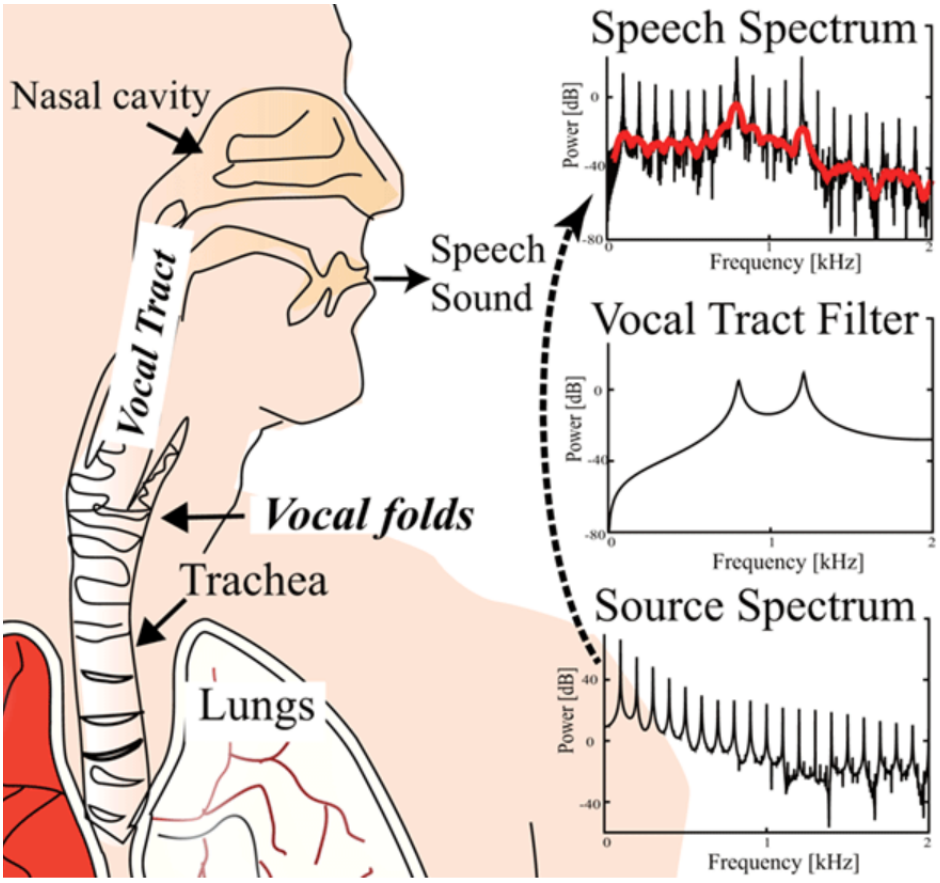

Figure 3.4: Speech source-filter decomposition

3.2.2 Linear prediction cepstral coefficients (LPCC)

As discussed in the section3.2, speech signal can be decomposed into source and filter response. The glottal excitation and vocal tract behavior can be mapped to the frequency domain. The aim of the Linear Prediction Coefficient is to capture the information in the spectral envelope from the FFT of the speech signal. The spectral envelope, as seen by the red curve in figure above, emphasizes the spectral peak in the speech segment. These peaks, also known as formats, are the resonant frequency of the vocal tract. Different overall resonant frequency shapes (formant structure) produce different sounds and can differentiate one sound from another. Linear Predictive coefficient takes a reverse filtering approach[55] to find these spectral peaks corresponding to the speech utterance. As the name suggests, it takes the linear combination of past speech samples to predict the current sample, as seen in equation 3.7. The predictor coefficients that are used for linear combination a_k are obtained by minimizing the mean of the square error function, that is, the error between the predicted sample value from the original sample, shown in equation 3.8. Taking speech samples of finite length and predicting the current sample from the past samples recursively, a set of equations is obtained. These equations are solved by using the Autocorrelation method or Durbin's Algorithm to get the LP coefficients for the speech segment.

The LP coefficients represent the vocal tract transfer function and hence the shape of the vocal tract over the speech segment. It is customary to obtain cepstrum from the LP coefficient to use the Linear Predictive Cepstral Coefficients (LPCC) as a feature for speech tasks . Once the cepstral analysis is performed over the

LP coefficients, liftering the first 13 Linear Predictive Cepstral coefficients can be used as the feature for speech processing tasks. The block diagram of LPCC implementation can be seen in 3.5.

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{3.7}$$

$$e(n) = \sum_{n=1}^{N} (s(n) - \hat{s}(n))^2 \tag{3.8}$$

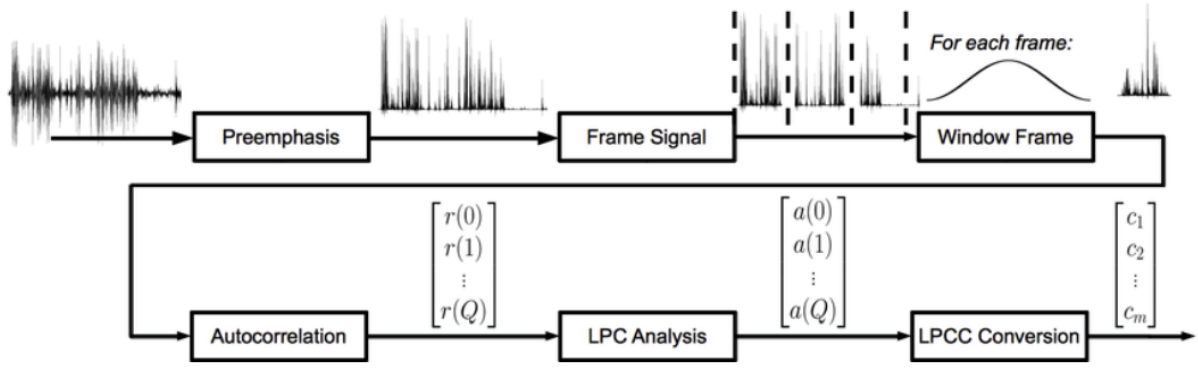

Figure 3.5: LPCC Feature Extraction

The approximation of formant structure by the LP analysis depends on the number of Linear Prediction Cepstral Coefficients that we choose, as shown in figure 3.6.

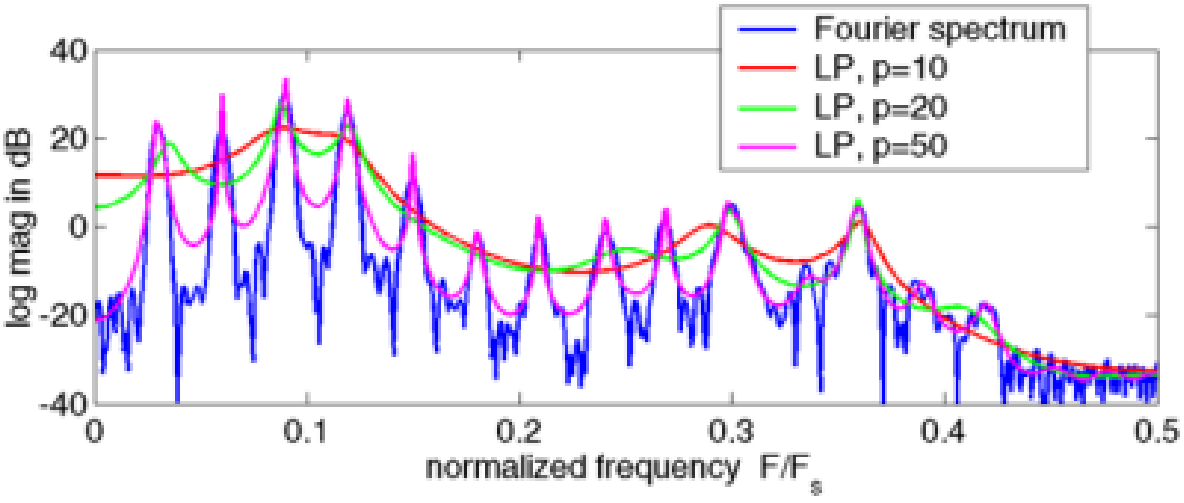

Figure 3.6: Formant approximation vs No of Cepstral Coefficient

3.2.3 Residual Cepstral Coefficients (ResCC)

A by-product of the LP analysis is the generation of prediction residues, or prediction errors e(n)[55]. The current speech sample can be predicted from the past speech sample only if there is no Excitation. Whenever there is excitation in the presence of a pitch pulse, the prediction goes wrong, and hence error prediction is an indicator of excitation. This pitch period can be determined by positions of the samples of **e(n)**, which are large, and defining the period as the difference between a pair of samples of **e(n)**, which exceeds a reasonable threshold.

3.2.4 Mel-frequency cepstral coefficients (MFCC)

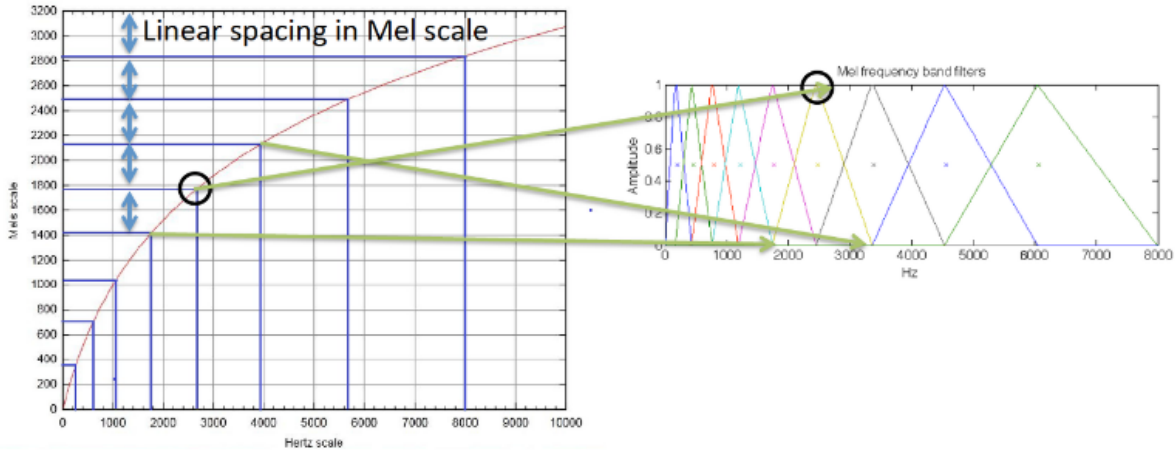

Figure 3.7: Mel Scale Filter Bank

The Mel Coefficients are obtained by averaging the spectrogram with a mel-filter bank[56]. The Mel-filter bank is based on the Mel-scale as seen in figure 3.7 Mel-scale mimics the human perception of auditory frequency. The Mel-coefficients are obtained by averaging the spectrogram with each filter in the Mel-filter bank(one coefficient for each filter). Once the Mel-coefficients are obtained, cepstral analysis is performed to obtain the MFCC. Cepstral analysis helps to get a better approximation of vocal tract information. The whole process of obtaining MFCC from the speech frames can be seen in the block diagram 3.7.

3.2.5 Glottal Volume Velocity (GVV)

Glottal Volume Velocity (GVV) refers to the rate of airflow passing through the glottis during vocalization. The glottis is the space between the vocal cords in the larynx. When we speak or sing, the vocal cords vibrate and open and close the glottis, allowing air to pass through in a controlled manner. The rate at which this airflow occurs is the glottal volume velocity, and it plays a crucial role in determining the voice's fundamental frequency (pitch) and the quality of the voice produced.

Extraction of Glottal Volume Velocity (GVV) can be done through inverse filtering[57]. It involves a multi-step process to isolate the fundamental source signal responsible for voice production in speech. It begins with segmenting the speech signal into short frames, estimating the vocal tract filter to remove its effects from the spectrum, and performing inverse filtering on each frame to isolate the GVV waveform. This GVV waveform represents the volume velocity of air passing through the glottis during speech production. Knowledge of glottal volume velocity and epoch locations helps obtain voice quality parameters such as pitch, glottal quotient, harmonic-to-noise ratio, jitter, and shimmer parameters.

3.2.6 Zero Frequency Filtering (ZFF)

Zero Frequency Filtering focuses on the source characteristic speech production system. It is based on the assumption that speech is produced by the convolution of impulse-like excitation with the all-pole filter vocal tract System. Unlike other speech algorithms that suppress the VT characteristics to bring out the Excitation source characteristic, ZFF evaluates the epochs directly from the speech signal[58]. Epochs (significant excitation in vocal tract through impulse) cause discontinuities in the speech signal in the time domain. Observing the discontinuity in the time domain is difficult due to the varying characteristics of the VT system. To monitor the epoch, the speech signal is passed through a stable zero-frequency resonator[59]. Deviation in the resonant frequency at the output of the resonator signifies the occurrence of Epochs. The zero-frequency resonators ensure high precision in the prediction of epoch location, as interference from vocal tract characteristics is negligible around zero frequency. The knowledge of epoch location can then be used to extract 76 dimensional intonation-related features as shown in table3.2.

Table 3.2: Details of the ZFF-based feature representation

Feature Name	Dimension
Statistical measures of F0	5
Jitter quotients of F0	22
Shimmer quotients of Strength of Excitation	22
Shimmer quotients of Energy of Excitation	22
F0 dispersion	1
Harmonic to noise ratio measure	4

3.2.7 Envelope Modulation Spectrum (EMS)

The Envelope Modulation Spectrum (EMS) represents the slow amplitude modulations in a signal and the distribution of energy in the amplitude fluctuations across designated frequencies, collapsed over time[60]. It has

been shown to be a useful indicator of atypical rhythm patterns in pathological speech. To calculate the EMS, the original speech signal is filtered into 9-octave bands with center frequencies of approximately 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680 Hz. Once the signal is filtered through the nine filters. An envelope is calculated for each filter output using the Analytical signal technique. From each of the 10 envelopes (one from the original and 9 from the octaves), 6 different parameters, including peak frequency and energy in various frequency bands, are calculated. The process results in a 60-dimensional feature for each speech utterance.

3.2.8 Long Term Averaging Spectrum (LTAS)

Long-Term Averaging Spectrum (LTAS) is the spectral analysis of a signal over an extended period; It averages the spectral characteristics to capture the long-term trend of a signal's frequency content. In LTAS, a signal is divided into overlapping windows, and FFT is applied to each. The spectra for each window are then averaged over time to produce a long-term average spectrum as in [61]. The choice of window size, overlap, and other parameters can influence the characteristics of the LTAS and should be chosen carefully based on the specific analysis requirements.

3.3 BERT: A language modelling approach for dementia detection

BERT stands for Bidirectional Encoder Representations from Transformers. BERT[62] is based on the encoder layer of the transformer. The inspiration behind BERT stems from the challenge of capturing word meaning within the broader context of a sentence. At the core of the transformer (encoder) is the multi-headed self-attention layer. The bi-directional nature of BERT, along with the self-attention mechanism, allows each word to attend to all other words simultaneously. Each word embedding is suitably tuned with the weight matrix to capture the contextual relationship with different words. The parallelism in learning embeddings allows BERT to learn long-range dependencies, avoiding the bottleneck due to sequential networks such as RNN (LSTM/GRU). Combining the bidirectional architecture, self-attention mechanism, and pre-training objectives (Masked Language Model and Next Sentence Prediction), BERT generates contextualized word embeddings that effectively represent the semantic and syntactic information in a sentence. In our work, we have utilized the contextual knowledge of the pre-trained, un-cassed hugging face BERT base model. We fine-tuned it for the downstream task of sentence classification and obtained an accuracy of 81% for Alzheimer's dementia classification from sentences.

3.3.1 BERT Architecture

BERT is stacked layers of "transformer-encoders". The core of the transformer and hence BERT is the "Self Attention" layer. Each self-attention layer is followed by feedforward layers. The use of self-attention allows parallel processing of all the words in a sentence. Parallel processing helps in capturing long-range contextual dependencies and efficient use of parallelism in computation.

As seen in Figure 3.9 and Figure 3.9 and discussed above, BERT is stacked layers of Transformer-encoders. Each transformer-encoder is composed of "Input Embedding" and "Positional Encoding" layers, followed by the most essential "Multi headed Attention" and "feed-forward" layers. The function of each layer is very symbolic of their names. All the layers together help BERT achieve state-of-the-art results over various NLP tasks. The distinctive feature of BERT is its unified architecture across different downstream tasks. Understanding the workings of the basic layer in BERT can help us to have a foundational understanding of the significance of BERT in Alzheimer's dementia classification using transcripts.

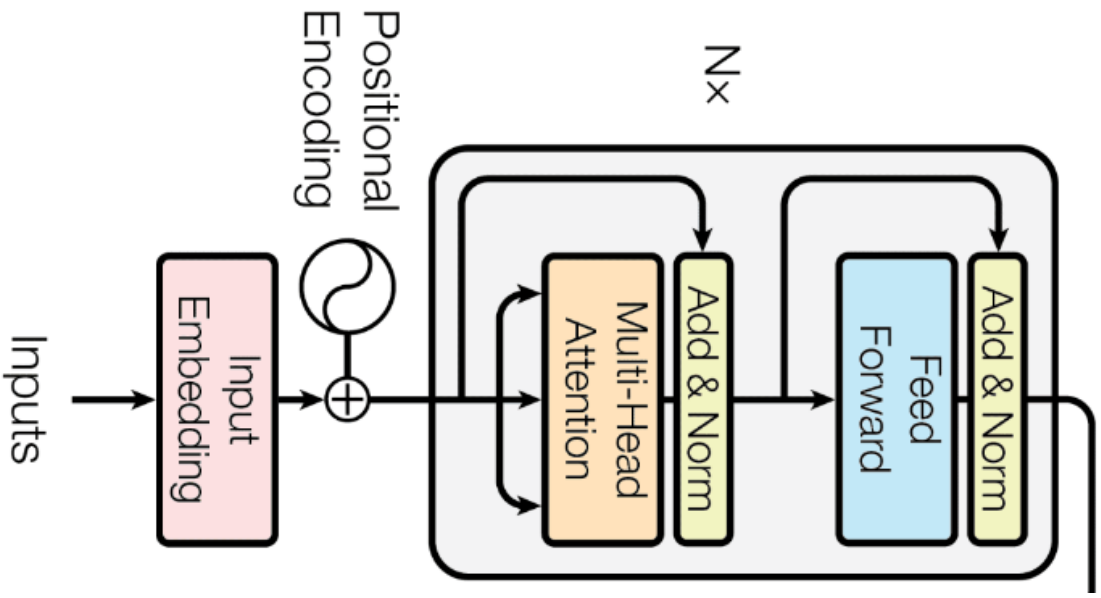

Figure 3.8: Transformer-encoder

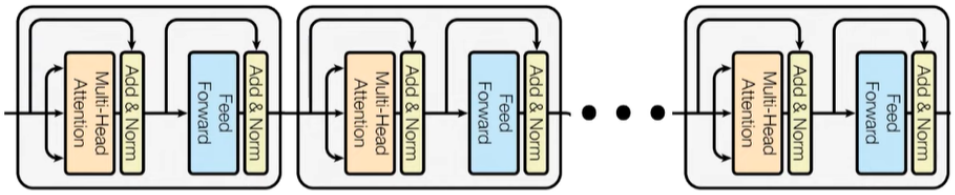

Figure 3.9: BERT

3.3.2 Input Embedding

The first layer in a transformer Encoder is the input embedding layer. The layer creates a fixed-size vector for each word token in a sentence. These tokens are learned by training an embedding matrix. The embedding matrix maps each token in the input sequence into an embedding vector suitable for the specific use case.

3.3.3 Positional Encoding

In BERT, the input embedding layer helps to feed vectorized tokens in a sentence parallelly to the following processing layers. There is no sequential information in these embeddings as they are fed parallelly. The order of words plays a crucial role in all Language processing tasks. Positional encoding is introduced to inject information about the positions of tokens into the model. It is usually a fixed-size vector added to the input embeddings to convey positional information.

3.3.4 Self Attention

The sequential model RNN/LSTM/GRU has been used for the longest time for sequential data processing. These models maintain an internal hidden state/memory that is modified recurrently as new tokens in the sequence arrive. Output at any time step depends on the current input and the previous hidden state, which depends on the previous hidden state, and so on. The recurrent mechanism captures the context while learning the embeddings for a sentence/sequence. Sequential models have reasonable performance on sequential tasks. However, they face problems with Capturing Long-Term Dependencies, vanishing and exploding gradients, computational complexity, lack of Parallelism, etc.

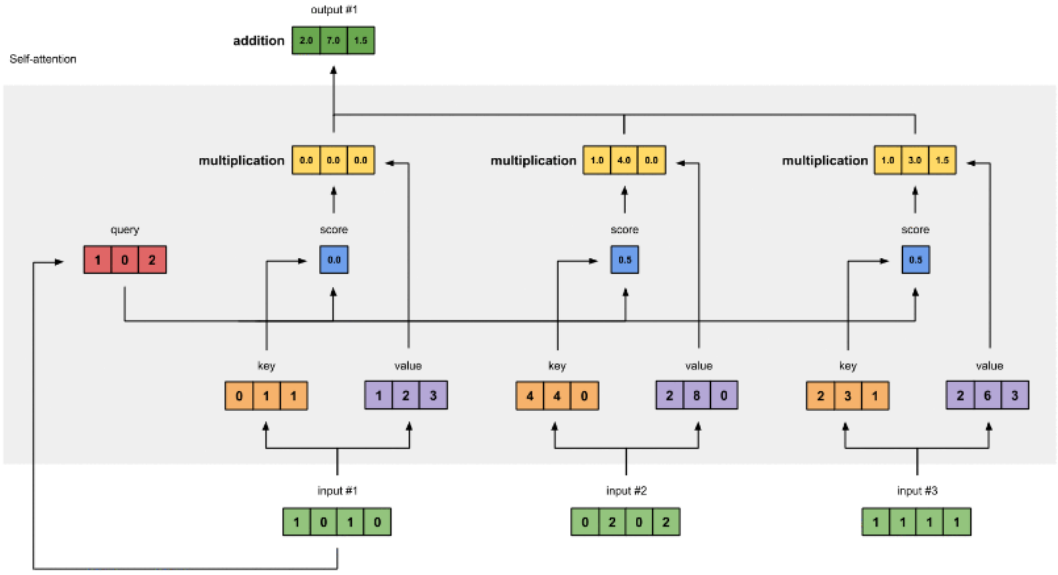

Figure 3.10: Self-Attention Mechanism

The attention mechanism reintroduced in the paper "Attention is all you need[63]" addressed most of the above problems and revolutionized the text processing field. The attention mechanism allows us to obtain context-aware embeddings. The contextual information is obtained by finding similarity scores (dot product) between each word in the sequence. The similarity score between any pair of words is independent of the separation of the words in the sequence (since a dot product). This independence from the positioning of words in a sequence is achieved by feeding all the input tokens in the sequence simultaneously, in parallel, as shown in figure 3.10. Parallel processing, along with other advantages, reduces the computational complexity by efficiently using GPUs. One easy way to find the relationship between words is through dot products. In "attention," each word in the sequence "dots" with every other word to obtain a similarity score. The similarity score is used to weigh each word to find their contribution to the specific word based on the context. Once the weighted embeddings are obtained for each word, they are combined together to obtain a final embedding to represent the sentence in the most efficient manner. Learning in BERT takes place by using the key, query, and value matrices. Though dot product is fundamental to self-attention, it doesn't enable the ability to learn. Each word embedding has a corresponding key, query and value vector as shown in fig3.11 for two words (can be extended for all the words in a sentence). It is the coefficients in the W^Q, Q^K and W^V matrix that are learned through the process of back-propagation. Once the Q, K and V is obtained for each word through matrix multiplication with the learned coefficients of the

matrices W^Q, Q^K, and W^V, the final embeddings for the words can be produced by one simple step as shown in figure . The final embedding Z obtained is the context-aware embedding for each word in the sentence (two words in this case). Z can then be passed to a sequence of feedforward and add-normalize layers before it can be used for sentence classification.

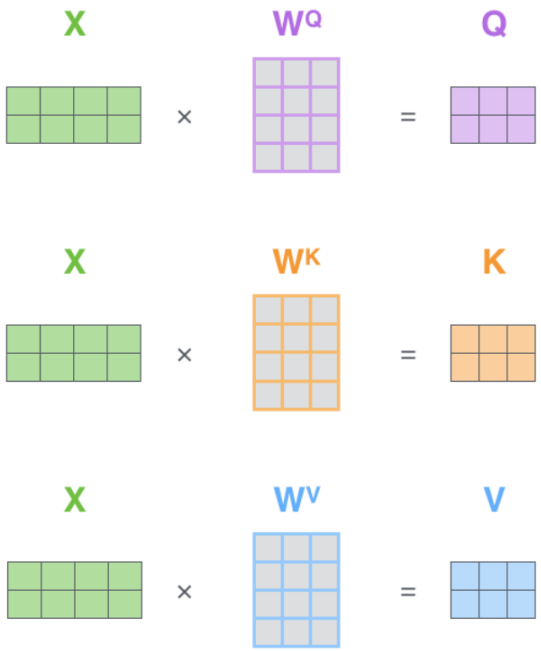

Figure 3.11: Key Query Value

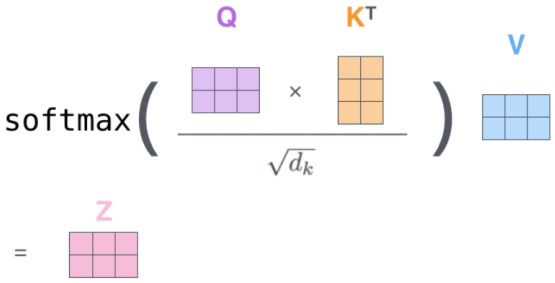

Figure 3.12: Self Attention Calculation

3.3.5 Multi-Head Attention

Multi-Head Attention is a crucial component of the Transformer architecture, specifically in the self-attention mechanism. The self-attention mechanism allows each element in a sequence to focus on different parts of the sequence, capturing dependencies and relationships between different elements. Multi-Head Attention enhances this mechanism by running multiple self-attention heads (layers) in parallel, enabling the model to capture different aspects of relationships within the sequence simultaneously.

3.3.6 Add & Norm

Add & Norm helps maintain stable and efficient training by incorporating residual connections and layer normalization within each layer of the Transformer encoder model.

Finally few other layers are added to top of the feedforward layers to enable sentence classification with BERT. The pre-trained model is fine-tuned with the manual transcripts provided with the recordings of Alzheimer's dementia Classification challenge [64]. The following section describes the experimental setup for obtaining results from acoustic and linguistic models.

3.4 Experimental Details

In the following section, we discuss the feature extraction from the acoustic algorithms. We briefly describe Machine Learning models for classifying Alzheimer's from acoustic features. The section gives details of the pre-trained BERT model that is fine-tuned for detecting Alzheimer's from the transcripts. All this follows with an overview of the dataset used and the workflow to obtain the results.

3.4.1 Feature Extraction

In this work, the statistical features are computed from the frame-level features. First, frame-level MFCC, LPCC, ZFFCC, GVVCC, and ResCC are estimated. Then, from each of the d-dimensional frame-level acoustic features, statistics (Stat), namely, mean, standard deviation, skew, and kurtosis, are computed to obtain the D–dimensional (D = 6 * d) utterance level feature vector, as in [65]. LTAS and EMS features are extracted as in [60, 66, 61]. In this study, nine acoustic features, including the state-of-the-art Computational Paralinguistic Challenge (ComParE) feature sets are referred. Details of all the features except ComParE are discussed in Section 3.2. TheComParE feature set is a brute-forced set [67] with a feature size of 6373. It is usually designed to extract paralinguistic information from the acoustic signal. The ComParE feature extraction is performed using the openSMILE toolkit. Other details of each acoustic feature considered for detecting Alzheimer's dementia from speech are presented in Table 3.1

3.4.2 Classifier

This work uses five different classifiers for the experiment. The most basic and interpretable euclidean distance-based K-Nearest Neighbours (KNN) classifier. Linear Discriminant Analysis (LDA), that assumes the class conditional density of data over the labels to be normally distributed and has a closed-form solution for the training data. Support Vector Machines with linear kernel, which handles linear classification with the margin of error for minimizing classification error on the training data. Finally, decision tree and random forest classifiers are used to have good accuracy for high-dimensional data. Decision tree is often prone to overfitting and bias error. Random forest using a large number of decision trees of arbitrary depth and taking a maximum vote over the trees mitigates the problem of error due to bias and variance. The classifiers used for the experiments are more or less complementary to each other and tries to capture variability in the data dimension and it's arrangement in space.

3.4.3 Classifying with BERT

This work uses pre-trained BERT model and fine-tunes it to detect dementia from speech transcripts. BERT is trained on two tasks:

- **Masked LM**:The authors have masked 15% of the tokens. Then, the model attempts to predict these masked tokens.
- **Next Sentence Prediction (NSP)**: The model is trained to predict the second sentence given the first. While training the model, the authors have made two kinds of pairs of sentences. One where the second sentence is related to the first and one where the second sentence is picked randomly from the corpus.

The text classification task, similar to NSP, can be done by adding another classification layer on top of the transformer [63] encoder's output. We propose to classify dementia speech transcripts using such a model.

- First, the speech transcripts are separated into train transcripts and test transcripts. To ensure no leakage between the train dataset and test dataset, the test dataset is nowhere used in the experiment except for final testing .
- Next, we have split each of the transcript files into sentences. These sentences were used to fine-tune a BERT model on a sentence classification task.
- Finally, to get the test accuracy, we ran a maximum voting algorithm on the test transcripts' sentence scores to get a single score for each transcript.

To implement this, we used cased base BERT model that the Hugging Face1 provides to tokenise and encode the sentences. To fine-tune the model, it is run for 8 epochs using Adam optimiser with a batch size of 16 and the learning rate set to $2 * 10^{-}5$. This was all done in a Pytorch environment.

3.4.4 Dataset

This work uses the ADReSS challenge Database, Alzheimer's dementia Recognition through Spontaneous Speech [19] for Acoustic models (classification with acoustic features). The dataset consists of speech recording and transcripts of spoken picture description elicited from the participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam. It consists of fifty-four AD (Alzheimer's dementia) and fifty-four NON-AD subjects in the training dataset. The test dataset consists of 24 AD and 24 Non-AD subjects. The dataset is balanced for age and gender to minimize the risk of bias in the prediction task.

The Dataset used for training the BERT Model uses all the single instances of a speaker (without repetition) in the entire Cookie Theft dataset once the test dataset is segregated. Larger training data as compared to the acoustic models allows us to fine-tune BERT better. Larger dataset is needed for fine-tuned BERT model to perform better. This has been documented in [68]. The test dataset is kept the same for the Acoustic models and Bert Model. The common dataset allows for direct comparison and combination of BERT Model and Acoustic results. Transcripts were annotated using the CHAT coding system and were converted into text files for processing through BERT Model.

3.4.5 Procedure

This work has a simple procedure that is very close to the procedure used in the baseline paper for the challenge [64]. The steps used for generating the result tables are as follows:

- Voiced audio chunks are produced from each speech wave instance (observation/recording) in the training and test dataset by Voice activity detection using the python interface "webrtc Voice Activity Detector". These audio chunks are normalized across all the chunks to control for variations caused by recording conditions.
- Class label is assigned to an observation in case of training or testing by taking a maximum vote (MV) over the class labels assigned to the audio chunks corresponding to the specific observation.
- Feature extraction is performed over the audio chunks for all the eight acoustic features and comParE features used in this work.
- 10-fold cross-validation along with MV is used to report the training accuracy of each ML model and feature pair.
- Finally, the models are trained over the audio chunks in the training dataset. The trained models perform target prediction over the features corresponding to the audio chunks in the test dataset. The test data accuracy is reported after doing MV over the class labels assigned to the audio chunks.
- Above two steps are followed to find accuracy over merged acoustic features as well.
- The BERT is trained over the sentences in transcripts of the training dataset, and the training accuracy over the sentence classification is reported. Also, the trained BERT Model is used to perform target predictions over the test dataset's sentences. The test accuracy is reported once target prediction is made through the maximum voting scheme over test data sentences.
- The class labels assigned to audio chunks (sentences) corresponding to observation are summed to estimate the probability that a given observation in the test dataset is AD (label=1).
- The complementary nature of the acoustic features to the BERT is observed by performing classification using the class labels obtained after the weighted combination of the above probability scores for the test dataset (weighted average ensemble).

3.5 Results and Discussion

This work intends to study the robust acoustic features for detecting Alzheimer's dementia from speech signals. As discussed above, all the features are extracted after applying voice activity detection and normalization on the WAV files. Further, the work also studies the performance of augmented features. The augmented features are obtained by merging the three best-performing individual features. Finally, 10-fold CV accuracy and confusion matrices are scanned to get the model and feature best suited for the Alzheimer's dementia classification task. The 10-fold Cross-Validation accuracy and test accuracy for MFCC-Stat, LPCC-stat, ResCC-stat, GVVCC-stat, ZFFCC-stat, EMS-Feat, LTAS-Feat, STAT-Feat, and the comParE feature are computed using LDA, KNN (K=1), SVM (linear kernel), DT (max-leaf-nodes=20), and RF (n-estimators=50, max-leaf-nodes=20). Consistency of the models with the baseline paper [19] provides us the freedom to compare the top-performing feature among the eight features with the performance of the ComParE feature used in the baseline. The train and test accuracy results are reported in 3.3

The accuracy of the individual features for different models for the training and test dataset are presented in Table 3.3. The first step is model-feature selection through 10-fold cross-validation (CV). The model feature with a CV accuracy of more than 51 % is selected and is followed by an elimination step. The model feature with

Table 3.3: Individual Feature train and test accuracy (0-100% scale)

	LDA		RF		KNN		SVM		DT	
	train	test	train	test	train	test	train	test	train	test
S1	57.9	51.1	52.1	48.9	47.3	55.3	58.9	55.3	48.2	48.9
S2	59.8	46.8	52.9	57.4	**51.2**	**62.5**	59.0	51.1	50.1	55.3
S3	50.2	48.9	50.0	41.0	51.9	55.3	56.1	51.0	57.0	46.8
S4	49.2	40.4	48.2	53.1	**54.3**	**60.4**	52.0	51.0	47.1	47.0
S5	43.4	55.3	53.9	53.1	**54.9**	**64.5**	46.4	61.7	52.1	51.0
S6	48.2	53.1	47.0	51.0	**54.1**	**62.5**	49.2	51.0	53.0	46.8
S7	52.9	46.8	45.0	46.8	52.2	46.8	49.1	51.0	53.0	46.8
S8	56.1	53.2	50.4	53.2	58.2	53.2	54.0	53.2	45.6	53.2
S9	56.0	62.5	50.9	54.2	57.4	45.8	52.8	50.0	52.8	62.5

a test accuracy of less than a chance probability of 50 % is eliminated. Finally, of all the models and features left, the combination that performs the best over test data is chosen as the best model feature pair for the task of Alzheimer's dementia classification.

From 3.3, it can be observed that the best accuracy of 64.5 % is obtained from ZFFCC-Stat (156) with the KNN algorithm; EMS-feat (48), LPCC-Stat (144), and GVVCC-Stat (156) all show promising performance with the accuracy of 62.5 %, 62.5 %, and 60.4 % respectively with KNN. Though all four features provide an accuracy of around 60 %, the superior performance of the ZFFCC-stat can be better observed from the confusion matrix in Table 3.4. It gives the least number of false-negative cases out of the four top-performing acoustic features. A low false-negative is highly desirable in medical diagnosis. Also, it is noteworthy that the best accuracy is obtained by KNN, one of the simplest classifiers. The accuracy of 64.5 % is comparable to the baseline performance of 62.5 % with the ComParE feature. ZFFCC-stat helps to track the epoch locations in the speech signal. Epochs represent the location of significant vocal tract vibrations at voiced frames. Parameters extracted from voiced and unvoiced frames, such as duration and density of voiced and unvoiced frames, have been used in the past for Alzheimer's dementia Detection. Therefore, the maximum accuracy of the ZFFCC-stat encourages us to look deeper into excitation source characteristics of speech signals for Alzheimer's dementia classification.

Since the top three features, ZFFCC-stat, EMS-feat, and LPCC-stat, all give the best accuracy with KNN; this work merges these three features in all possible combinations and trains KNN over them to observe any change in accuracy. The 10-fold cross-validation and test accuracies for merged features are noted down in Table 3.5

Table 3.4: Confusion Matrix Test-data ZFFCC-stat

	positive	negative
positive	15	9
negative	8	16

Table 3.5: : Classification accuracy for combinations of features

Feature	Model	train	test
S5 + S6	KNN	54.8	63.8
S5 + S2	KNN	54.8	59.5
S6 + S2	KNN	56.8	59.5
S5 + S6 + S2	KNN	54.8	63.8

The last part of the experiment involves training the BERT Model with the Pitt Cookie Theft dataset. For training the BERT Model uses all single instances of a speaker in the entire Cookie Theft dataset once the test dataset (same as test data for the acoustic model) is segregated. The cookie theft has multiple wavfiles corresponding to a specific speaker which corresponds to multiple visits of the patient to the doctor over several years. Any data leakage from the test set to the training set is removed. Also, in cross-validation, K-folds over chunks are created by splitting over the observation so that no two folds have audio chunks corresponding to the same speaker. Though the BERT model's training data is a super-set of the training data used for the acoustic models, test data is consistent for both BERT Model and acoustic models. The common test dataset allows for the direct comparison of Linguistic features' role to acoustic features for the Alzheimer's dementia classification task. It also enables us to get the combined performance of BERT Model and acoustic features over the test data. Table 3.6 summarizes BERT Model's performance over training and test dataset.

Table 3.6: BERT Classification Accuracy

	Train acc.	Test acc.
Bert Model	84.4	79.1

Table 3.7: Score fusion accuracy BERT Model and Acoustic models.

combination	model	test accuracy
0.1*EMS + 0.9*Bert	KNN	82.9
0.1*LPCC + 0.8*Bert	DT	82.9
0.2*LPCC + 0.8*Bert	KNN	85.2
0.1*LTAS + 0.9*Bert	DT	82.9
0.1*STAT + 0.9*Bert	KNN	82.9
0.2*ZFF + 0.8*Bert	KNN	82.9

3.6 Summary and Conclusion

This work studies the attributes of the speech production systems that are affected due to Alzheimer's dementia using different acoustic features. Among all the acoustic features, the top three classification accuracies for dementia are 64.5%, 62.5%, and 62.5%, obtained for ZFFCC_Stat (excitation source), EMS_feat (prosody), and LPCC_Stat (vocal-tract system), respectively. These individual features' performance is at par with the high dimensional ComParE feature. The performance of KNN (k=1) is consistently better than other classifiers, indicating some level of clustering of the features in the Euclidean space. In addition to the above, the study explores the BERT model to detect dementia. BERT provides an accuracy of around 79% on the test dataset. BERT model shows better accuracy than the acoustic features, which indicates that the characteristics of Alzheimer's dementia are manifested better in linguistic features than in the acoustics features. However, combining the acoustic feature based model's output with that of the linguistic model shows an increase in classification accuracy, which demonstrates the complementary information captured by the acoustic features. Extending this work in the future, we intend to explore higher-order features like intonation, duration, pause patterns, articulatory impairments, slowness, rhythm, etc., for dementia classification. We also plan to explore different techniques to combine acoustic and linguistic features to improve the overall accuracy of Alzheimer's dementia classification task.

Chapter 4

Single Frequency Filtering Representation for Alzheimer's Dementia

4.1 Introduction

Speech is the output of the dynamic vocal tract system. Speech features manifest at different resolutions at varying amplitudes or energy levels within speech utterances. Traditional speech signal analysis methods have primarily relied on block processing, where the signal is windowed into short time frames of around 20-30 milliseconds. However, this approach averages out significant spectral and temporal variations present in speech, which might be crucial for detecting conditions like Alzheimer's disease. In the context of Alzheimer's disease, it is hypothesized that these abnormal spectral temporal variations in speech hold valuable diagnostic information. To capture the spectral temporal variations effectively, an alternative approach using instantaneous spectral features obtained from filter banks is proposed. Specifically, a method known as Single Frequency Filtering (SFF) is employed. SFF provides not only the magnitude or envelope of the speech signal but also its phase at any desired frequency with high-frequency resolution. The goal is to identify robust speech-specific features and develop methods to extract them from speech signals. The approach, based on Single Frequency Filtering, offers a higher-resolution perspective on speech signals, capturing both spectral and temporal details for Alzheimer's classification from the speech signal.

4.2 Motivation

The motivation for this study is to explore the potential of Single Frequency Filtering Cepstral Coefficients (SFCC) for automatically detecting Alzheimer's disease. Unlike traditional Short-Time Fourier Transforms (STFTs), the SFCC-based feature exhibits superior temporal and spectral resolution, enabling it to more appropriately capture transient characteristics in speech. It offers an efficient and compact way to derive the formant structure in speech signals. Experiments are conducted using the ADReSSo dataset[21], and a support vector machine(SVM) classifier was trained for Alzheimer's disease classification. While many studies have focused on acoustic features in speech analysis, most have used block processing, which might not effectively capture instantaneous spectral changes. A single frequency filter (SFF) based spectrum was proposed in the literature to capture both spectral and temporal resolution. In [69, 70, 66, 71], Single frequency filtering (SFF) representation is used for investigating robust epoch extraction, speakers separation, voice activity detection (VAD), speakers spoofing, and voice disorder classification. The better temporal and spectral resolution helps the SFF spectrum to capture the transient characteristics of speech signals more appropriately. The high spectral resolution indicates clear harmonic structures and is also called the timber spectrum. The importance of such patterns is to track formant structure

accurately. The conventional feature extraction is based on block processing with a typical frame size of 20-25 ms and an overlap of 10 ms. A shorter window gives high temporal resolution but low spectral resolution [72], whereas the larger window provides high spectral resolution but low temporal resolution. This motivates us to apply the SFF-based technique for the Alzheimer's detection task because the non-stationarity in speech signals can be captured easily. Given the above motivation, this work's main contribution is a novel auditory scaling incorporated into a single-frequency filtering approach and corresponding cepstral coefficient generation(SFCC) for the Alzheimer's detection task. Later, the experimental validations are performed on the ADReSSo Challenge database. [21].

4.3 Single Frequency Filter Bank

Speech signals have correlations among samples along time at each given frequency and across frequencies for a given time sample. In Single Frequency Filtering, the amplitude envelope of the signal is obtained at each frequency with high temporal and spectral resolution. Since the method is based on extracting energy at a single frequency, it is called the single frequency filtering (SFF) method. The envelope is computed at every 20 Hz in the range of 300 Hz to 4000 Hz as a function of time. The SFF method uses a near-zero bandwidth resonator and extracts information from the speech signal (if present) at a particular frequency with high power. The amplitude envelopes derived using the SFF method are processed for auditory enhancement and to obtain the cepstral coefficients. The speech feature obtained is the Single Frequency Cepstral Coefficients(SFCC). The following section initially explains a mathematical formulation of the single-frequency filtering (SFF) approach. Later, we describe the extraction of cepstral coefficients SFCC from SFF.

4.3.1 Proposed Feature Extraction

This section initially explains a mathematical formulation of the single-frequency filtering (SFF) approach. Later, we describe the extraction of cepstral coefficients SFCC from SFF.

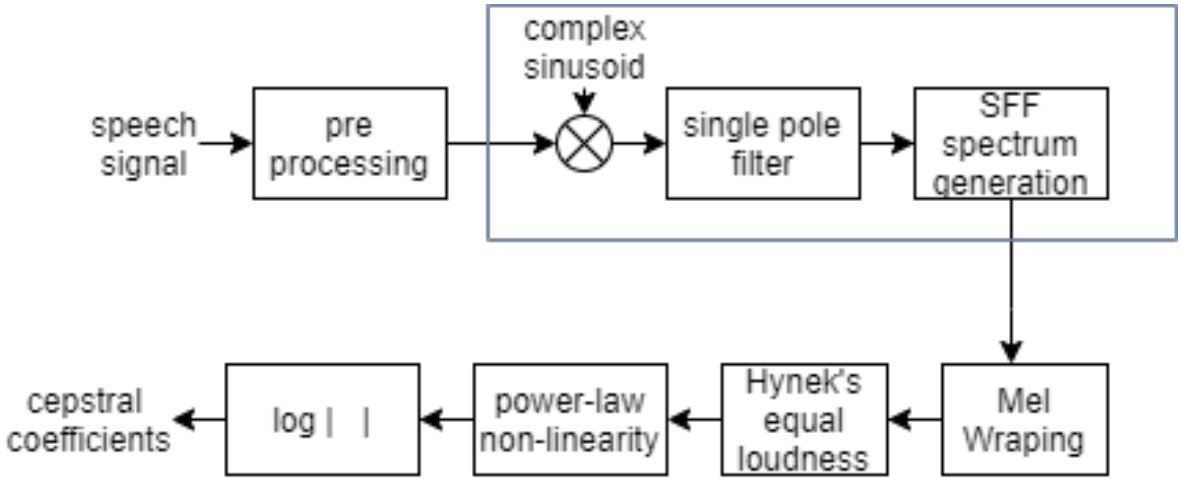

Figure 4.1: Functional block diagram of SFCC feature extraction

4.3.2 Single frequency filtering

The single frequency filtering (SFF) approach involves generating an amplitude envelope at every sample for a selected frequency f_k for the pole location(r) closer to the unit circle so that the information is extracted at half the sampling frequency. The appropriate spectro-temporal resolution can be achieved by varying the r value, i.e., the pole location in the z-plane. We state the procedure to extract SFF as follows:

1. A discrete-time speech signal $x[n]$ is passed through a pre-emphasis filter to remove low-frequency components present in it.i.e.,

$$s[n] = x[n] - x[n-1] \tag{4.1}$$

2. The emphasized signal $s[n]$ is multiplied with a complex sinusoidal ($e^{j\overline{\omega}_k n}$), with normalized shifted frequency, mathematically expressed as:

$$s[n,k] = s[n](e^{j\overline{\omega}_k n}) \tag{4.2}$$

where, s[n,k] is the resultant output for n^{th} sample and
$\overline{\omega}_k$ is the the normalized frequency at k^{th} filter.

$$\overline{\omega}_k = \frac{2\pi\overline{f}_k}{f_s} \tag{4.3}$$

where, $\overline{f}_k$ is the shift in frequency ie., $\overline{f}_k$ = $\frac{f_s}{2} - f_k$.
In Equation 2, n varies from 1 to N, and k varies from 0 to K, where N and K are the total number of samples and filters, respectively. If the spacing between each filter is Δ fHz, then the total number of filters can be computed as,

$$K = \frac{f_s/2}{\Delta f}$$

3. Now, $s[n,k]$ is passed through a single-pole filter $H(z)$

$$H(z) = \frac{1}{1 + rz^{-1}} \tag{4.4}$$

4. From Equation 4, r gives the pole location on the negative axis at $z = -r$. The stability of the transfer function is maintained by selecting the pole location inside the unit circle (radius $r = 0.994$). SFF output can be mathematically expressed as,

$$y[n,k] = -ry_k[n-1] + s[n,k] \tag{4.5}$$

where, $y[n,k]$ is a complex number with real part $y_r[n,k]$ and imaginary part $y_i[n,k]$

5. The amplitude envelopes ($v[k,n]$) is computed at k^{th} filter to produce the filtered output, which is

$$v[k,n] = \sqrt{y_r^2[n,k] + y_i^2[n,k]} \tag{4.6}$$

4.3.3 SFCC Extraction

The functional block diagram of the feature extraction framework is depicted in Fig 4.2. The detailed mathematical formulation is explained as follows:

1. **Mel Warping**: Mel frequency warping is performed on the SFF spectrum $s(n,k)$ as the human auditory perception works with non-linear frequency scaling.

$$S_{warp}(n,k) = \Phi_m\{s(n,k)\} \tag{4.7}$$

   Here $\Phi_m\{\bullet\}$ denotes Mel warping operator

2. **Equal loudness pre-emphasis:** To model non-uniform sensitivity of human hearing across the range of frequencies, we perform equal loudness pre-emphasis on the warped spectrum. In this work, Hynek's magic equal-loudness [73] is applied on warped spectrum, it is mathematically expressed as

$$S_{loudness}(n,k) = \Phi_{eqloudness}\{S_{warp}(n,k)\} \tag{4.8}$$

   Here $\Phi_{eqloudness}\{\bullet\}$ denotes Heynek's magic equal loudness.

3. **Power law non-linearity:** For speech segments of low power, the logarithmic non-linearity can produce large output changes even if the input changes are small. Due to this, degradation in speech recognition is observed as the input approaches zero. With a power-function non-linearity, the output is close to zero if the input is very small, which is observed in human auditory processing. The reason for choosing power law non-linearity over logarithmic is that the dynamic behavior of the output does not depend critically on the input amplitude. This non-linearity, which is used in SFCC feature extraction, is described by the equation,

$$S_{powerlaw}(n,k) = (S_{loudness}(n,k))^{\gamma} \tag{4.9}$$

   Where γ is some constant varying between 0 to 1.

4. Inverse Fourier transform is computed for the logarithm of power-law nonlinearity spectrum and passed through liftering to obtain the 13-dimensional cepstral coefficients, mathematically described as,

$$c(n,k) = IFFT\{log\{S_{powerlaw}(n,k)\}\} \tag{4.10}$$

   From static coefficients c(n, k), delta (Δ) and double delta ($\Delta\Delta$) coefficients are derived and appended to have 39-dimensional SFCCs.

To summarize the contributions of this work, we have explored the SFF spectrum and proposed a novel approach to extract cepstral coefficients so that it can be used for the Alzheimer's detection task. The proposed framework has been evaluated on the ADReSSo dataset [21].

4.4 Experimental Details

This section describes the database and features used for detecting Alzheimer's Dementia, including the baseline features. Further, it also describes the classifier and its parameters used in this study.

4.4.1 Dataset

The dataset used in this study for Alzheimer's Dementia detection is the ADReSSo dataset (Alzheimer's Dementia Recognition through Spontaneous Speech only). It consists of a total of 237 audio files sampled at 44kHz. The audio files were split into the cross-validation dataset and test datasets. The cross-validation dataset contains 70% of the total dataset with 79 control(healthy) speakers and 87 speakers with AD. The test dataset contains 36 control speakers and 35 AD speakers. Each speech recording in the dataset is a description of "Cookie theft picture" by the subject. All the speech samples were downsampled to 8KHz for our experimental setup.

4.4.2 Features used for Alzheimer's detection task

The mathematical formulation of SFCC is explained in Section 2, the parameters considered while extracting 39 dimension features for every 10 ms are tabulated in Table 4.1.

Table 4.1: parameter consider while extracting SFCC's

Parameters	**Values**
Radius	0.994
Sampling frequency (f_s)	8000Hz
f1 (start frequency)	100 Hz
Frequency step	20 Hz
gamma (γ)	0.45

All the features in this work are extracted in MATLAB and loaded into the framework for classification. The benchmarked features used for performance evaluation is as follows:

- **Envelope Modulation Spectrum(LPCC)** Speech envelope modulation spectra quanities the rhythmicity of speech signal in different frequency bands. Temporal regularities in amplitude envelope of speech signal and requires no segmentation. Envelope modulation spectra is the spectral analysis of low-rate amplitude of speech signal within select frequency bands. The envelope is obtained by half wave rectification and then passing the signal through a low pass filter with a cutoff of 30Hz. The resulting envelope consists of temporal variations in amplitude such as those corresponding to syllables. In EMS the amplitude envelope is caluclated for the original signal and the octave freuquency ranging from 125 Hz - 8000 Hz. This allows one to observe the rhythmic patterns that corresponds to vowel nuclei, voicing, bursts and fricatives, and so forth. It is understood that the amplitude envelopes extracted from the speech at different frquency bands are only partially correlated and hence has orhogonal information. Though EMS gives a detailed idea of the modulation at different frequency bands, it is not possible to use that to train models. 6 parameters are evaluated for the 7 octaves including the original signal. Resulting in 48 dimensional features representative of the speech rhythm.

- **Mel-frequency cepstral coefficients (MFCC) and Perceptually Linear Prediction (PLP)** are computed for 25 ms window with 5 ms frame shift. The first 13 static coefficients and corresponding delta and delta-delta features were computed which results in a total of 39 dimensions.

- **Linear prediction cepstral coefficient(LPCC)** were computed using 25 ms with 10 ms frame shift. The first 12 static coefficients, and their first and second derivative were calculated, which resulted in a total 36 dimension feature vector.
- **Mel frequency cepstral coefficients of LP-residual (MFCC-WR) and ZFF signal (MFCC-ZF)** features were obtained from excitation source signals, namely zero frequency signal and residual signal respectively. It is also computed using the frame size of 25 ms with frame shift of 5 ms [65, 74].

Then four statistical averages like mean, standard deviation, kurtosis, and skewness were calculated, which resulted in a total of 156 dimension feature vectors of MFCC, PLP, MFCC-ZF and MFCC-WR, whereas LPCC is 144 dimension feature vector.

4.4.3 Classifier

Alzheimer's disease detection was carried out using the support vector machine classifier. AD detection task was also performed using all other classifiers like Naive Bayes, logistic regression, decision trees, and KNN. For most features, SVM outperformed other classifiers in the AD detection task. Hence in this study, linear SVM classifier is used for experimentation. The experiments are conducted using 5-fold cross-validation, and finally, average classification accuracy from all the folds is reported in the work. The model performance is compared on the test data for SFCCs and SFCCs combined with baseline features.

4.5 Results and Discussions

In this study, we performed Alzheimer's disease detection on the ADReSSo dataset using single frequency filtering cepstral coefficients (SFCC). The performance was also compared with baseline features like MFCC, PLP, LPCC, MFCC-ZF, and MFCC-WR. The linear support vector machine classifier is used to perform the experiments. Performance was also evaluated by pairing SFCC features with the baseline features.

Table 4.2 shows the performance of Alzheimer's disease detection in terms of classification accuracy on the cross-validation dataset and test dataset. From Table 4.2 it can be observed that SFCC gives the highest classification accuracy of 65.1%, and 60.6% for cross-validation and test datasets, respectively. Apart from SFCC, LPCC provides cross-validation accuracy of 62.7% and test accuracy of 53.5%. It is noteworthy the huge margin by which the performance of SFCC exceeds baseline features on the test data.

Also, Alzheimer's disease detection was performed by pairing SFCC feature with the baseline features which is shown in Table 4.3. It can be observed from the table that, SFCC when combined with LPCC, shows the best CV accuracy of 65.7% among all the other pairing. For test dataset, the best performance of 63.4% can be observed by combination of SFCC with MFCC.

Later, we will examine in more details the behavior and efficacy of Single frequency filtering based features by asking the question described below.

- **For what value of *r* the SFF filter is stable?** The stability of the filter depends on the pole location that should ideally be inside the unit circle. So the optimum value of r for AD detection is estimated by arbitrarily changing the value as follows:

  - We first observe the SFF representation for $r = 0.85$. We observe that the time resolution is good, but the formant information is smeared. The value of r is changed from *0.85* to *0.90*. In comparison to $r = 0.85$, there is a slight improvement. It is observed that as r increases from 0.85 to 1.0, smearing

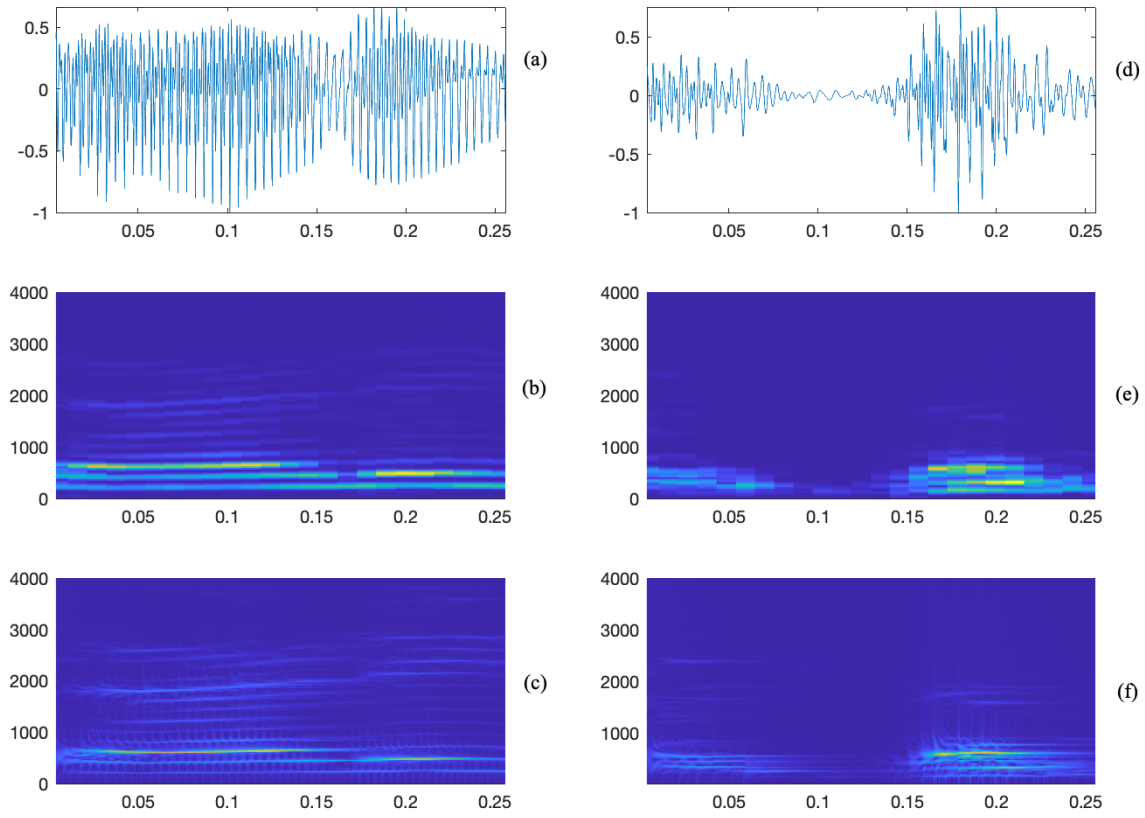

Figure 4.2: An illustration of STFT and SFF spectrogram: (a)-(c) represents speech segment from Control Speaker and it's corresponding STFT spectrogram, SFF spectrogram respectively (d)-(f) represents speech segment from AD Speaker and it's corresponding STFT spectrogram, SFF spectrogram, respectively.

Table 4.2: Classification accuracy(in percentage) of individual acoustic features for Alzheimer's Detection on ADReSSo Dataset(Cross validation and test dataset)

Features	CV	Test
MFCC	60.2	53.5
PLP	63.9	49.3
SFCC	**65.1**	**60.6**
LPCC	62.7	53.5
MFCC-ZF	56.6	50.7
MFCC-WR	59	46.5
eGeMAPs	58.5	57.1

Table 4.3: Classification accuracy(in percentage) of combined acoustic features for Alzheimer's Detection on ADReSSo Dataset(Cross validation and test dataset)

Feature	CV	Test
SFCC+MFCC	63.9	**63.4**
SFCC+PLP	65.1	59.2
SFCC+LPCC	**65.7**	61..97
SFCC+MFCC-ZF	63.3	63.38
SFCC+MFCC-WR	63.9	60.6
SFCC+eGeMAPs	63.1	58.2

in the formant is reduced. In order to hold the law of generality between time and frequency, the optimum value of *"r"* lies between 0.95 and 0.995, preserving both time and frequency information.

- **Efficiency of SFF** To understand the efficiency of SFF we have performed the analysis by considering an audio sample from the database for analysing the time-frequency representations are depicted in Fig.2. In the figure the second row represents a STFT based representation. The STFT is computed using window length of 20 ms with 10 ms overlap. The third row gives information about SFF based time-frequency representation for the same signals where left image is for the control speakers and right for the Alzheimer's speakers. It is evident from the figure that STFT based representation has more frequency spread when compared to SFF representation. So this concludes that SFF based representation gives better time-frequency resolution.

4.6 Conclusion

This work explored the various acoustic features for Alzheimer's Dementia detection. The performance of the SFCC feature was compared to several baseline features commonly used in speech analysis, such as Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), linear prediction cepstral coefficient (LPCC), Mel frequency cepstral coefficients of LP-residual (MFCC-WR), ZFF signal (MFCC-ZF), and eGeMAPS. The experiment results show the best classification accuracy of 65.1% and 60.6% obtained using the SFCC features on cross-validation and test data, respectively. Pairing SFCCs with other baseline features improves performance over the test data, though the CV accuracy stays the same. The highest test accuracy of 63.4% for the combined features is obtained for SFCCs+MFCC. The linear SVM generalizes very well with the SFCCs+MFCC as the CV accuracy is close to the test accuracy. We conclude that proposed acoustic features leverage the SFF characteristics to capture more comprehensive speech subtleties at the temporal scale, which aids in classifying Alzheimer's Dementia.

.

Chapter 5

Conclusion and Future Work

5.1 Conclusion and Future Work

This study examined the impact of Alzheimer's Dementia on the attributes of the speech production system using a range of acoustic features. Two standard releases of the datasets were used for conducting all the experiments. The second dataset[20] is an improved version of the first dataset[19]. The second dataset is speech-only and emphasizes detecting Alzheimer's directly from the speech signal. The dataset for Alzheimer's dementia detection is comparatively smaller. Conceptually, a smaller dataset would need a smaller feature size. Fortunately, we have a legacy of acoustic features whose length is almost 100 folds smaller than the size of the baseline comParE (openSMILE) feature. On the other hand, the target of this work is also to explore the acoustic features that are unexplored and set the ground performance of the acoustic algorithms related to speech production, prosody, filter-bank, and statistical features for Alzheimer's Dementia detection.

The key findings from this work included the top three classification accuracies for Dementia, with scores of 64.5% for ZFFCC_Stat (excitation source) and 62.5% for both EMS_feat (prosody) and LPCC_Stat (vocal-tract system). These individual features performed comparably to the high-dimensional ComParE(openSMILE) feature. The KNN (k=1) classifier consistently outperformed others, suggesting feature clustering in the Euclidean space. Moreover, the study explored using the BERT language model on the manual transcripts provided in the first dataset. BERT could achieve an accuracy of approximately 79% on the test dataset. BERT outperformed the acoustic features, indicating that Alzheimer's Dementia characteristics may be better manifested in linguistic features. Notably, combining the outputs of the acoustic feature-based model with the linguistic model increased classification accuracy, highlighting the complementary information captured by both feature types. The research also explored the potential of filter bank features, particularly SFCCs, in capturing speech subtleties related to Alzheimer's Dementia. Looking at speech utterance at improved frequency-time resolution can boost the accuracy and provide more valuable insights for further research and improvements in dementia detection techniques. The thesis also outlined future research directions, including exploring higher-order features such as intonation, duration, pause patterns, articulatory impairments, slowness, and rhythm for Alzheimer's classification.

In the future, we plan to investigate techniques for integrating acoustic and linguistic features better to enhance overall accuracy in Alzheimer's dementia classification. We can carry forward this work by incorporating SFF-based features with prosodic features. It would help capture supra-segmental signs in the speech utterances that may characterize Alzheimer's. A bigger dataset can give us the freedom to apply deep learning models to the problem of Alzheimer's Dementia Classification. Increased accuracy can solidify our hypothesis of speech being a significant indicator of the pre-clinical stage of Alzheimer's Dementia. Hopefully, with intensive research and development, we can curb or at least slow down the progression of Alzheimer's Dementia.

Chapter 6

List of Publications

1. **Nayan Vats**, Aditya Yadavalli, Krishna Gurugubelli, and Anil Kumar Vuppala, "Acoustic Features, BERT Model and their Complementary Nature for Alzheimer's Dementia Detection", ACM IC3, August, Noida, 2021.
2. **Nayan Vats**, Purva Barche, Ganesh S Mirishkar, and Anil Kumar Vuppala,"Exploring High Spectro-Temporal Resolution for Alzheimer's Dementia Detection", IEEE SPCOM 2022, IISC Bangalore, India.

Bibliography

[1] Martin Prince, Renata Bryce, Cleusa Ferri, et al. *World Alzheimer Report 2011: The benefits of early diagnosis and intervention*. Alzheimer's Disease International London, 2011.

[2] Alzheimer's Association. 2019 alzheimer's disease facts and figures. *Alzheimer's & dementia*, 15(3):321–387, 2019.

[3] Karen Ritchie, Isabelle Carriere, Li Su, John T O'Brien, Simon Lovestone, Katie Wells, and Craig W Ritchie. The midlife cognitive profiles of adults at high risk of late-onset alzheimer's disease: The prevent study. *Alzheimer's & Dementia*, 13(10):1089–1097, 2017.

[4] Jill Rasmussen and Haya Langerman. Alzheimer's disease–why we need early diagnosis. *Degenerative neurological and neuromuscular disease*, 9:123, 2019.

[5] Ingrid Arevalo-Rodriguez, Nadja Smailagic, Marta Roqué i Figuls, Agustín Ciapponi, Erick Sanchez-Perez, Antri Giannakou, Olga L Pedraza, Xavier Bonfill Cosp, and Sarah Cullum. Mini-mental state examination (mmse) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (mci). *Cochrane Database of Systematic Reviews*, (3), 2015.

[6] Christoph Laske, Hamid R Sohrabi, Shaun M Frost, Karmele López-de Ipiña, Peter Garrard, Massimo Buscema, Justin Dauwels, Surjo R Soekadar, Stephan Mueller, Christoph Linnemann, et al. Innovative diagnostic tools for early detection of alzheimer's disease. *Alzheimer's & Dementia*, 11(5):561–578, 2015.

[7] Alan D Baddeley, Sergio BRESSI, Sergio Della Sala, Robert Logie, and Hans SPINNLER. The decline of working memory in alzheimer's disease: A longitudinal study. *Brain*, 114(6):2521–2542, 1991.

[8] Julian Appell, Andrew Kertesz, and Michael Fisman. A study of language functioning in alzheimer patients. *Brain and language*, 17(1):73–91, 1982.

[9] Laura M Gonnerman, Justin M Aronoff, Amit Almor, Daniel Kempler, and Elaine S Andersen. From beetle to bug: Progression of error types in naming in alzheimer's disease. In *Proceedings of the annual meeting of the cognitive science society*, volume 26, 2004.

[10] Lars Bäckman, Sari Jones, Anna-Karin Berger, Erika Jonsson Laukka, and Brent J Small. Cognitive impairment in preclinical alzheimer's disease: a meta-analysis. *Neuropsychology*, 19(4):520, 2005.

[11] Shona D'Arcy, Viliam Rapcan, Nils Penard, Margaret E Morris, Ian H Robertson, and Richard B Reilly. Speech as a means of monitoring cognitive function of elderly speakers. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[12] Steven H Ferris and Martin Farlow. Language impairment in alzheimer's disease and benefits of acetylcholinesterase inhibitors. *Clinical interventions in aging*, 8:1007, 2013.

[13] Vicent Deramecourt, Florence Lebert, Brigitt Debachy, MA Mackowiak-Cordoliani, Stéphanie Bombois, Olivier Kerdraon, Luc Buée, C-A Maurage, and Florence Pasquier. Prediction of pathology in primary progressive language and speech disorders. *Neurology*, 74(1):42–49, 2010.

[14] Romola S Bucks, Sameer Singh, Joanne M Cuerden, and Gordon K Wilcock. Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91, 2000.

[15] Katrina E Forbes, Annalena Venneri, and Michael F Shanks. Distinct patterns of spontaneous speech deterioration: an early predictor of alzheimer's disease. *Brain and Cognition*, 48(2-3):356–361, 2002.

[16] Samrah Ahmed, Anne-Marie F Haigh, Celeste A de Jager, and Peter Garrard. Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain*, 136(12):3727–3737, 2013.

[17] Konrad Maurer, Stephan Volk, and Hector Gerbaldo. Auguste d and alzheimer's disease. *The lancet*, 349(9064):1546–1549, 1997.

[18] Vanesa Nieves López. Evaluación del discurso de las personas con enfermedad de alzheimer: una revisión. 2016.

[19] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: the adress challenge. *arXiv preprint arXiv:2004.06833*, 2020.

[20] Jun Chen, Jieping Ye, Fengyi Tang, and Jiayu Zhou. Automatic detection of alzheimer's disease using spontaneous speech only. In *Proc. Interspeech*, pages 3830–3834, 2021.

[21] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Detecting cognitive decline using speech only: The adresso challenge. *medRxiv*, 2021.

[22] Saturnino Luz, Sofia de la Fuente, and Pierre Albert. A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*, 2018.

[23] Fasih Haider, Sofia De La Fuente, and Saturnino Luz. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2019.

[24] Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference Mechatronics and Automation, 2005*, volume 3, pages 1569–1574. IEEE, 2005.

[25] Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. Detecting Signs of Dementia Using Word Vector Representations. In *Proc. Interspeech 2018*, pages 1893–1897, 2018.

[26] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016.

[27] Muhammad Shehram Shah Syed, Zafi Sherhan Syed, Margaret Lech, and Elena Pirogova. Automated screening for alzheimer's dementia through spontaneous speech. In *INTERSPEECH*, pages 2222–2226, 2020.

[28] Harold Goodglass, Edith Kaplan, and Sandra Weintraub. *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[29] Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *INTERSPEECH*, pages 2162–2166, 2020.

[30] Karen Croot, John R Hodges, John Xuereb, and Karalyn Patterson. Phonological and articulatory impairment in alzheimer's disease: a case series. *Brain and language*, 75(2):277–309, 2000.

[31] Ildikó Hoffmann, Dezso Nemeth, Cristina D Dye, Magdolna Pákáski, Tamás Irinyi, and János Kálmán. Temporal parameters of spontaneous speech in alzheimer's disease. *International journal of speech-language pathology*, 12(1):29–34, 2010.

[32] Juan José G Meilán, Francisco Martínez-Sánchez, Juan Carro, Dolores E López, Lymarie Millian-Morell, and José M Arana. Speech in alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334, 2014.

[33] Juan JG Meilán, Francisco Martínez-Sánchez, Juan Carro, José A Sánchez, and Enrique Pérez. Acoustic markers associated with impairment in language processing in alzheimer's disease. *The Spanish journal of psychology*, 15(2):487–494, 2012.

[34] Aharon Satt, Alexander Sorin, Orith Toledo-Ronen, Oren Barkan, Ioannis Kompatsiaris, Athina Kokonozi, and Magda Tsolaki. Evaluation of speech-based protocol for detection of early-stage dementia. In *Interspeech*, pages 1692–1696, 2013.

[35] Ali Khodabakhsh, Serhan Kuscuoglu, and Cenk Demiroglu. Detection of alzheimer's disease using prosodic cues in conversational speech. In *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, pages 1003–1006. IEEE, 2014.

[36] Sabah Al-Hameed, Mohammed Benaissa, and Heidi Christensen. Simple and robust audio-based detection of biomarkers for alzheimer's disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 32–36, 2016.

[37] Saeideh Mirzaei, Mounim El Yacoubi, Sonia Garcia-Salicetti, Jérôme Boudy, C Kahindo, V Cristancho-Lacroix, Hélène Kerhervé, and A-S Rigaud. Two-stage feature selection of voice parameters for early alzheimer's disease prediction. *Irbm*, 39(6):430–435, 2018.

[38] Francisco Martínez-Sánchez, Juan JG Meilán, Juan Antonio Vera-Ferrandiz, Juan Carro, Isabel M Pujante-Valverde, Olga Ivanova, and Nuria Carcavilla. Speech rhythm alterations in spanish-speaking individuals with alzheimer's disease. *Aging, Neuropsychology, and Cognition*, 24(4):418–434, 2017.

[39] Saturnino Luz. Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 45–46. IEEE, 2017.

[40] James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594, 1994.

[41] Roozbeh Sadeghian, J David Schaffer, and Stephen A Zahorian. Speech processing approach for diagnosing dementia in an early stage. 2017.

[42] Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W Schuller, Mathew Magimai-Doss, Helmer Strik, et al. A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. In *Interspeech 2020*, pages 2182–2186. ISCA-International Speech Communication Association, 2020.

[43] Matej Martinc and Senja Pollak. Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. In *INTERSPEECH*, pages 2157–2161, 2020.

[44] Morteza Rohanian, Julian Hough, and Matthew Purver. Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech. *arXiv preprint arXiv:2106.09668*, 2021.

[45] Erik Edwards, Charles Dognin, Bajibabu Bollepalli, Maneesh Kumar Singh, and Verisk Analytics. Multiscale system for alzheimer's dementia recognition through spontaneous speech. In *INTERSPEECH*, pages 2197–2201, 2020.

[46] Anna Pompili, Thomas Rolland, and Alberto Abad. The inesc-id multi-modal system for the adress 2020 challenge. *arXiv preprint arXiv:2005.14646*, 2020.

[47] Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. *arXiv preprint arXiv:2009.00700*, 2020.

[48] Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee. Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition. *arXiv preprint arXiv:2009.04070*, 2020.

[49] Lara Gauder, Leonardo Pepino, Luciana Ferrer, and Pablo Riera. Alzheimer disease recognition using speech-based embeddings from pre-trained models. In *Proc. INTERSPEECH*, pages 3795–3799, 2021.

[50] Yilin Pan, Bahman Mirheidari, Jennifer M Harris, Jennifer C Thompson, Matthew Jones, Julie S Snowden, Daniel Blackburn, and Heidi Christensen. Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. In *Proc. Interspeech*, pages 3810–3814, 2021.

[51] Aparna Balagopalan and Jekaterina Novikova. Comparing acoustic-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2106.01555*, 2021.

[52] Zafi Sherhan Syed, Muhammad Shehram Shah Syed, Margaret Lech, and Elena Pirogova. Tackling the adresso challenge 2021: The muet-rmit system for alzheimer's dementia recognition from spontaneous speech. *Proc. Interspeech 2021*, pages 3815–3819, 2021.

[53] Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velázquez, Piotr Zelasko, Jesús Villalba, and Najim Dehak. Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. In *Proc. Interspeech*, pages 3825–3829, 2021.

[54] Ning Wang, Yupeng Cao, Shuai Hao, Zongru Shao, and KP Subbalakshmi. Modular multi-modal attention network for alzheimer's disease detection using patient audio and language data. *Proc. Interspeech 2021*, pages 3835–3839, 2021.

[55] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[56] Zrar Kh Abdul and Abdulbasit K Al-Talabani. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 2022.

[57] Manu Airaksinen, Tuomo Raitio, Brad Story, and Paavo Alku. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):596–607, 2013.

[58] K Sri Rama Murty and Bayya Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613, 2008.

[59] Krishna Gurugubelli and Anil Kumar Vuppala. Stable implementation of zero frequency filtering of speech signals for efficient epoch extraction. *IEEE Signal Processing Letters*, 26(9):1310–1314, 2019.

[60] Julie M Liss, Sue LeGendre, and Andrew J Lotto. Discriminating dysarthria type from envelope modulation spectra. 2010.

[61] Elvira Mendoza, Nieves Valencia, Juana Muñoz, and Humberto Trujillo. Differences in voice quality between men and women: Use of the long-term average spectrum (ltas). *Journal of voice*, 10(1):59–66, 1996.

[62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[64] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The adress challenge, 2020.

[65] Purva Barche, Krishna Gurugubelli, and Anil Kumar Vuppala. Towards automatic assessment of voice disorders: A clinical approach. In *INTERSPEECH*, pages 2537–2541, 2020.

[66] KNRK Raju Alluri, Sivanand Achanta, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty, and Anil Kumar Vuppala. Sff anti-spoofer: Iiit-h submission for automatic speaker verification spoofing and countermeasures challenge 2017. In *Interspeech*, pages 107–111, 2017.

[67] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition. 2015.

[68] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer, 2019.

[69] G Aneeja and B Yegnanarayana. Single frequency filtering approach for discriminating speech and non-speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):705–717, 2015.

[70] Sudarsana Reddy Kadiri and B Yegnanarayana. Epoch extraction from emotional speech using single frequency filtering approach. *Speech Communication*, 86:52–63, 2017.

[71] Sudarsana Reddy Kadiri, Rashmi Kethireddy, Paavo Alku, et al. Parkinson's disease detection from speech using single frequency filtering cepstral coefficients. In *Interspeech*, pages 4971–4975, 2020.

[72] Yegnanarayana Bayya and Dhananjaya N Gowda. Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. *Speech Communication*, 55(6):782–795, 2013.

[73] Hynek Hermansky and Jean-Claude Junqua. Optimization of perceptually-based asr front-end (automatic speech recognition). In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 219–220. IEEE Computer Society, 1988.

[74] Sudarsana Reddy Kadiri and Paavo Alku. Analysis and detection of pathological voice using glottal source features. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):367–379, 2019.