University of Padova

Department of Mathematics

Master Thesis in Data Science

Applying Machine Learning Techniques to forecast the level of dementia from spontaneous speech conversations

Supervisor

Lamberto Ballan

University of Padova

Co-supervisor

Federico Ghirardelli

Head of AI Co-Founder @ DeepVibes

Master Candidate

Hau Yee, Cheung

Academic Year

2021-2022

Dedication.

This thesis dedicates to my parents, Without their support and understanding, this thesis would never have been completed.

Abstract

This report summarizes the duties and results conducted in the work placement. The project in the work placement was about applying machine learning to mobile applications, which analyzed human language to forecast speakers' level of dementia. It is split into four parts, including collecting audio data, exploratory data analysis, machine learning analysis, and applying the machine learning model to the existing product. Firstly, collected dementia group and normal group audio data samples as the input for further analysis usage. In the second part, I explored the data to find insights and preprocess the dataset before building machine learning models. After that, building machine learning models analyzed the characteristics of the speech and forecasted speakers' level of dementia. At last, combining machine learning results in mobile application development.

Keywords: Dementia, Spontaneous Speech Conversation, Machine Learning, Classification

Contents

Listing of figures

Listing of tables

Listing of acronyms

NLP . . . . . Natural Language Processing

TTR . . . . . Type-token ratio

MATTR . . . . . Moving-average-type-token ratio

MTLD . . . . . Measure of lexical textual diversity

RAVDESS . . . . . Ryerson Audio-Visual Database of Emotional Speech and Song

MFCC . . . . . Mel-frequency cepstral coefficients

MFC . . . . . Mel-frequency cepstrum

STT . . . . . Speech to text

SVM . . . . . Support vector machine

CNN . . . . . Convolutional neural networks

POS . . . . . Part of speech

1 Introduction

1.1 DeepVibes - Company introduction

DeepVibes is a company that aims to improve the quality of life for people affected by Alzheimer's and dementia and also their beloved families by creating a voice archive with the best memories. The app stimulates and records conversations between family members and applies Artificial Intelligence to monitor the disease progression and enhance the therapy journey. It analyzes audio records to predict or keep track of cognitive abilities. Keeping the brain active with Reminiscence therapy is one of the most effective activities used during the early stages of the disease as a means to keep the brain active. In the late stages of dementia, relistening emotive moments or families' voices bring joy to the daily life of individuals. The 'smart play' function automatically selects and plays a mash-up of the most emotive memories. [1]

1.2 Topic introduction

Previous research has shown that automatic speech recognition can be used to distinguish between healthy people and people affected by dementia. [2] This project aims to add speakers' dementia level estimation to the existing DeepVibes mobile app. Dementia is a deadly and costly disease that has negative emotional, mental, and physical implications for those afflicted with the disease and their loved ones. Discovering dementia in an early stage can slow down

the disease with regular medical support.

In this project, I started by reading related literature to gather different symptoms to distinguish dementia patients from healthy people. Control group and dementia group data were extracted and pre-processed for the model training before using existing data collected from the DeepVibes application. Then, different functions and models were built to prepare a score for each of the items in the final score calculation. Summarizes every scoring item and a final score is presented in the dashboard at the end. The idea of the whole picture of my project is shown in Figure 1.1 below.

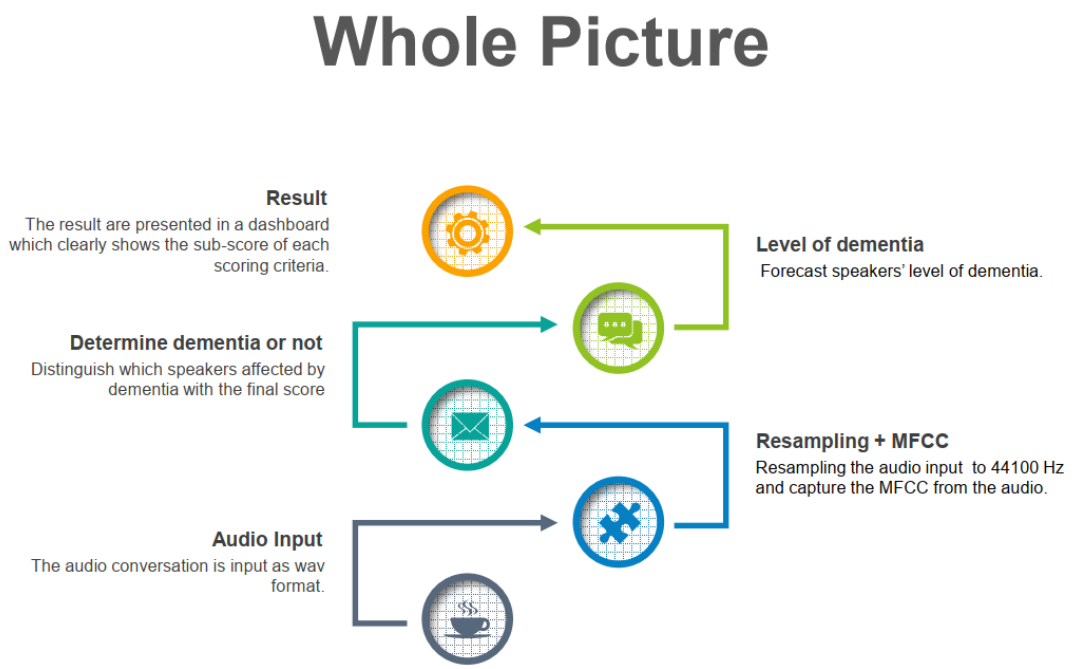


Figure 1.1: Whole picture of the project.

1.3 What is Dementia?

Dementia is the loss of cognitive functioning, including thinking, remembering, and reasoning, to such an extent that it interferes with a person's daily life and activities. Some people with dementia cannot control their emotions, and their personalities may change. Dementia ranges in severity from the mildest stage, when it is just beginning to affect a person's functioning, to the most severe stage, when the person must depend completely on others for basic activities of living.

People with dementia have problems with

- Memory loss
- Attention
- Communication
- Reasoning, judgment, and problem-solving
- Visual perception beyond typical age-related changes in vision

Signs that may point to dementia include

- Getting lost in a familiar neighborhood
- Using unusual words to refer to familiar objects
- Forgetting the name of a close family member or friend
- Forgetting old memories
- Not being able to complete tasks independently

[3]

1.4 Timeline

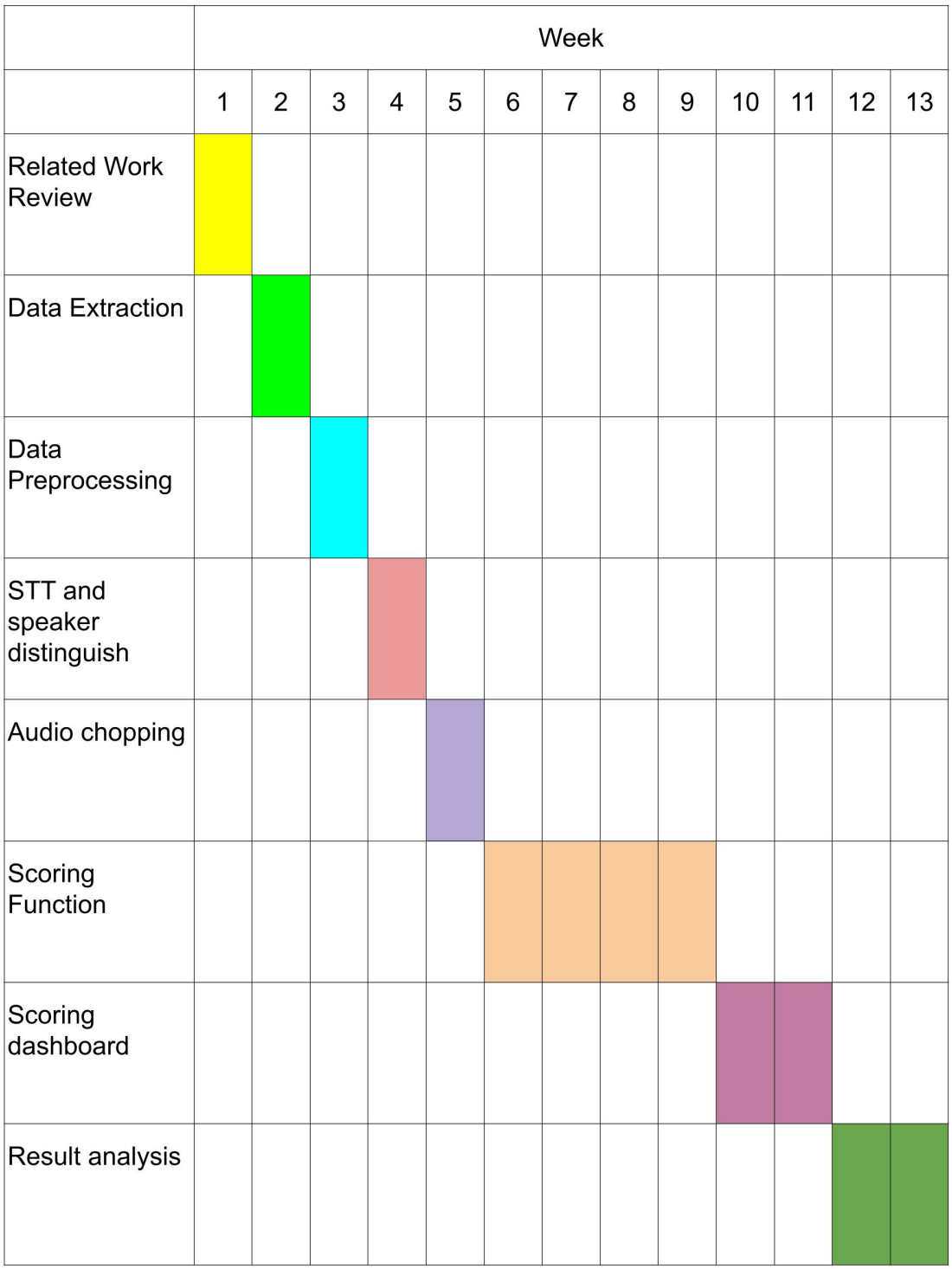


Figure 1.2: Timeline.

1.5 Technical Tools used

We tried to build our speech-to-text and speaker detection model and also compare other different companies' API resources. AssemblyAI API gave us the best accuracy at the end, hence we used it for our input data preparation. AssemblyAI's Speech-to-Text APIs support audio conversion to text by providing timestamp and speaker detection, which were the input for our models. [4]

All code was written in Python language. For the functions and model implementation, we used the open-source libraries listed in Table 1.1 below.

Functions/ Models	Libraries used
Emotion detection	Sklearn
Lexical richness	LexicalRichness
Speech fluency	myprosody
NLP	Nltk, textstat
Signal feature	librosa, Scipy
Voice quality	timbral models

Table 1.1: Python Libraries summary

2 Related Work

In the following section, we are going to refer to related papers which discuss factors that contribute to distinguishing people affected by dementia.

2.1 Natural Language Features

2.1.1 Question Rate

People affected by dementia are less willing to speak and thus we can predict that other people in the conversation will ask more open-end questions to initiate the people with dementia to talk more. The question words such as "'which"', "'what"' etc. are tagged and the total number of questions is calculated in each conversation. The question rate is computed by dividing the number of question words by the total number of words in the conversation. [5]

$$\text{Ratio of question rate in dementia group} < \text{Ratio of question rate in control group} \quad (2.1)$$

2.1.2 Nominal Phrases

Ratios of the number of nominal phrases out of the total number of words differed significantly between the dementia group and the control group. The previous paper shows that the control

group people produce more noun phrases and fewer pronoun phrases. In order words, the ratio of pronouns to noun phrases is also significantly higher in the dementia group than in the control group. [6]

$$\frac{\text{Ratio of Pronoun phrase in dementia group}}{\text{Ratio of noun phrase in dementia group}} > \frac{\text{Ratio of Pronoun phrase in control group}}{\text{Ratio of noun phrase in control group}} \tag{2.2}$$

2.1.3 Verb Phrases

The ratio of the number of verb phrases over the total number of words occurred less in the dementia group than in the control group. In addition, people affected by dementia produced less ratio of adverbs than the control group.

$$\text{Ratio of verb phrase in dementia group} < \text{Ratio of verb phrase in control group} \tag{2.3}$$

$$\text{Ratio of adverb phrase in dementia group} < \text{Ratio of adverb phrase in control group} \tag{2.4}$$

2.1.4 Clausal Domain

The clausal domain is the most complex structural unit distinguished here, as it includes noun phrases and verb phrases as proper parts. Within this domain, highly significant group differences emerged in patterns of clausal connectivity, both between the dementia group and the control group. Previous research measures of clausal connectivity showed a smaller proportion of usage of embedded adjunct clauses among people in the dementia group. The use of coordinated clauses appears less in the dementia group. [7]

$$\text{Ratio of conjunction in dementia group} < \text{Ratio of conjunction in control group} \tag{2.5}$$

2.1.5 Passive voice vs Active voice

Passive voice considers the form of auxiliary and presence of an agent. The passive voice makes the subject, the person or thing acted on or affected by the action represented by the verb. The active voice asserts that the person or thing represented by the grammatical subject performs the action represented by the verb. Research shows that people affected by dementia decrease the usage of sentences with passive voice. It also means that the ratio of the number of active voice sentences over the number of passive voice sentences of a text from people affected by dementia is higher than people from the control group. [8]

$$\frac{\text{No. of active voice sentence in dementia group}}{\text{No. of passive voice sentence in dementia group}} > \frac{\text{No. of active voice sentence in control group}}{\text{No. of passive voice sentence in control group}} \tag{2.6}$$

[9]

2.2 Lexical Diversity

Research shows that people affected by dementia showed a lower lexical diversity in their conversations than the people in the control group. Type-token ratio (TTR), Moving-average-type-token ratio (MATTR), and Measure of lexical diversity (MTLD) are common factors used for measuring lexical diversity, hence it is common that people affected by dementia have a lower score in TTR, MATTR, and MTLD, in other words, dementia groups treated with lower lexical diversity, are at higher risk of dementia.

The details of the elements used for measuring lexical diversity are discussed in the following subsection.

2.2.1 Type-token ratio (TTR)

The type-token ratio (TTR) is one of the measurement criteria of lexical diversity, which is obtained by dividing the number of different words (word types) by the total number of words (word tokens).

2.2.2 Moving-average-type-token ratio (MATTR)

MATTR calculates the lexical diversity of a sample using a moving window that estimates TTRs for each successive window of fixed length. Initially, a window length is selected—for example, 10 words—and the TTR for words 1–10 is estimated.

2.2.3 Measure of lexical textual diversity (MTLD)

MTLD reflects the average number of words in a row for which a certain TTR is maintained. To generate a score, MTLD calculates the TTR for increasingly longer parts of the sample. Every time the TTR drops below a predetermined value, a count (called the factor count) increases by 1, and the TTR evaluations are reset. The algorithm resumes from where it had stopped, and the same process is repeated until the last token of the language sample has been added and the TTR has been estimated. Then, the total number of words in the text is divided by the total factor count. Subsequently, the whole text in the language sample is reversed and another score of MTLD is estimated. The forward and the reversed MTLD scores are averaged to provide the final MTLD estimate.

MTLD and MATTR Variables associated with MTLD and MATTR were found to be very strong indicators of the lexical diversity of a language sample in absolute terms. Scores that were generated using MATTR and MTLD were influenced only by the content factors and did not demonstrate systematic effects from construct-irrelevant sources. Therefore, holding all else constant, MTLD and MATTR provide a more accurate reflection of the lexical diversity for the testing samples.

[10]

2.3 Speech Fluency

2.3.1 Rate of Pauses in Utterances

Studies found that people affected by dementia tend to stop more in the middle of sentences to think about what to say next. These pauses can be used as a marker for cognitive problems. The total number of occurrences of pauses inside each utterance is extracted. Utterances are assumed to be segments where the subjects talk without having very long pauses or an interruption. The pause rate is computed by dividing the number of pauses by the number of segments.

The rate for the whole recording is computed by averaging the pause rate computed in each utterance.

2.3.2 Disfluency

Other studies include approaches from conversation analysis that look at more interaction aspects, showing that disfluencies such as fillers and repairs, and purely nonverbal features such as inter-speaker silence, can be key features of conversation of people affected by dementia.

2.3.3 Rate of speech

Previous research studied the temporal organization of speech in people affected by dementia is usually slower. Thus, the rate of speech of the people affected by dementia is expected to be higher. Speech rate and articulation rate are both defined as "the number of output units per unit of time" often expressed in words per minute (wpm). While speech rate includes pause intervals while articulation rate does not.

$$\text{Speech Rate} = \frac{\text{Number of words}}{\text{Time with pause intervals}} \tag{2.7}$$

$$\text{Articulation Rate} = \frac{\text{Number of words}}{\text{Time without pause intervals}} \tag{2.8}$$

[11]

2.4 Emotion Detection

Previous Research gave an aggregated score to indicate the status of people affected by dementia have a higher tendency of more negative mood than positive mood. It is commonly shown that people with dementia commonly show fear, anger, and disgust. Other studies have also shown that depression, anxiety, and post-traumatic stress disorder are also linked with dementia. People with dementia often experience changes in their emotional responses. They may have less control over their feelings and how to express them. For example, someone may overreact to things, have rapid mood changes, or feel irritable. They may also appear unusually distant or uninterested in things. [12]

2.5 Voice Quality

Dementia weakens muscles caused by problems in the brain. Spasms can affect dementia speaker throat muscles and make their voice hoarse or weak, which affects their voice quality. Voice quality is that component of speech that gives the primary distinction to a given speaker's voice when pitch and loudness are excluded. It has been defined as the characteristic auditory coloring of an individual's voice, derived from a variety of laryngeal and supralaryngeal features and running continuously through the individual's speech. The following summarizes part of the elements which can measure voice quality. [13]

- Roughness: Roughness correlates to how noticeable or annoying a sound is as heard by the human ear. More specifically, roughness is a hearing sensation related to loudness modulations at frequencies too high to discern separately, such as modulation frequencies greater than 30 Hz.
- Booming:A boom is a very loud, deep sound that echoes.
- Hardness: It's a combination of the loudness and "harshness" of a sound and the pickup range and pattern of your microphone.
- Sharpness: Sharpness considers the relationship between the loudness of high frequency components and total loudness, and roughness evaluates modulation characteristics.
- Warmth: "Warmth" and "brilliance" refer to the reverberation time at low frequencies relative to that at higher frequencies.

[14]

3 Dataset

Forecasts of dementia are integrated with a variety of factors. In such cases, using the same dataset does not generalize well in all the factors. Moreover, because we collected a large-scale dataset, the dementia bank dataset, at a very late stage, the work must go on, hence we used alternative datasets which are suitable for each factor during the model-building process. The data we used in this project to train and/ or test each specific scoring criteria with different datasets are described below.

3.1 Dementia Bank

DementiaBank is a medical domain task, the largest publicly available that contains dementia speakers' audio recordings and transcripts. It contains 117 people diagnosed with Alzheimer's Disease, and 93 healthy people, reading a description of an image, and the task is to classify these groups. This release contains only the audio part of this dataset, without the text features. It contains a publicly available dataset that contains audio recordings and transcripts of participants (people with dementia and healthy controls) describing the Cookie Theft picture. Participants were asked to describe all events in the image. Patients were asked to perform various tasks; for example, in the "Boston Cookie Theft" description task, patients were shown in Figure 3.1 and asked to describe what they saw. Other tasks include the 'Recall Test' in which patients were asked to recall attributes of a story they had been told previously. Each transcript

in DementiaBank comes with automatic morphosyntactic analysis, such as standard part-of-speech tagging, description of tense, and repetition markers. [15]

Note that these features are generic, automatically-extracted linguistic properties and are not AD-specific. We broke each transcript into individual utterances to use as data samples.

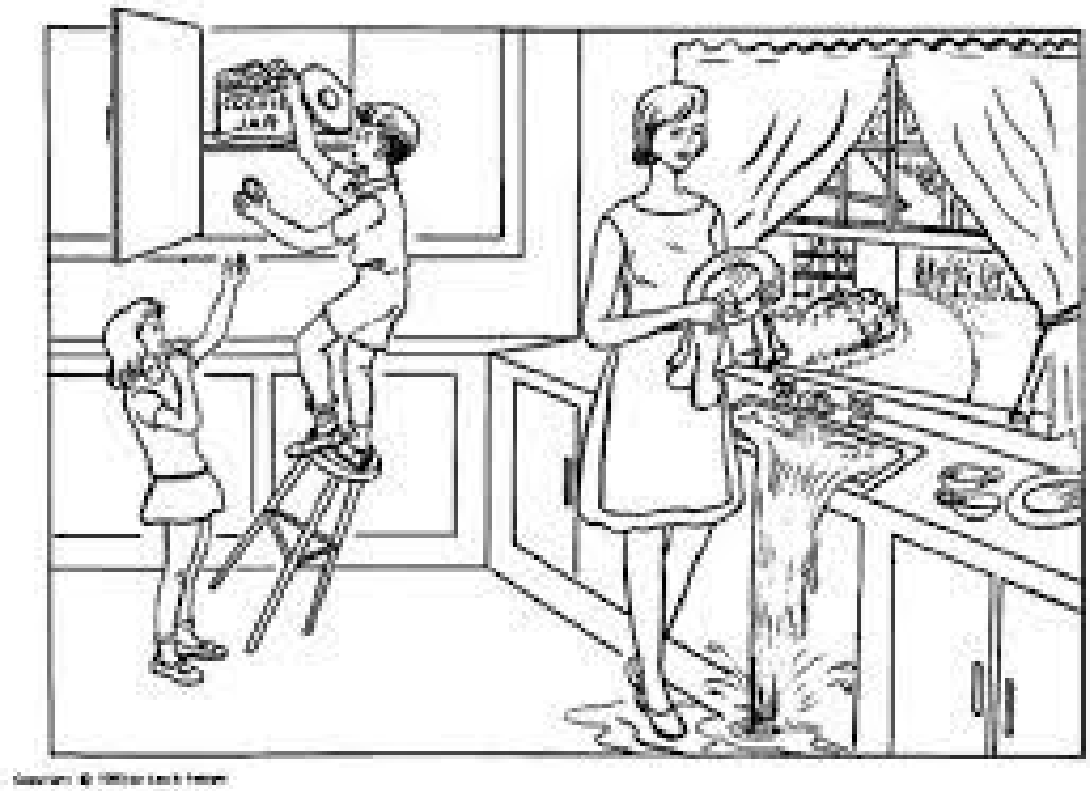


Figure 3.1: Dementia Bank - Cookie Theft picture.

3.2 RAVDESS

The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprised, and disgust expressions and the song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings was each rated 10 times on emotional validity, intensity, and gen-

uineness. Ratings were provided by 247 individuals who were characteristic untrained research participants from North America. A further set of 72 participants provided test-retest data. All recordings are made freely available under a Creative Commons license and can be downloaded at https://doi.org/10.5281/zenodo.1188976. [16]

3.3 Audio used

Access to DementiaBank is password protected and restricted to members of the DementiaBank consortium group, thus it took time for us to collect it at the end. We do not have any other data resources at the beginning, but we need audio with a transcript to test our functions built. Similar types of audio were extracted from youtube and used for testing before we received the dementia bank dataset. The testing results of those audios and/ or dementia bank datasets are also included in the following sections. The audios we used are listed in Table 3.1 below.

Audio Title	Source
Federico at phone with Memory Lane	Company Audio
Mom, Alzheimers, and a Conversation	Youtube
Interviewing My 100 Year Old Great Grandmother	Youtube
Interviewing my grandpa	Youtube

Table 3.1: Audio source

4

Exploratory data analysis

To better understand the data for pre-processing as well as building the networks, we did data exploration and visualization for both DementiaBank and RAVDESS datasets. The results are attached below.

Audio signal samples were visualized in the time-series domain in Figure 4.1 below. An audio signal can be represented in the time domain which indicates the amplitude of the signal at each point in time, or more precisely at each time when the signal was observed.

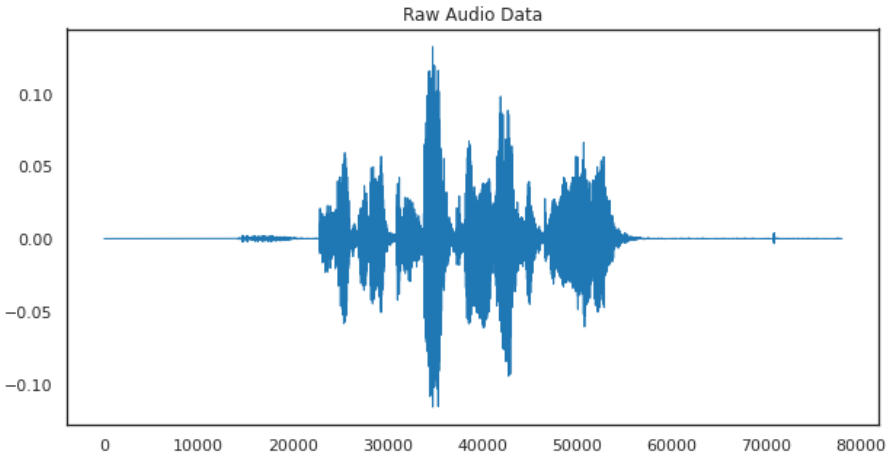


Figure 4.1: waveplot.

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies

with time. Not only can one see whether there is more or less energy at, for example, 2 Hz vs 10 Hz, but one can also see how energy levels vary over time. The spectrogram graph of one of the audio is plotted in Figure 4.2. [17]

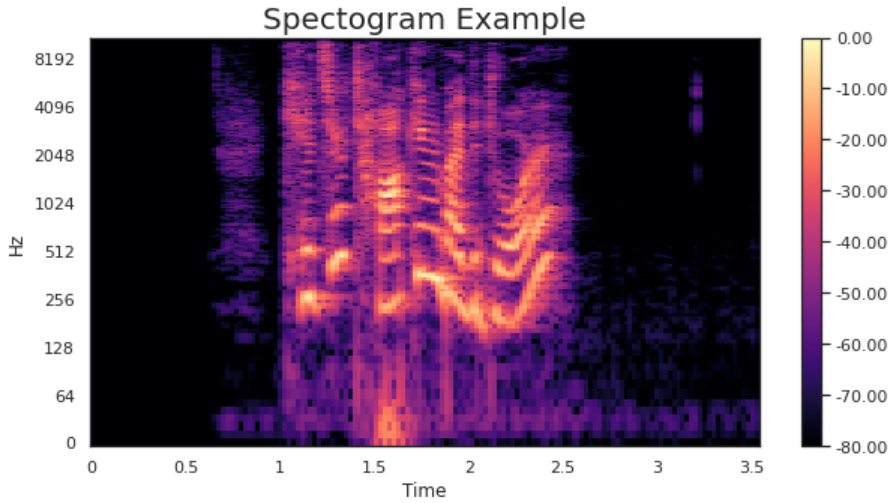


Figure 4.2: Spectrogram.

The Mel spectrogram is used to provide our models with sound information similar to what a human would perceive. The raw audio waveforms are passed through filter banks to obtain the Mel spectrogram. The Mel spectrogram graph of one of the audio is plotted in Figure 4.3. [18]

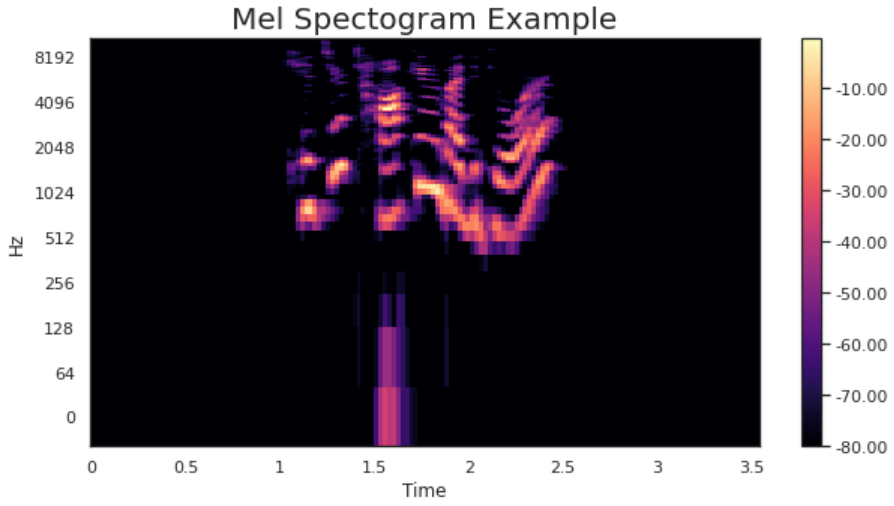


Figure 4.3: Mel spectrogram.

In sound processing, the Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. The MFCC heatmap of one of the audio is plotted in Figure 4.4. [19]

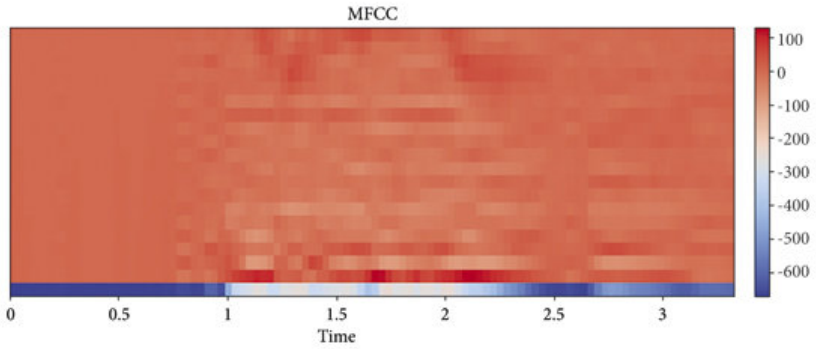


Figure 4.4: MFCC.

4.1 DementiaBank

The frequency of utterances of both dementia group and control group are plotted in Figure 4.5.

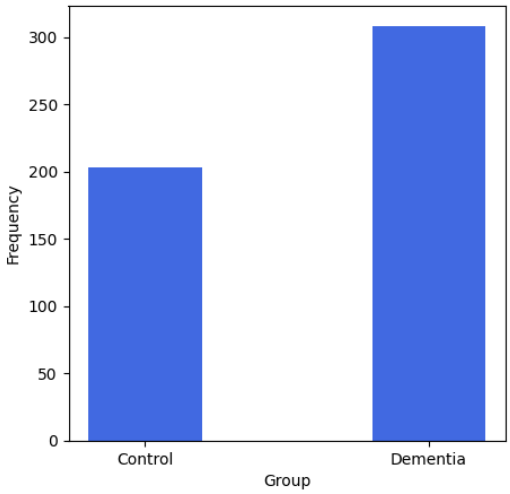


Figure 4.5: Number of recordings of each group.

The lengths of utterances range from 20s to 270s. From the Figure 4.6 shows that the duration of conversation in dementia group is usually longer than the the control group.

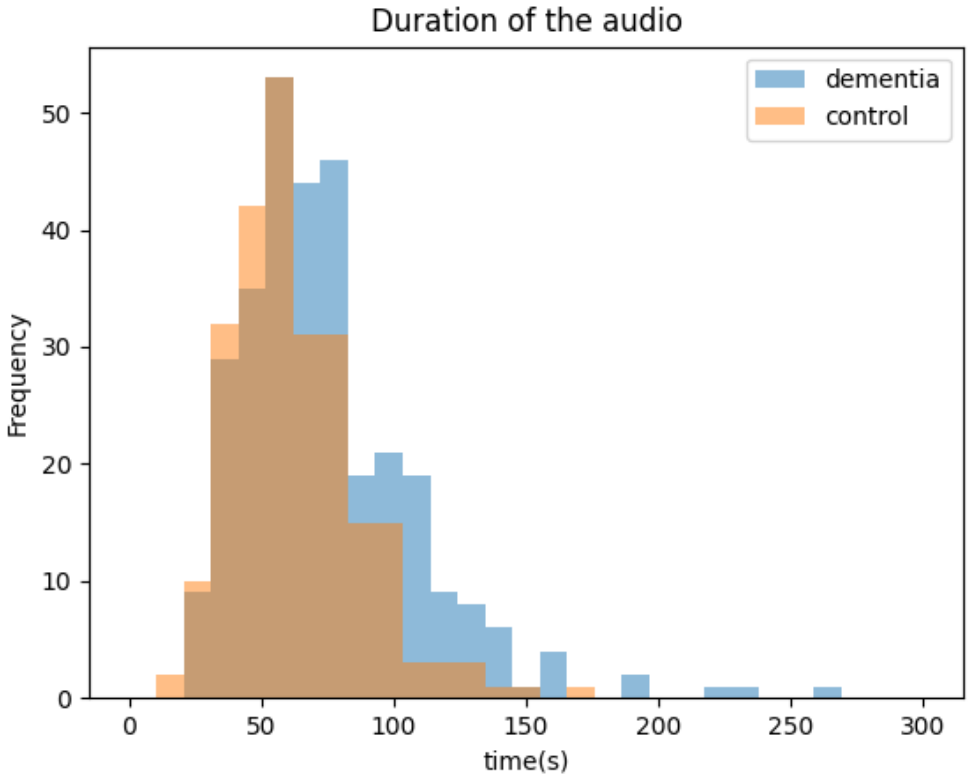


Figure 4.6: Duration Distribution.

In Figure 4.7 shows the age distribution of both dementia and control groups. There are more elderly people suffering from cognitive impairments than younger people.

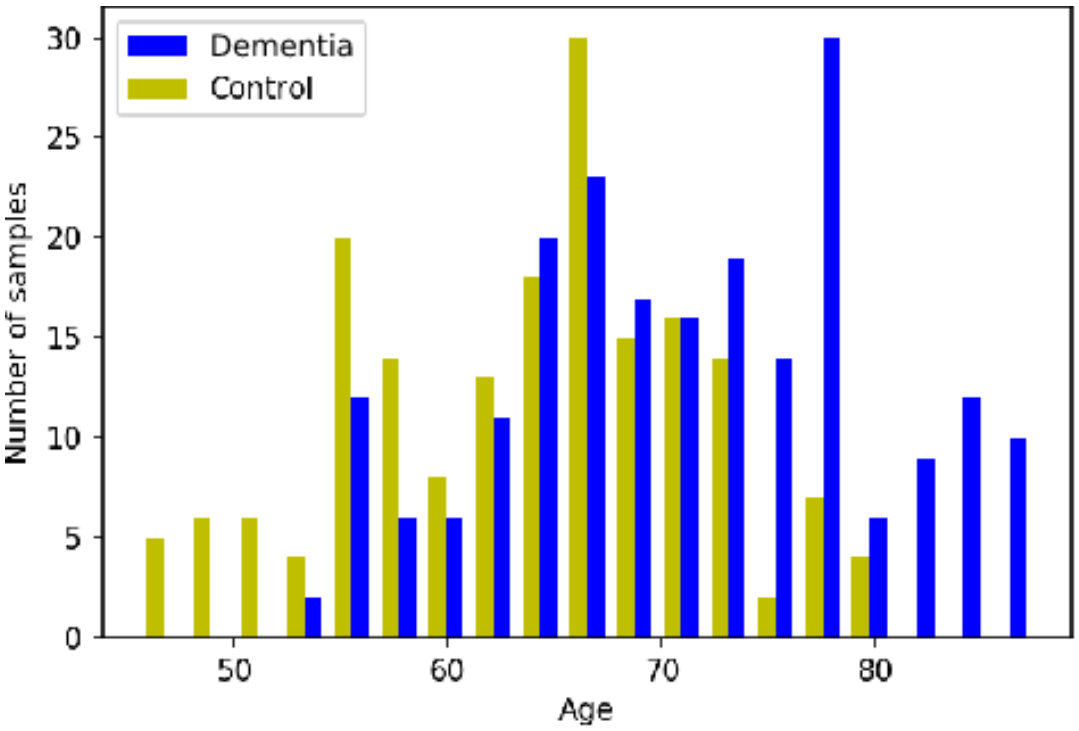


Figure 4.7: Age Distribution.

4.2 RAVDESS

RAVDESS is a gender-balanced set of validated speeches and songs that consists of eight emotions of 24 professional actors speaking similar statements in a North American accent. Figure 4.8 shows that angry, calm, disgust, fear, happy and sad emotion classes constituted 192 audio files each. The surprise emotion had 184 files and the neutral emotion had the lowest number of audio files 96.

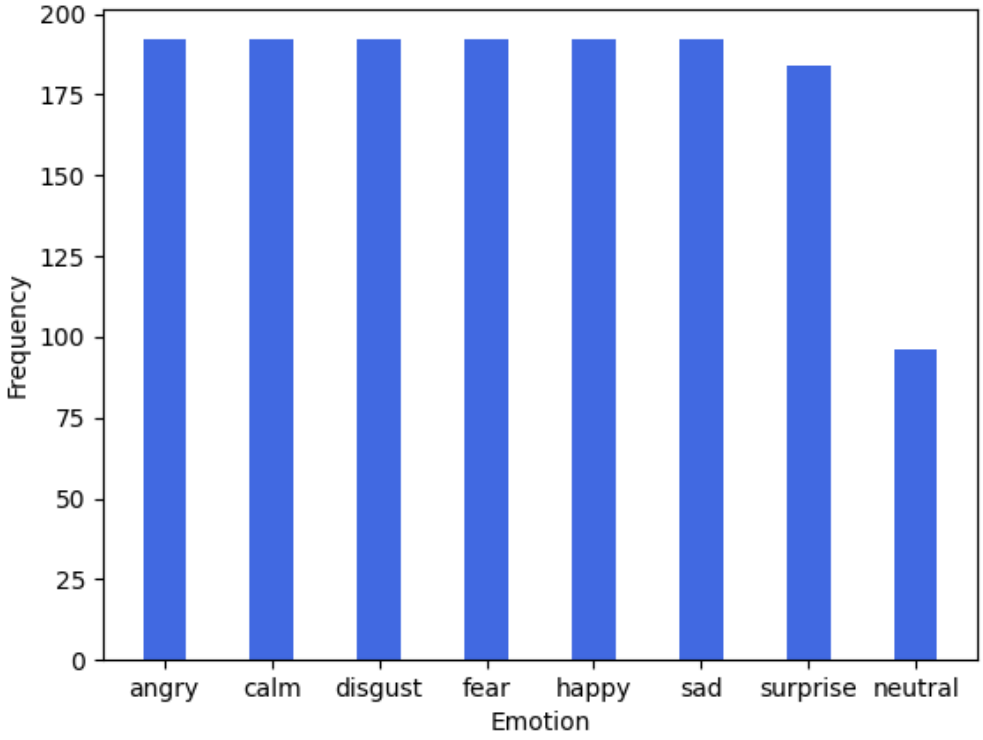


Figure 4.8: Emotion Distribution.

5 Data Preprocessing

5.1 Resampling

Since the DementiaBank data we downloaded are in wav-format files, they should be sampled by a certain sampling rate. The sample rate is the number of samples of a sound that are taken per second to represent the event digitally. The more samples taken per second, the more accurate the digital representation of the sound can be. Amplitude is the maximum extent of a vibration or oscillation, measured from the position of equilibrium, hence the sampling rate can be shown in a wave plot with different amplitudes along time series.

Resampling, in terms of audio files, is also known as Sample Rate Conversion. This is done when you need to convert a digital audio file from a given sample rate into a different sample rate. Sample rate is the number of samples of audio carried per second, measured in Hz. One of the open source libraries, myprosody, requires the sample rate of the audio input as 44100 Hz. The audio was resampled to 44100 Hz to fit its requirement.

5.2 Speech to text (STT) and Speaker diarization

To obtain the transcript of the audio and speakers diarization act as the input for the model training. We tried to build our own speech-to-text and speaker detection model and also compare other different companies' API resources. The comparison of the API resources are shown

in Table 5.1 below. By using a few youtube videos with transcripts as testing materials, comparing the API conversion result and the actual transcripts to calculate the accuracy (number of correct words/ total number of words). After comparing the accuracy and cost, AssemblyAI API gave us the best package at the end, hence we used it for our input data preparation.

API resources
AssemblyAI – Speech to Text API
Deepgram
Google cloud – Speech to text
IBM – Watson speech to text

Table 5.1: STT API resources

We used both open source libraries Silero and vosk for building our own speech to text and speaker diarization model. Silero is a speech-To-Text model that provides enterprise grade STT in a compact form-factor for several commonly spoken languages. The models consume a normalized audio in the form of samples (i.e. without any pre-processing except for normalization to -1 ... 1) and output frames with token probabilities. It provides a decoder utility for simplicity. Vosk is an offline open source speech recognition toolkit.

5.3 Capture Mel-frequency cepstrum (MFCC)

MFCCs aim to extract the features from the audio signal. In general, 12 to 20 cepstral coefficients are typically optimal for speech analysis. From each segment, we extracted 20 features. The main advantage of using MFCCs techniques is because of less complexity and accurate results. To prepare the input dataset of the dementia classification model, we captured the information of frequency based on the Mel-filterbank. The window size is 25ms, the frame shift is 10ms and MFCC features size is 20. In this program, the sampling rate is 16000 Hz, thus the input size for the following models is 99 in the time dimension and 20 in the frequency dimension. In sound processing, the Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. [19]

5.4 Audio Segmentation into specific parts

It is better for us to compare the differences, such as mood, between different sections of time in the conversation. We segmented the audio into a specific number of parts and split them according to the speakers involved in the conversation. Three main steps are involved. A 9 minutes conversation example with two speakers is going to be used to explain in detail as follows.

Figure 5.1: Example

Step 1: Find the segmenting timestamp
According to the specific number of parts, we found the timestamp should be segmented in the audio. In the example, we are going to segment it into 3 parts. The timestamp are shown as blue dotted lines in our example.

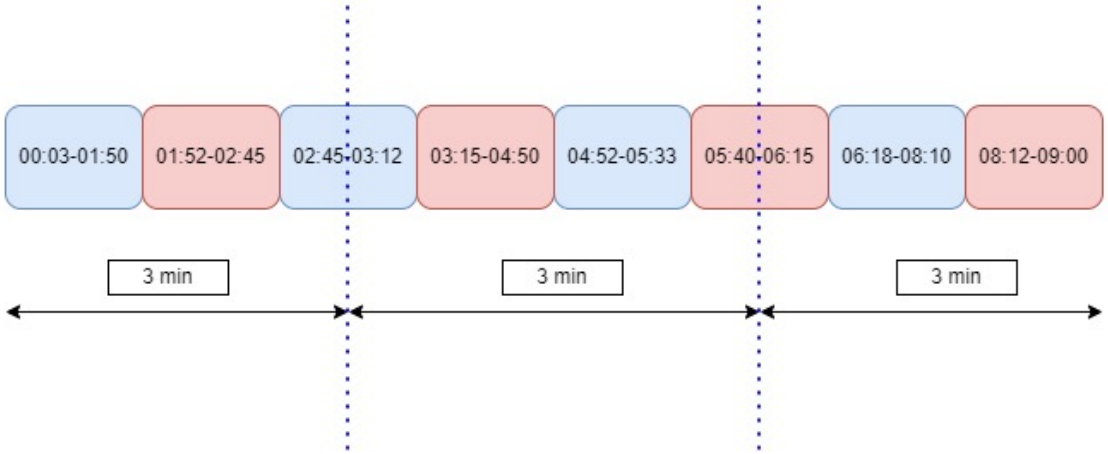


Figure 5.2: Example

Step 2: Find the nearest timestamp of the sentence end and the start time of the next sentence according to the timestamp found in step 1
To avoid we segment the audio in the middle of the sentence. Hence, we are going to find the timestamp of the last sentence involved in that section. Moreover, we also find the nearest start time of the sentence in the next section. The new timestamp is shown as yellow lines in our example.

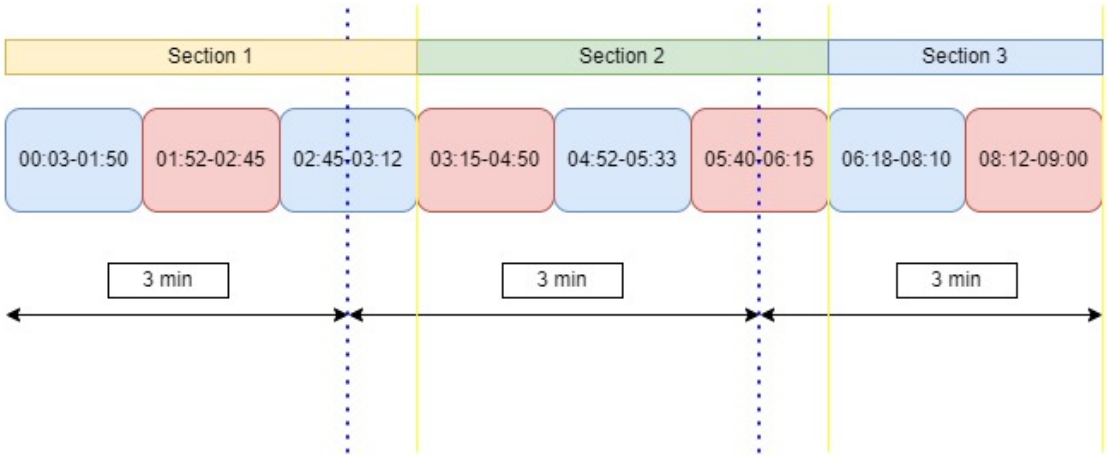


Figure 5.3: Example

Step 3: Segment the audio parts according to the speakers involved
We split each of the segmented audio according to the speakers involved in that section.

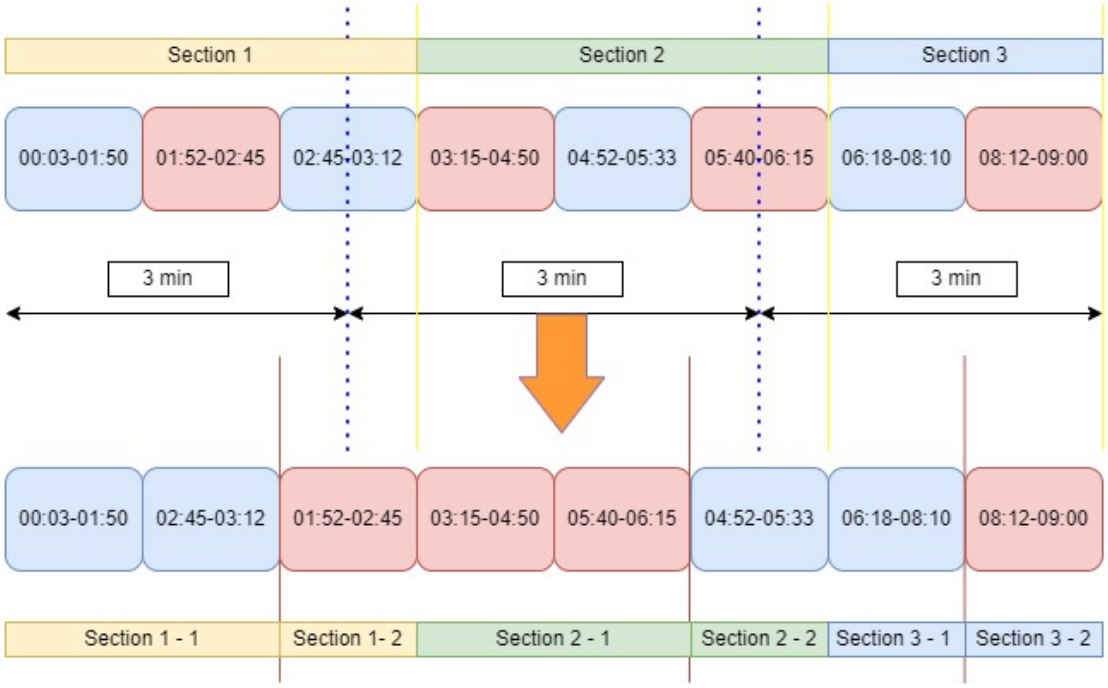


Figure 5.4: Example

In our example, we segmented the audio into 6 audio parts at the end. The details of the parts are as follows:

Audio parts id	Timestamp
Section 1 - 1	00:03 - 01:50; 02:45 - 03:12
Section 1 - 2	01:52 - 02:45
Section 2 - 1	0315 - 04:50; 05:40 - 06:15
Section 2 - 2	04:52 - 05:33
Section 3 - 1	06:18 - 08:10
Section 3 - 2	08:12 - 09:00

Table 5.2: Segmented audio parts

5.5 Remove samples with insufficient information

A histogram in Figure 5.5 below shows the distribution of the duration of recordings. Audio samples less than 10 seconds provide insufficient information for the training model, in which audio samples of less than 10 second were removed from the dataset. We can notice that most of the audio samples lasted around 10 second. Hence, most of the samples were kept in the dataset.

Dementia:

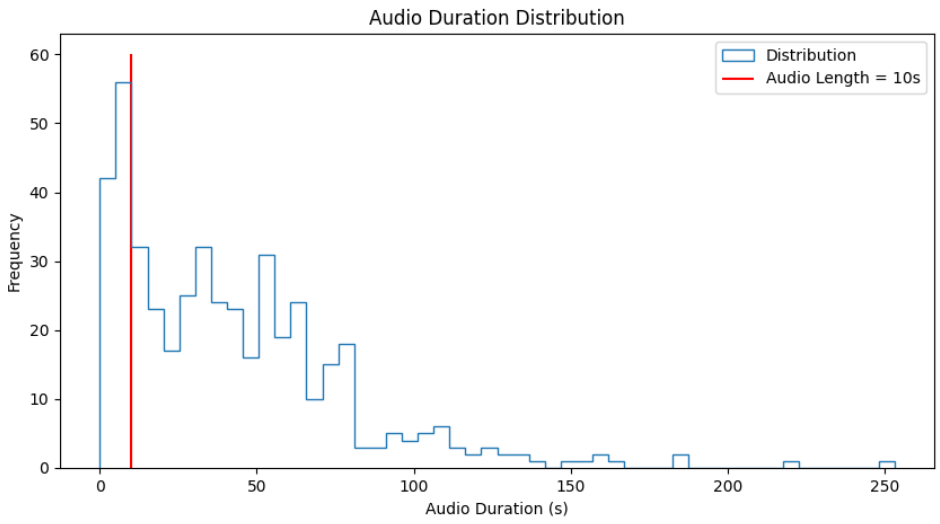


Figure 5.5: Example

Control:

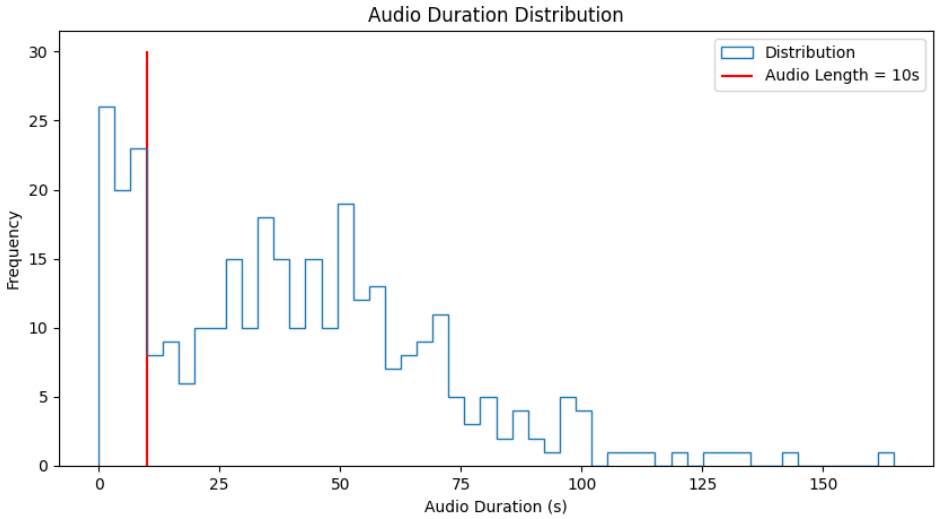


Figure 5.6: Example

6 Hypothesis

From the previous related work result shown, people with dementia usually have a higher chance to have the following characteristics:

Items	Dementia group characteristics
Lexical Diversity	Decrease
Lexical Repetition	Increase
Speech fluency	Decrease
Word Class distribution	Smaller proportion of noun and verb, More proportion of pronoun
Voice Quality	Decrease
Emotion	Negative mood
Complexity	Decrease

Table 6.1: Characteristics of dementia group

In this project, we aim to conclude similar results in the final dashboard as described in Table 6.1 above. Moreover, we are going to try other elements to find more factors for classifying people affected by dementia.

7

Learning Framework

The entire pipeline of this program consists of 2 main parts: data preprocessing, and models for predicting people affected by dementia and the level of dementia through conversational audio.

In the following section, I am not only going to build functions mentioned in the previous paper that can show differences between dementia and control groups but also try other factors to see any other elements that can be used to classify people affected by dementia. Different types of data, including audio time series, text, and MFCC are used as input. ”load” function of the python library ”Librosa” is used to load an audio file as a floating point time series. ”mfcc” function in the ”feature” class of the same library is used to capture the MFCC sequence of the audio. Assembly AI Speech-to-text API is used to prepare the transcript of the audio. The Table 7.1 below shows the input of the datatype used in the functions built in this section:

Function	Input Data
”nlp_result”	Text
”lexical_richness_result”	Text
”speech_fluency_result”	Audio
”signal_result”	Audio
”voice_quality_result”	Audio
”complexity_result”	Text
”emotion”	MFCC

Table 7.1: Input Datatype

7.1 Natural Language Features

As previous research showed that people affected by dementia suffered from language deficits, including but not limited to sentence comprehension on deficits, and tend to use more pronouns and fewer noun phrases and verb phrases.

Natural Language Toolkit suite of tools and python libraries were applied to extract Natural Language Processing (NLP) features for each text, and a set of seven features was extracted. Each text was tokenized by breaking it up into words and punctuation marks and the set of unique words, consisting of its vocabulary, was further extracted. Part-of-speech (POS) tagger is then used to assign grammatical information to each word of the sentence. The following Table 7.2 shows a detailed explanation of the POS tagger.

NLTK POS Tagger	POS	Example
NN	Noun	Book
VBG	Verb	Eat
PRP	Personal pronoun	Him
PRP$	Possessive pronoun	Mine
JJ	Adjective	Beautiful
RB	Adverb	Slowly
CC	Conjunction	And

Table 7.2: POS tagger

Before calculating the ratio of each part of speech in the text, the total number of words in the text is found. The following Table 7.3 shows the formulas for calculating the ratio of each part of speech in the text:

Item	Formula
Ratio of noun	$\text{Ratio of noun} = \frac{\text{Number of noun}}{\text{Total number of words}}$
Ratio of verb	$\text{Ratio of verb} = \frac{\text{Number of verb}}{\text{Total number of words}}$
Ratio of pronoun	$\text{Ratio of pronoun} = \frac{\text{Number of pronoun}}{\text{Total number of words}}$
Ratio of adjective	$\text{Ratio of adjective} = \frac{\text{Number of adjective}}{\text{Total number of words}}$
Ratio of adverb	$\text{Ratio of adverb} = \frac{\text{Number of adverb}}{\text{Total number of words}}$
Ratio of conjunction	$\text{Ratio of conjunction} = \frac{\text{Number of conjunction}}{\text{Total number of words}}$

Table 7.3: Formulas - POS

Research also showed that people affected by dementia pronounced less passive voice. By using the same library "sent_tokenize" function to find the total number of sentences in the text to further calculate the ratio of active voice and ratio of passive voice in the text. The following Table 7.4 shows the details of both equations:

To detect passive voice in sentences, we used the spacy module to tag each token in the sentence, then build a classifier to classify it as passive or active based on conventional grammatical rules.

If a clause has all of the following, then it is in the passive voice:

- A form of an auxiliary verb (usually be or get)
- The past participle of a transitive verb
- No direct object
- The subject of the verb phrase is the entity undergoing an action or having its state changed

Item	Formula
Ratio of active voice	$\text{Ratio of active voice} = \frac{\text{Number of active voice sentences}}{\text{Total number of sentences}}$
Ratio of passive voice	$\text{Ratio of passive voice} = \frac{\text{Number of passive voice sentences}}{\text{Total number of sentences}}$

Table 7.4: Formulas - voice of sentence

A person with dementia may find it difficult to initiate a conversation or an activity themselves. Hence we can predict that the one without dementia will initiate the conversation more with question-ending sentences. In other words, people with dementia in the conversation have a lower ratio of sentences with question endings.

$$\text{Question ratio} = \frac{\text{Number of question endings sentences}}{\text{Total number of sentences}} \tag{7.1}$$

"nlp_result" function is going to conclude the combination of POS characteristics and features of the sentences of the conversation.

7.2 Lexical Diversity

People affected by dementia demonstrated a lower lexical diversity in their speaking than the control group of people. Lexical richness is used interchangeably with lexical diversity, lexical variation, lexical density, and vocabulary richness and is measured by a wide variety of indices. "lexical_richness_result" function summarizes the lexical diversity characteristic of the text.

Python library "LexicalRichness" is used to measure the lexical diversity of the text, including type-token ratio (TTR), the measure of textual lexical diversity (MLTD), and moving-average type-token ratio (MATTR). The equation of the formula in the library refers to the paper "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment" written by McCarthy and Jarvis. Before calling the LexicalRichness library, we cleaned the data by calling the python library audio_preprocess. It helps to transfer the text into tokens.

The TTR index gives the ratio between the number of 'types' and 'tokens' in a text. The number of 'types' is the number of unique word forms; the number of 'tokens' is the total number of words. A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite. The range falls between a theoretical 0 (infinite repetition of a single type) and 1 (the complete non-repetition found in a concordance).

MATTR and MTLD are the varieties of TTR. If the length of tokens of the text is more than 5, the window size used in the MATTR calculation is set to 5. Otherwise, the window size is set to be the length of tokens of the text. It is the moving window that estimates TTRs for each successive window of fixed length. A window length of 5 words is selected and the TTR for words 1–5 is estimated. Then the TTR is estimated for words 2–6, then 3–7, and so on to the end of the text. For the final score, the estimated TTRs are averaged. The measure

of textual lexical diversity is computed as the mean length of sequential words in a text that maintains a minimum threshold TTR score. Iterates over words until TTR scores fall below a threshold, then increase the factor counter by 1 and start over. McCarthy and Jarvis' research recommends a factor threshold in the range of 0.660 to 0.750. The library default threshold "0.72" is used for measuring the MLTD value in our function.

7.3 Speech Fluency

Previous research shows that people affected by dementia usually speak at a slower rate of speech and with more pauses involved in their conversation. The "Speech fluency result" function is built to summarize the prosody of audio. Prosody refers to intonation, stress pattern, loudness variations, pausing, and rhythm. We express prosody mainly by varying pitch, loudness, and duration. In our "speech fluency result" function, we focus on measuring the rate of speech, articulation, and pauses of the audio.

MyProsody is a Python library for measuring these acoustic features of speech. An acoustic algorithm breaks recorded utterances into 48 kHz, 32-bit sampling rate, and bit depth respectively to detect syllable boundaries, fundamental frequency contours, and formants. We use its built-in functions to measure: Rate of speech, Articulation, and rate of pauses. A higher rate of speech and articulation with a lower rate of pauses is classified as a lower chance to be detected in people with dementia. The equation of rate of speech, articulation ,and rate of pauses are listed below:

$$\text{Rate of speech} = \frac{\text{Number of words}}{\text{Duration of an audio}} \tag{7.2}$$

$$\text{Articulation} = \frac{\text{Number of words}}{\text{Duration of an audio without pauses}} \tag{7.3}$$

$$\text{Rate of pauses} = \frac{\text{Number of Pauses}}{\text{Duration of an audio}} \tag{7.4}$$

7.4 Signal Features

People with dementia with a lower voice or voice disorder. People with weak voices will usually have low volume and may have a change in pitch. Voice volume refers to how loud or soft

a speaker's voice level is, the lower the voice volume detect, the higher indicator showing the possibility of the weak voice of the speaker.

Volume is an acoustic feature that is correlated to the samples' amplitudes within a frame. To define volume quantitatively, we can employ 10 times the 10-based logarithm of the sum of sample square method to compute the volume of a given frame:

$$volume = 10 * \log_{10} \sum_{i=1}^{n} s_i^2$$

This method requires floating-point computations, but it is linearly correlated to our perception of the loudness of audio signals. Quantity computing is also referred to as the "log energy" in the unit of decibels. The decibel (dB) is a logarithmic unit used to measure the sound level. It is also widely used in electronics, signals, and communication. The dB is a logarithmic way of describing a ratio.

Volume is referred to as loudness, and loudness is measured in sones. The loudness level is measured in phons, whereas the pitch is measured in mels. Pitch refers to the degree of highness or lowness of a tone. Auto-correlation function (ACF) is used for pitch tracking. This is a time-domain method that estimates the similarity between a frame s(i),i=0,...,n−1, and its delayed version via the auto-correlation function:

$$acf(\tau) = \sum_{i=0}^{n\ 1\ \tau} s(i)s(i+\tau)$$

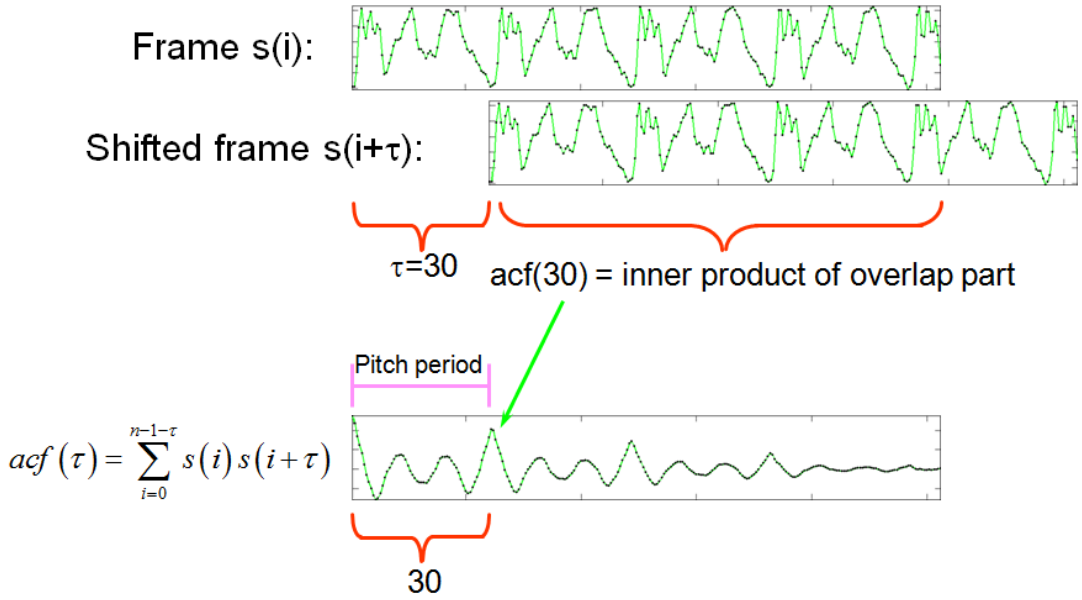


Figure 7.1: Pitch illustration.

In other words, we shift the delayed version n times and compute the inner product of the overlapped parts to obtain n values of ACF. Eventually by measuring the differentiation of the pitch track in audio to estimate the degree of changes of tone in the audio.

Variance, skewness, kurtosis and mean are applied to the measurement of volume and pitch to analyze the characterization of the location and variability of the signal features of audio. Variance is a measure of variability, it is calculated by taking the average of squared deviations from the mean. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. Kurtosis is a measure of whether the data are nearly-tailed or light-tailed relative to normal distribution. Mean is the average of set of values. The "signal_result" function concluded the variance, skewness, kurtosis and mean of pitch and volume of the testing audio.

7.5 Voice Quality

Dementia can include symptoms of anxiety and depression, sleep disorders, agitation, and verbal or physical aggression, which may result in behavioral disturbance, with verbal aggression and physical aggression. Screaming is common among people who have dementia, especially for those who live in nursing centers, tends to occur along with the development of other related agitated behaviors, and has been attributed to a variety of causes, including vulnerability,

suffering, sense of loss, loneliness, physical pain (including hunger), clinical depression, and, more abstractly, as a call for help or need to fill a void with sound or emotion. Thus, screaming, yelling and crying sound is more possible to show in a dementia group of people, which affect the voice quality of their conversation.

Python library "timbral_model" is used to build the function for measuring voice quality, the distribution contains scripts for predicting eight timbral characteristics: hardness, depth, brightness, roughness, warmth, sharpness, booming, and reverberation. I chose to return the factors that can describe more the intensity of anger, including booming score, hardness score, sharpness score, and warmth score in the voice quality function for summarizing the voice quality of the speaker. The output values are returned ranging from 0 to 100, which is the mean score of the audio, based on the regression models trained on subjective ratings.

Roughness correlates to how noticeable or annoying a sound is heard by the human ear which is a hearing sensation related to loudness modulations at frequencies too high to discern separately, such as modulation frequencies greater than 30 Hz. While sharpness considers the relationship between the loudness of high-frequency components and total loudness, roughness evaluates modulation characteristics. To prevent double counting similar scoring factors are included in the function to result in a double counting effect. Roughness and sharpness are measuring similar items, but sharpness also considers the loudness of the voice, hence I chose to only include sharpness but not roughness in the voice quality function.

In the "voice quality result" function, the booming score, hardness score, sharpness score, and warmth score of the audio are returned finally. The description of each item is listed in the table below:

- Booming score: A boom is a very loud, deep sound that echoes.
- Hardness score: A combination of the loudness and harshness of a sound
- Sharpness score: Sharpness is a hearing sensation related to frequency and independent of loudness.
- Warmth score: A warm voice has lower overtones. Breathy and air leaks through the vocal cords and sounds like a sigh.

7.6 Complexity

People with dementia usually difficult to find words to communicate and tends to speak with simple wording, which can predict the content of the conversation with a lower complexity score. The "complexity_result" function summarizes the complexity of the conversation, including readability, the ratio of polysyllabic wordings in the text ,and the average height of the parse tree of the sentences in the text.

7.6.1 Readability

Readability is a measure of how easy a piece of text is to read. The level of complexity of the text, its familiarity, legibility, and typography all feed into how readable your text is. I used the python library "textstat" to build a function that measures the readability of the text. Textstat is a library to calculate statistics from the text. It helps determine readability, complexity, and grade level. "readability_result" function is going to conclude the complexity of the text.

Different methods can be used to calculate the readability of the text, including flesh reading, flesch kincaid grade, gunning fog, smog index, automated readability index, coleman liau index, linsear write formula, dale chall readability score, text standard, spache readability and mcalpine eflaw. The description of each readability scoring method is described below:

- Flesch reading: It designed to indicate how difficult a passage in English is to understand.
- Flesch kincaid grade: The Flesch–Kincaid readability tests are readability tests designed to indicate how difficult a passage in English is to understand. There are two tests: the Flesch Reading-Ease, and the Flesch–Kincaid Grade Level.
- Gunning fog: It is a weighted average of the number of words per sentence, and the number of long words per word. An interpretation is that the text can be understood by someone who left full-time education at a later age than the index.
- Smog index: SMOG stands for "Simple Measure of Gobbledygook". It is a readability framework. It measures how many years of education the average person needs to have to understand a text.
- Automated readability index: It produces an approximate representation of the US grade level needed to comprehend the text.
- Coleman liau index: A readability test designed by Meri Coleman and T. L. Liau to gauge the understandability of a text.

- Linsear write formula: The Lensear Write readability formula is a text-based formula. It scores monosyllabic words and strong verbs.
- Dale chall readability score: It is a readability test that provides a numeric gauge of the comprehension difficulty that readers come upon when reading a text. It uses a list of 3000 words that groups of fourth-grade American students could reliably understand, considering any word not on that list to be difficult.
- Spache readability: A readability scoring method for primary-grade reading materials.
- Mcalpine eflaw: It is a function scoring the readability according to the number of words in the document, the number of mini-words (3 or fewer characters), and the number of sentences.
- Overall readability: It takes into account all the above scoring methods and then takes the average.

7.6.2 Ratio of polysyllabic words

Apart from readability, the ratio of polysyllabic words used is also a measurement of the complexity of the text. Monosyllabic is a word with only one syllable or a person who uses short, abrupt words in conversation. The word cat is an example of a monosyllabic word. Polysyllabic words have many syllables and we consider that words with more than three syllables are polysyllabic words. For example, the word librarian is polysyllabic, but the word book is not. Function "poly_mono_ratio" is used to measure whether the speaker tends to speak polysyllabic words or not.

$$\text{poly mono ratio} = \frac{\text{Number of polysyllabic words}}{\text{Number of monosyllabic words}} \tag{7.5}$$

7.6.3 Average parse tree height of the sentences

A parse tree is an ordered, rooted tree that represents the syntactic structure of a string according to some context-free grammar. The more complex a sentence is, the higher the height of the parse tree. Average tree height is used to represent the mean of the tree height of the sentences in a text.

The characteristics of the parse tree are listed as follows:

- Each word in the sentence is a node in the graph.
- The tree has a root, a node from which the rest of the tree flows. This is the main verb of a sentence.
- As the arrows suggest, the edges are directional. If an edge goes from one node to the other, the first node is the "parent" and the second node is the "child". In parse trees, each node (except the root) has one parent and can have zero or more children.
- Each node has a label. Separately, each edge has a label. These are two different sets of labels.

For example "Tom lives in a big tent with Jerry", the text and its corresponding encoding tree with a height of 3. The green node represents the root node of the encoding tree, and the yellow points refer to other non-leaf nodes.

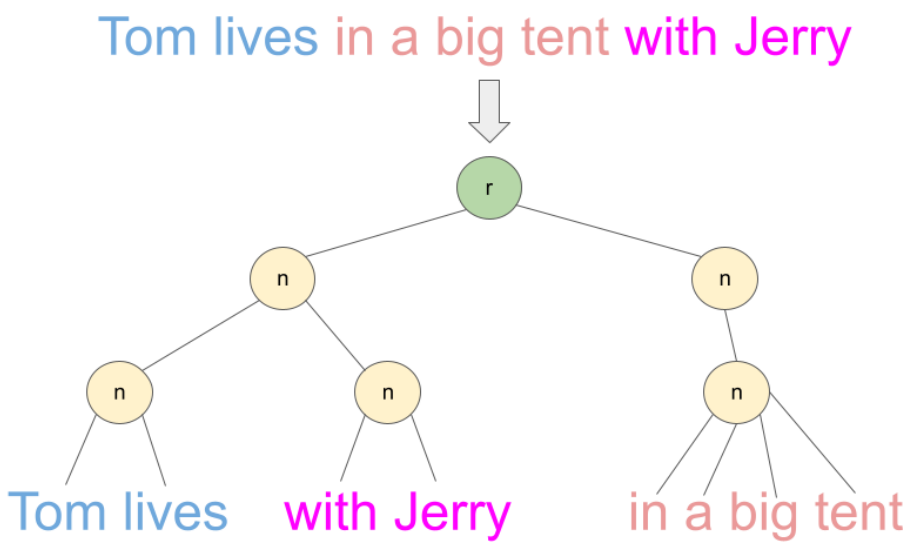


Figure 7.2: Example.

7.7 Emotion Recognition

As mentioned in the voice quality subsection, people with dementia usually suffer from more negative emotions and have more emotion changes from time to time. The objective of the

emotion recognition function is to apply machine learning models to the problem of classifying emotion on audio. The best CNN architecture and SVM model are going to find in this subsection to find a model that can perform the best emotion classification by comparing their accuracy.

Since the dementia bank data do not label the emotion of the audio, I used the RAVDESS dataset instead to train the model. It contains 2452 audio files, with 12 male speakers and 12 Female speakers, the lexical features (vocabulary) of the utterances are kept constant by speaking only 2 statements of equal lengths in 8 different emotions by all speakers. We converted the audio samples to MFCC used as input data features for both SVM and CNN model training. By using MFCC for feature extraction are also suggested to improve the speaker recognition efficiency as it's extraction is based on the human peripheral auditory system.

We first trained our data in the normal 2D-CNN model, the model requires large number of operations and parameters in order to be trained. However, the results show low accuracy rate, 65%, and we tried to find more suitable approach. We adopted the SVM model to overcome to drawbacks of using normal CNN. SVM provided the higher accuracy in the dataset. As the kernel trick in the SVM model can help to handle nonlinear input spaces, which can convert non-separable problems to separable problems by adding more dimension, in order to build a more accurate classifier. In this subsection, we showed the experiments to verify that normal CNN is not a nice approach for emotion detection. We found a simple SVM model works well in our best model with almost 90% accuracy for classifying three emotions, including happy, neutral and sad.

7.7.1 SVM

SVM plots the training vectors in high-dimensional feature space, and label each vector with its class. A hyperplane is drawn between the training vectors that maximizes the distance between the different classes. The hyperplane is determined through a kernel function, which is given as input to the classification software. The kernel function may be linear, polynomial, radial basis, or sigmoid. The shape of the hyperplane is generated by the kernel function, though many experiments select the polynomial kernel as optimal.

The details of the training flow chart of SVM model is shown in Figure 7.3 below.

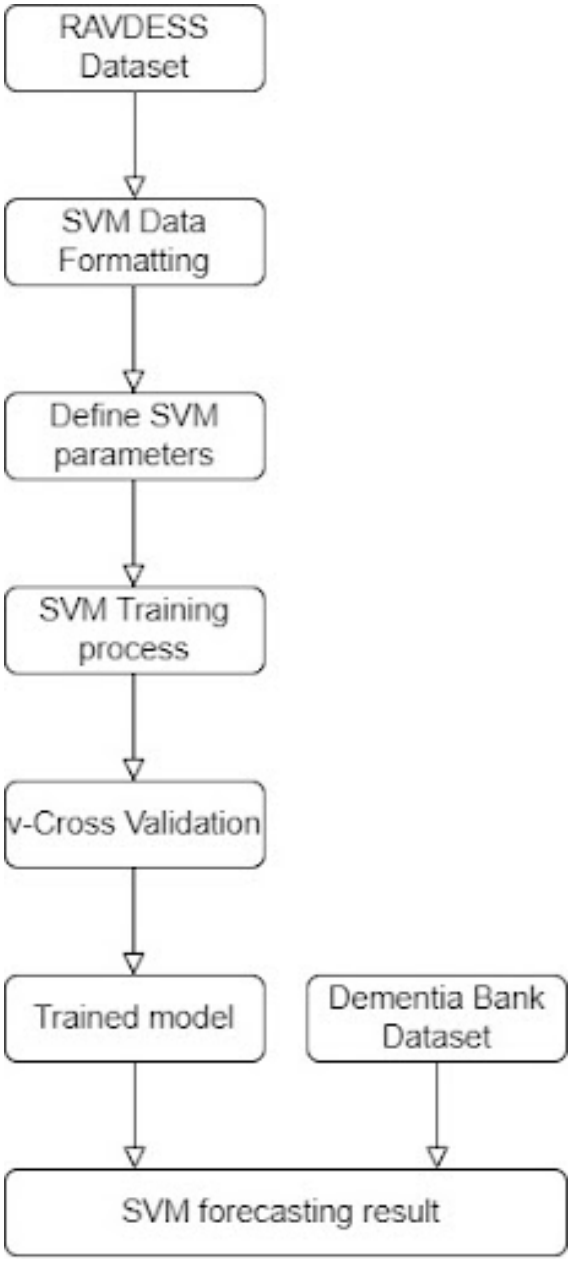


Figure 7.3: SVM flow chart.

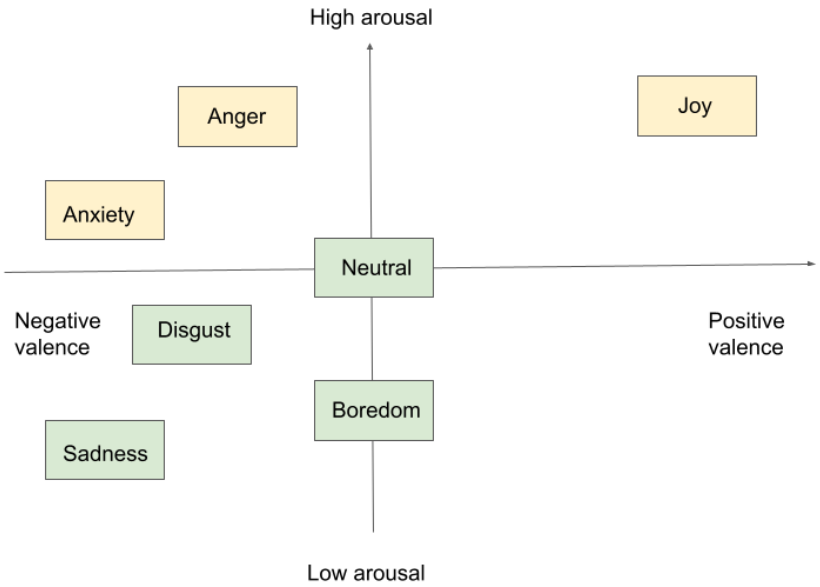
As the major purpose of the function aims to detect the degree of negative emotions of the speaker. To strike a balance between the number of emotion combinations, computation resources and accuracy, different numbers of combinations of emotions are tested, including 3 emotions, 5 emotions and 8 emotions. We used the subset of the dataset of the respected combination of emotions included to train the SVM model for classifying 3 emotions and 5

emotions. The set of emotion combinations is listed below:

- 3 emotions: happy, sad, neutral
- 5 emotions: happy, angry, neutral, sad, and fearful
- 8 emotions: happy, angry, neutral, sad, fearful, disgust, calm, surprised

SVM classifier models used MFCC features as model input. After considering the computational resources and accuracy, 3 emotions is used these embeddings to train a speech emotion recognizer with a Support Vector Machine (SVM) implemented in sklearn with an 'RBF' kernel and a regularization parameter of C = 0.001.

The emotion sets were selected based on the two-dimensional space of arousal and valence. 'Happy' was selected as it falls in the highest range of both arousal and positive valence. In contrast, 'sad' was selected as it falls in the lowest range of arousal and the highest range of negative valence. 'Neutral' was selected as this emotion is balanced between happy and sad in a two-dimensional space of arousal and valence. For the second set, five emotions were selected for evaluation: happy, angry, neutral, sad, and fearful.



Accuracy

Emotion Class	Best parameters after tuning	Accuracy Unstandardized	Accuracy Standardized	Accuracy Standardized + PCA
3 emotions	{"C":100, "Gamma": 0.0001, "Kernel": rbf	86.70%	89.59%	89.02%(PCA=40)
5 emotions	{"C":100, "Gamma": 0.0001, "Kernel": rbf	81.25%	85.07%	86.81%(PCA=59)
8 emotions	{"C":100, "Gamma": 0.0001, "Kernel": rbf	73.10%	82.43%	78.46%(PCA=50)

Confusion Matrix

Emotion	Happy	Neutral	Sad
Happy	**91.23%**	0%	8.77%
Neutral	5.17%	**91.38%**	3.45%
Sad	8.62%	5.17%	**86.21%**

Emotion	Angry	Fearful	Happy	Neutral	Sad
Angry	**86.21%**	1.72%	10.34%	0%	1.72%
Fearful	1.75%	**94.74%**	1.75%	0%	1.75%
Happy	3.45%	8.62%	**84.48%**	3.45%	0%
Neutral	0%	1.72%	3.45%	**91.38%**	3.45%
Sad	3.51%	12.28%	3.51%	3.51%	**77.19%**

Emotion	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
Angry	**77.19%**	0.00%	10.53%	0.00%	7.02%	1.75%	0.00%	3.51%
Calm	0.00%	**89.47%**	0.00%	0.00%	0.00%	3.51%	7.02%	0.00%
Disgust	6.90%	1.72%	**75.86%**	0.00%	3.45%	1.72%	3.45%	6.90%
Fearful	1.72%	0.00%	1.72%	**77.59%**	6.90%	1.72%	8.62%	1.72%
Happy	5.17%	1.72%	13.79%	6.90%	**58.62%**	3.45%	10.34%	0.00%
Neutral	0.00%	0.00%	3.51%	3.51%	0.00%	**92.98%**	0.00%	0.00%
Sad	1.72%	0.00%	3.45%	3.45%	3.45%	5.17%	**75.86%**	6.90%
Surprised	0.00%	0.00%	1.72%	6.90%	3.45%	1.72%	3.45%	**82.76%**

7.8 Convolutional neural networks (CNN)

CNNs are widely used for any deep learning solution with images as data. Thanks to our audio data can be transformed into a time-frequency representation through MFCC, CNNs consider this kind of data representation as a grayscale image to learn. There are many hyperparameters in the CNNs, like filter striding, pooling sizes, kernel sizes, etc, so the number of multiplications of models can be changed to meet the computational constraints by adjusting the hyperparameters.

Since CNN models have many parameters that can be changed in training, we first used hyperparameter optimization to find the best hyperparameter combination of the model. Our new CNN model consists of a standard deep 2D convolutional neural network of 4 2D convolutional layers with the number of filters 32, 32, 32, 32, and kernel sizes are (4, 10), (4, 10), (4, 10), (4,10). We added a 2D maxpooling layer and 4 dropout layers with a 0.2 dropout rate to avoid overfitting. Finally, we used relu to produce the results. Figure 7.4 shows loss graphs for this model, and it keeps decreasing slowly as can be seen in the graph. This suggests that the network keeps improving. However, we only achieved an accuracy of around 65%, which is far lower than that SVM model.

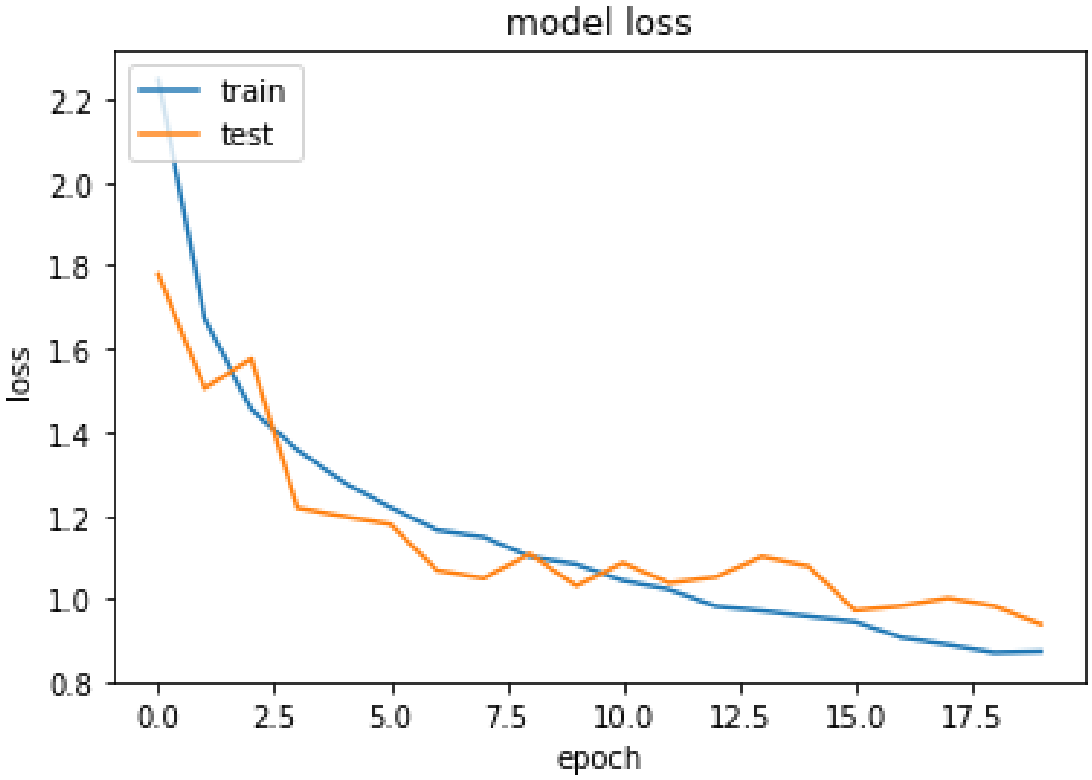


Figure 7.4: Loss graph.

8 Finding and Result

The following section is going to show the distribution plot between the dementia group and control group of each scoring factor we tried. The factors which have significant differences between groups will be selected as factors for calculating the level of dementia of the speaker's diagnosis as dementia. The higher differences in the factors between groups, the higher weighting of the factors will be assigned in the equation for the level of dementia calculation.

8.1 Mann-Whitney U Test

Mann-Whitney U test is used to test whether dementia and control groups are likely to derive from the same population. The null hypothesis for this test is that the two groups have the same distribution, while the alternative hypothesis is that one group has larger or smaller values than the others. The Mann–Whitney U test is agnostic to outliers and concentrates on the center of the distribution.

8.1.1 Procedure of Mann-Whitney U test

- 1. Combine all data points and rank them
- 2. Compute $U1 = R1 - n1(n1 + 1)/2$, where $R1$ is the sum of the ranks for data points in the first group and $n1$ is the number of points in the first group.
- 3. Compute $U2$ similarity for the second group

- 4. The test statistic is given by $stat = min(U1, U2)$

8.1.2 Mann-Whitney result interpretation

A significance level of 0.05 is chosen for the p-value. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference. If the p-value is less than or equal to the significance level, the decision is to reject the null hypothesis. The lower the p-value of the scoring factors, means the higher weighting contributes for distinguish people affected by dementia.

[20]

8.2 Result

The following subsection shows the distribution of dementia and the control group of each scoring factor by using histograms.

8.2.1 Natural Language Features

Polysyllabic, monosyllabic ratio

- Mann-Whitney result: [U statistic: 41258.0, p-value: 0.014552976616715313]

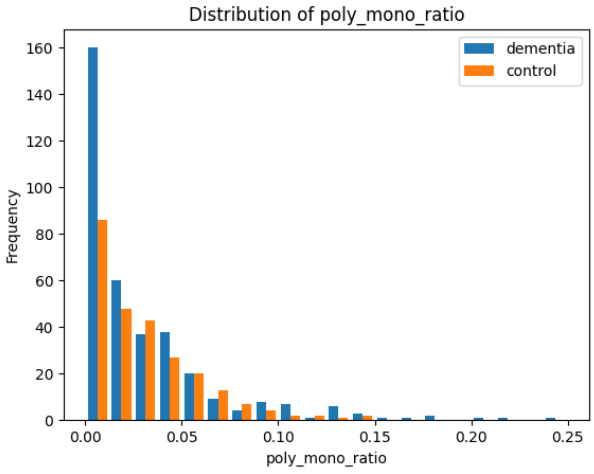


Figure 8.1: Polysyllabic monosyllabic ratio distribution.

Question ratio

- Mann-Whitney result: [U statistic: 33024.0, p-value: 1.4488201146178206e-09]

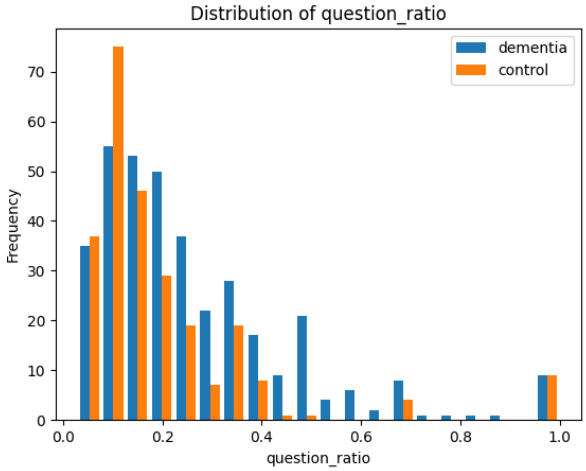


Figure 8.2: Question distribution.

Passive voice, Active voice ratio

- Mann-Whitney result: [U statistic: 45701.5, p-value: 0.45016880935555728]

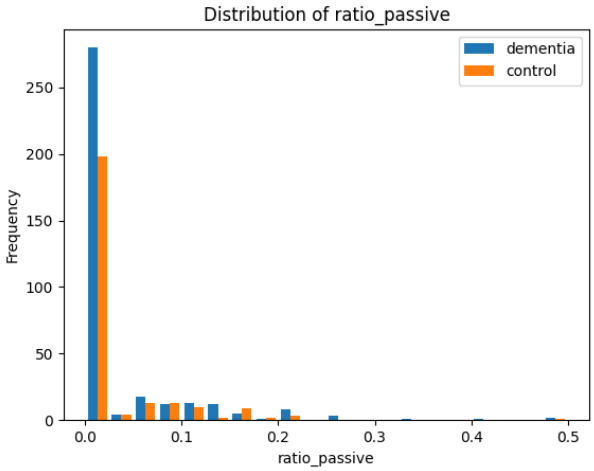


Figure 8.3: Passive voice, Active voice ratio distribution.

Ratio of conjunction

- Mann-Whitney result: [U statistic: 38760.0, p-value: 0.0003287406461775372 6]

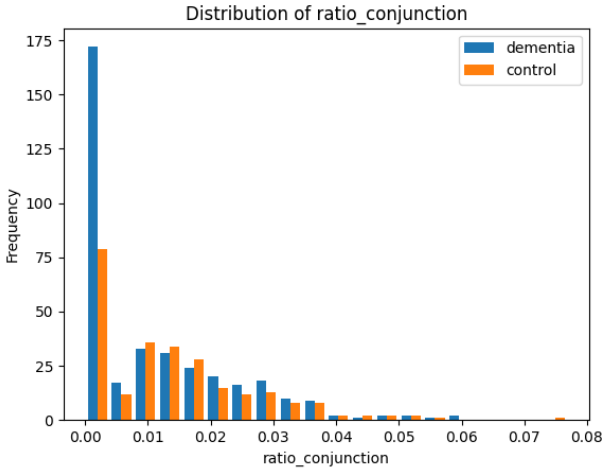


Figure 8.4: Ratio of conjunction distribution.

Ratio of noun

- Mann-Whitney result: [U statistic: 32394.5, p-value: 2.4637176656533 27e-10]

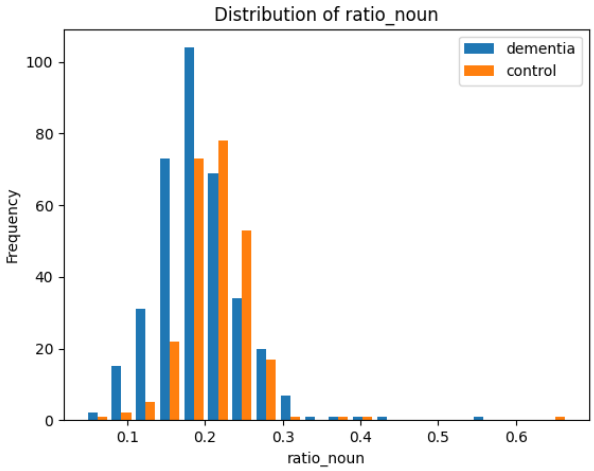


Figure 8.5: Ratio of noun distribution.

Ratio of personal pronoun

- Mann-Whitney result: [U statistic: 44302.0, p-value: 0.2304108122348981 8]

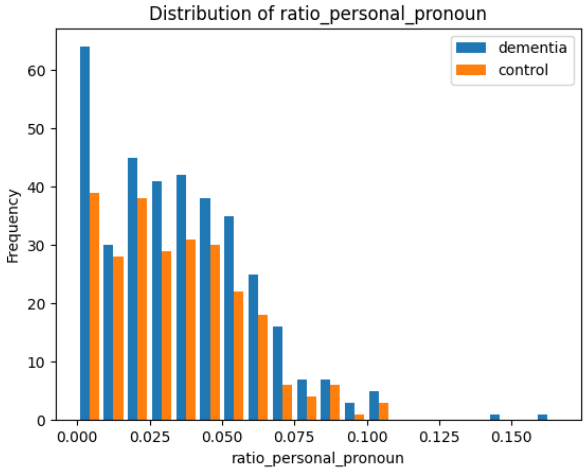


Figure 8.6: Ratio of personal pronoun distribution.

Ratio of pronoun

- Mann-Whitney result: [U statistic: 33636.0, p-value: 8.044274815499266e-09]

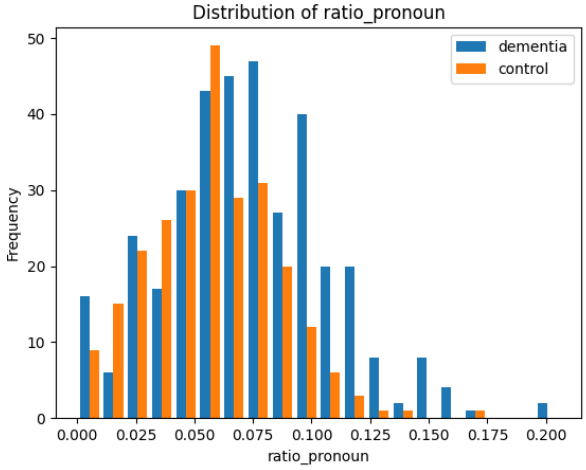


Figure 8.7: Ratio of pronoun distribution.

Ratio of verb

- Mann-Whitney result: [U statistic: 35260.5, p-value: 4.766447136117611 4e-07]

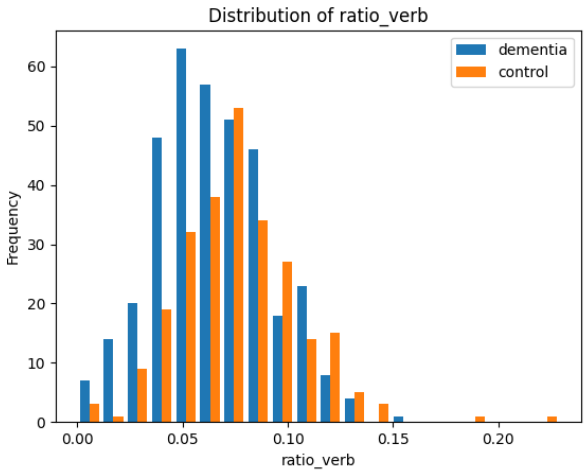


Figure 8.8: Ratio of verb distribution.

8.2.2 Lexical Diversity

mattr

- Mann-Whitney result: [U statistic: 42759.5, p-value: 0.07397751 66088744]

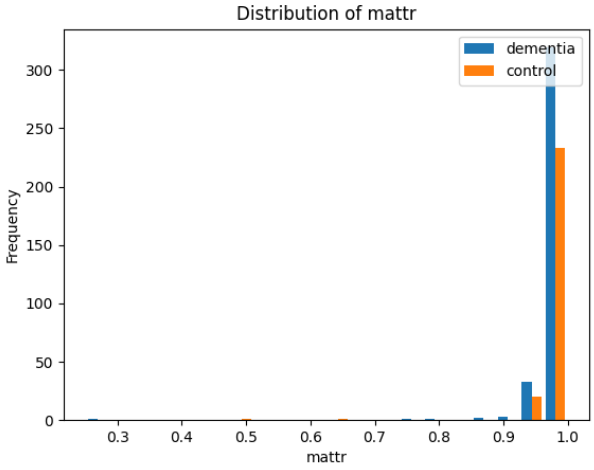


Figure 8.9: mattr distribution.

MLTD

- Mann-Whitney result: [U statistic: 42306.5, p-value: 0.04894743318910905]

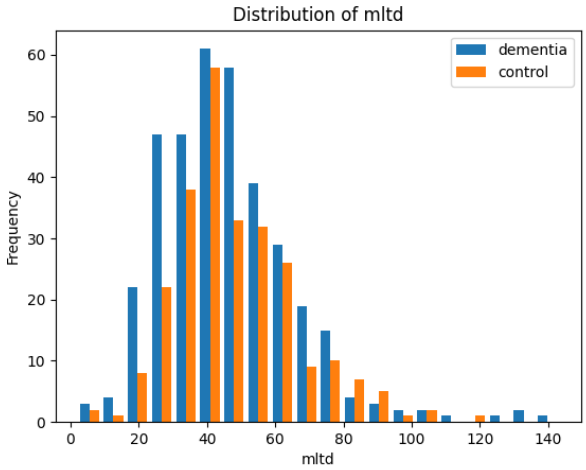


Figure 8.10: mltd distribution.

TTR

- Mann-Whitney result: [U statistic: 45832.0, p-value: 0.4875970359619717]

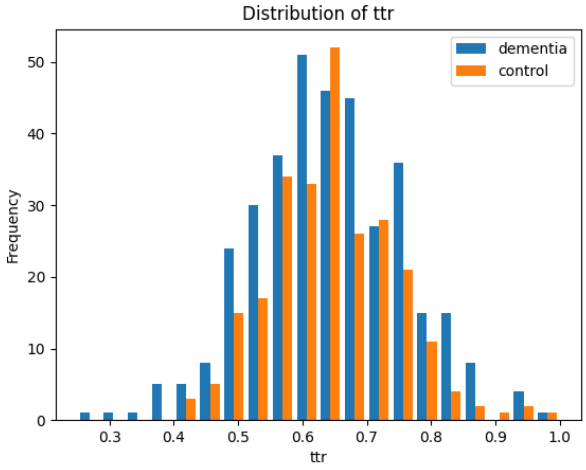


Figure 8.11: ttr distribution.

8.2.3 Speech Fluency

Articulation

- Mann-Whitney result: [U statistic: 32212.5, p-value: 0.10123459751174885]

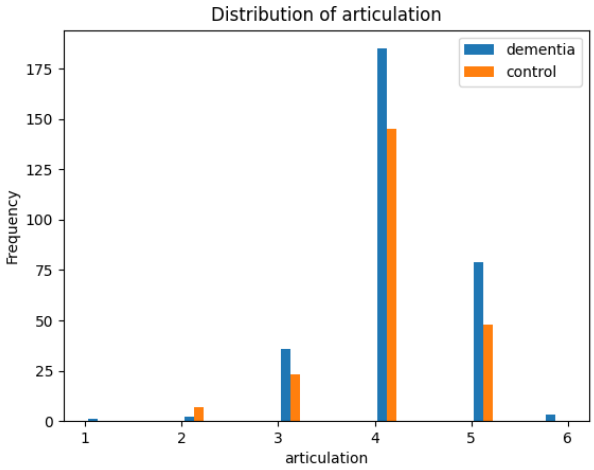


Figure 8.12: Articulation distribution.

Pauses

- Mann-Whitney result: [U statistic: 31487.0, p-value: 0.06478450631703 06]

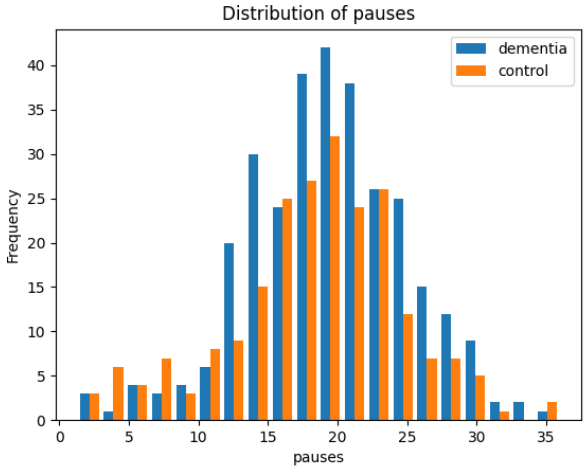


Figure 8.13: Pauses distribution.

Rate of speech

- Mann-Whitney result: [U statistic: 29995.5, p-value: 0.0045479351966737 46]

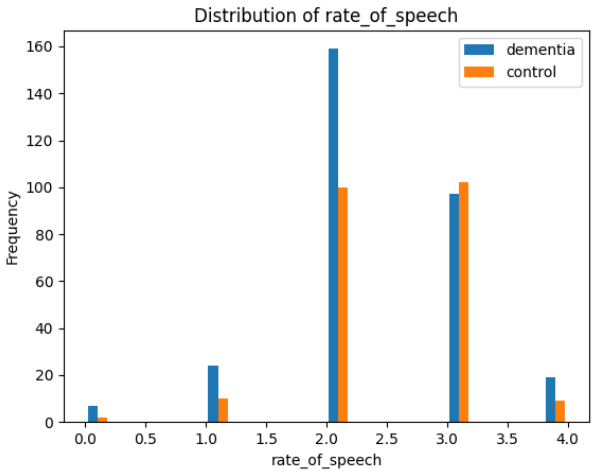


Figure 8.14: Rate of speech distribution.

8.2.4 Signal Features

Kurtosis of pitch

- Mann-Whitney result: [U statistic: 30103.0, p-value: 0.1944347899320 43]

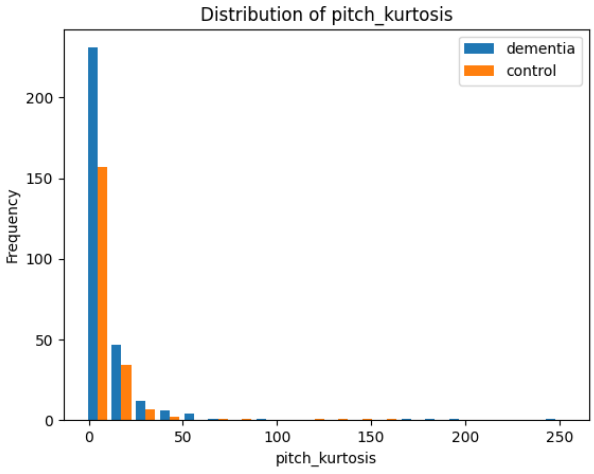


Figure 8.15: Kurtosis of pitch distribution.

Mean of pitch

- Mann-Whitney result: [U statistic: 30143.0, p-value: 0.20121129164820417]

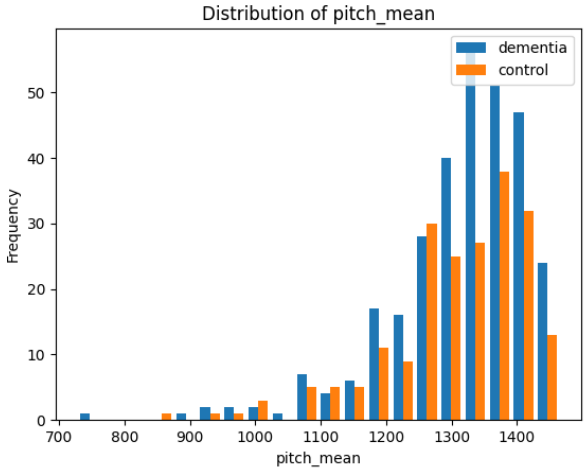


Figure 8.16: Mean of pitch distribution.

Skewness of pitch

- Mann-Whitney result: [U statistic: 29985.0, p-value: 0.17526774692000935]

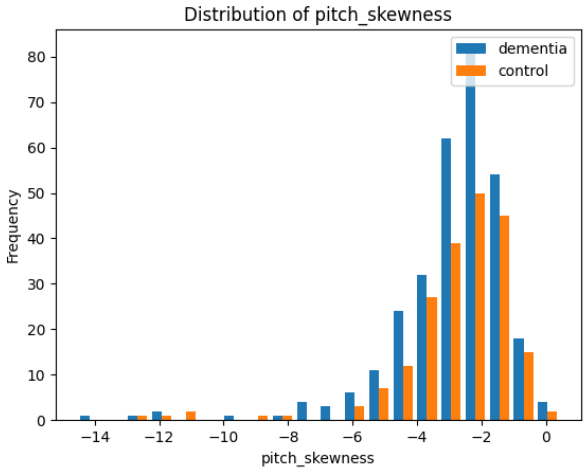


Figure 8.17: Skewness of pitch distribution.

Variance of pitch

- Mann-Whitney result: [U statistic: 30770.0, p-value: 0.3244274204857 69]

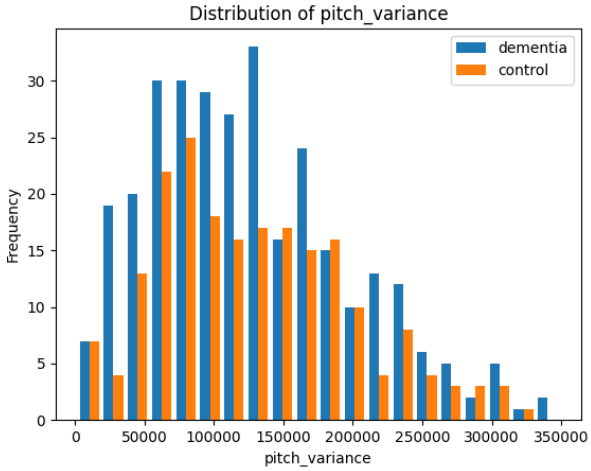


Figure 8.18: Variance of pitch distribution.

Kurtosis of volume

- Mann-Whitney result: [U statistic: 35985.0, p-value: 0.10569096942321299]

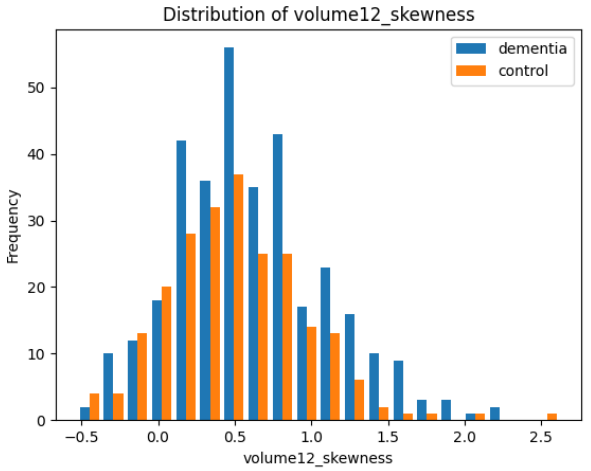


Figure 8.19: Kurtosis of volume distribution.

Mean of volume

- Mann-Whitney result: [U statistic: 34127.0, p-value: 0.012991535307823904]

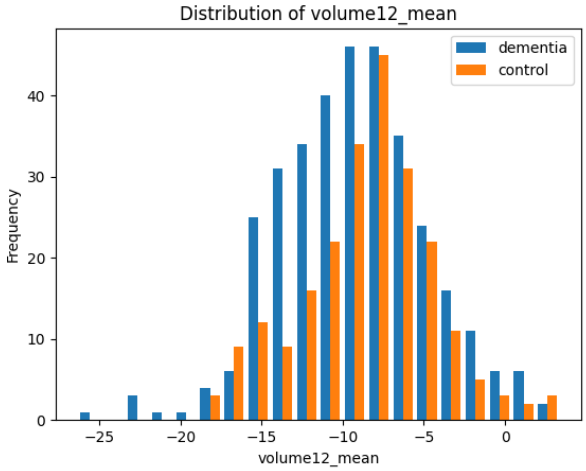


Figure 8.20: Mean of volume distribution.

Skewness of volume

- Mann-Whitney result: [U statistic: 33111.0, p-value: 0.0028852690525494 94]

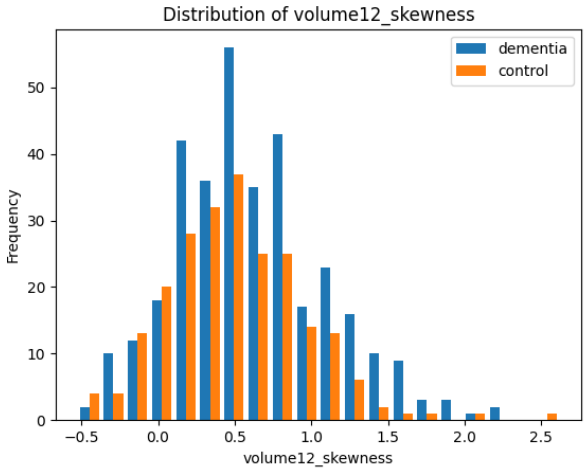


Figure 8.21: Skewness of volume distribution.

Variance of volume

- Mann-Whitney result: [U statistic: 37701.0, p-value: 0.3640218379545 63]

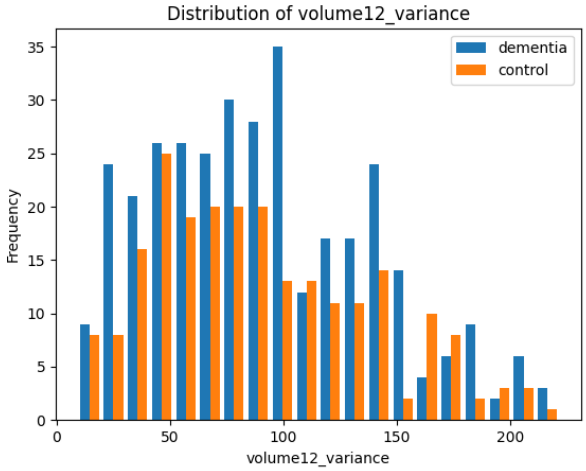


Figure 8.22: Variance of volume distribution.

Kurtosis of ZCR

- Mann-Whitney result: [U statistic: 45288.0, p-value: 0.3890896271640787]

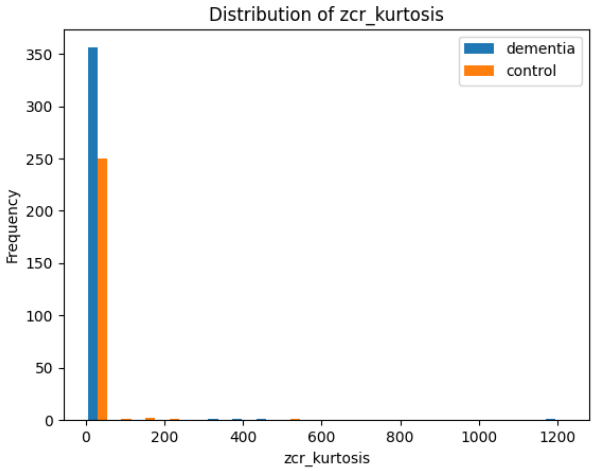


Figure 8.23: Kurtosis of ZCR distribution.

Mean of ZCR

- Mann-Whitney result: [U statistic: 45800.0, p-value: 0.4817206803894213]

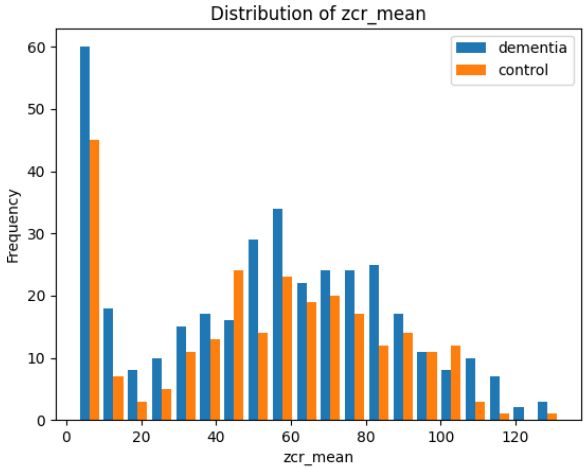


Figure 8.24: Mean of ZCR distribution.

Skewness of ZCR

- Mann-Whitney result: [U statistic: 44686.0, p-value: 0.28807782983485464]

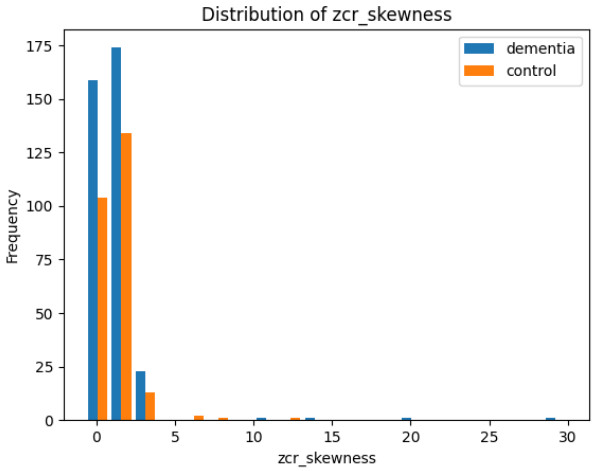


Figure 8.25: Skewness of ZCR distribution.

Variance of ZCR

- Mann-Whitney result: [U statistic: 45756.0, p-value: 0.47364732605623 31]

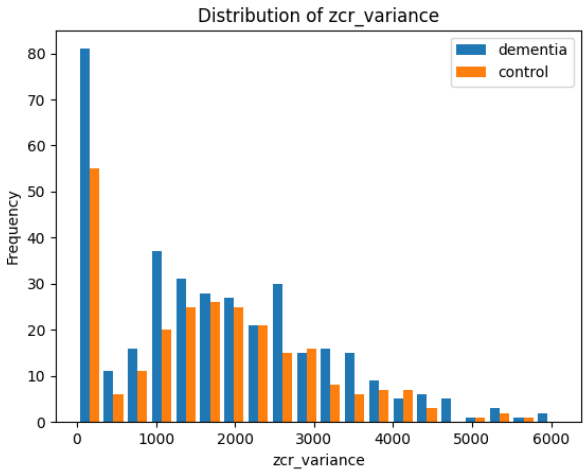


Figure 8.26: Variance of ZCR distribution.

8.2.5 Voice Quality

Hardness

- Mann-Whitney result: [U statistic: 45099.0, p-value: 0.3561544324819 47]

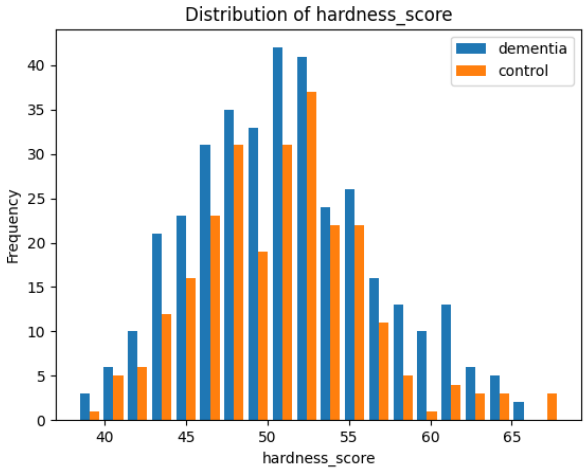


Figure 8.27: Hardness score distribution.

Sharpness

- Mann-Whitney result: [U statistic: 43742.0, p-value: 0.1601429976173 78]

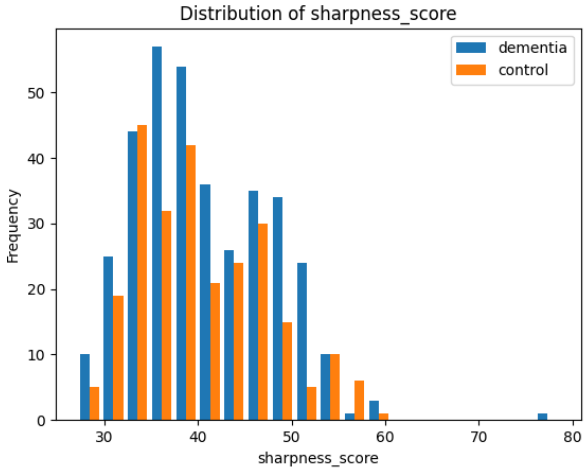


Figure 8.28: Sharpness score distribution.

warmth score

- Mann-Whitney result: [U statistic: 43810.0, p-value:0.16788781068470804]

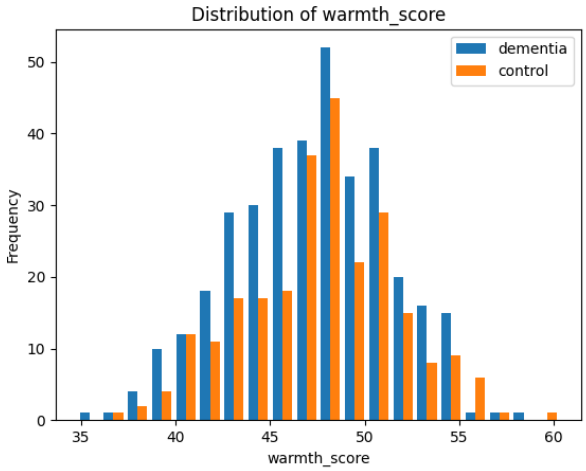


Figure 8.29: Warmth score distribution.

Booming score

- Mann-Whitney result: [U statistic: 450311.0, p-value: 0.34454815010766 34]

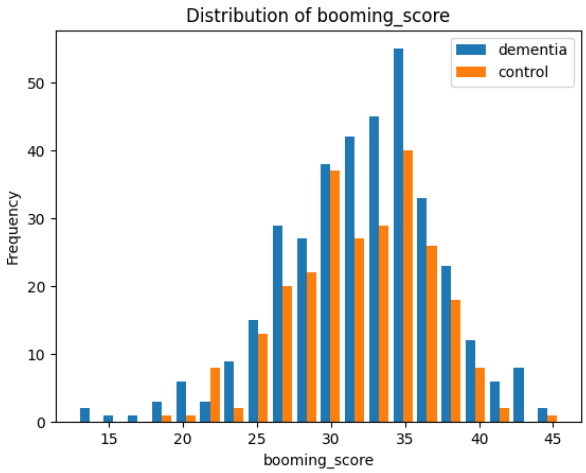


Figure 8.30: Booming score distribution.

8.2.6 Complexity

Automated readability index

- Mann-Whitney result: [U statistic: 40454.5, p-value: 0.0060578898344879 85]

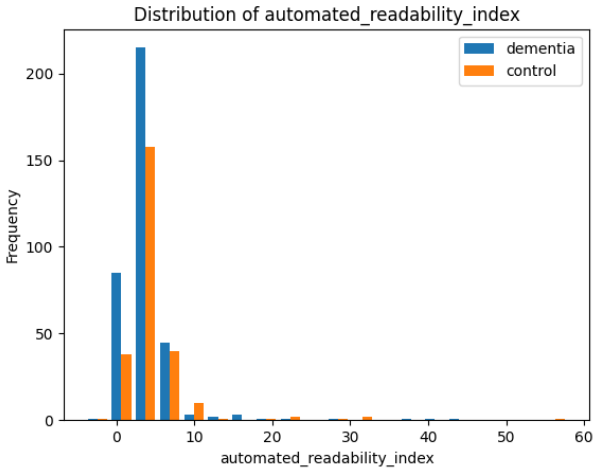


Figure 8.31: Automated readability index distribution.

Coleman liau index

- Mann-Whitney result: [U statistic: 37079.0, p-value: 2.4194560430450334e-05]

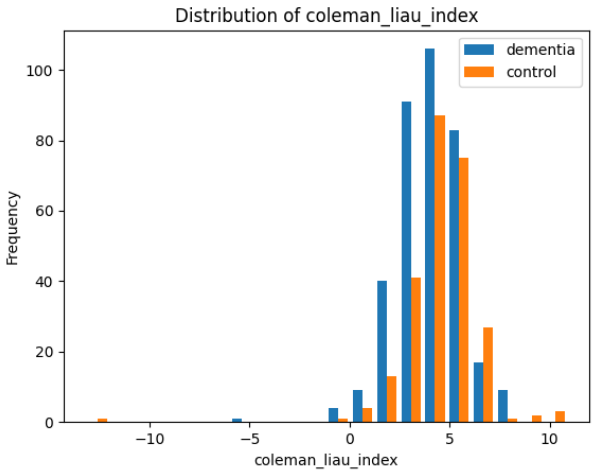


Figure 8.32: Coleman liau index distribution.

Dale chall readability score

- Mann-Whitney result: [U statistic: 32206.5, p-value: 1.4143331911863448e-10]

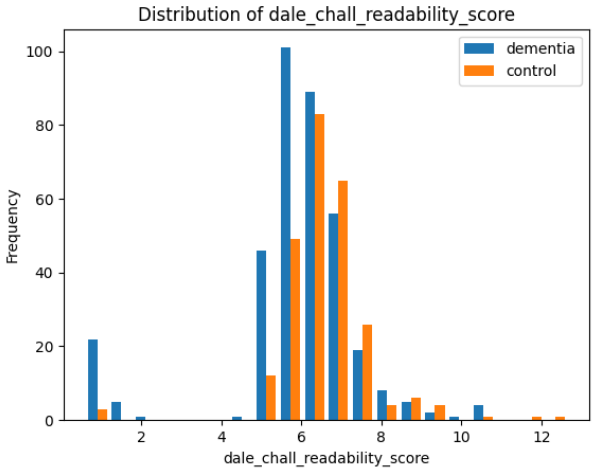


Figure 8.33: Dale chall readability distribution.

Flesch kincaid grade

- Mann-Whitney result: [U statistic: 35923.5, p-value: 2.1471217096784 02e-06]

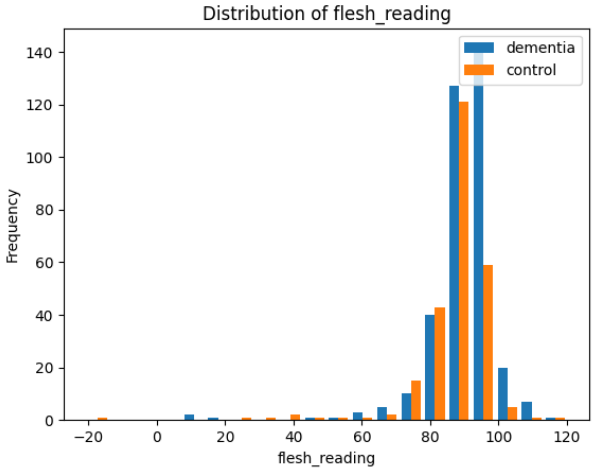


Figure 8.34: Flesch kincaid grade distribution.

Flesh reading

- Mann-Whitney result: [U statistic: 33141.0, p-value: 2.0834943820986 65e-09]

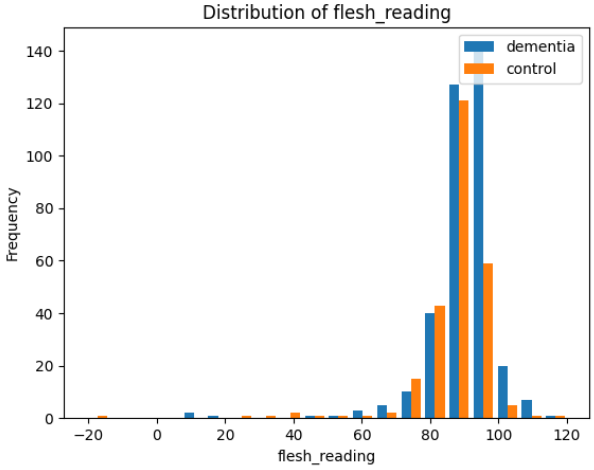


Figure 8.35: Flesh reading distribution.

Gunning fog

- Mann-Whitney result: [U statistic: 40362.5, p-value: 0.005375341087445494]

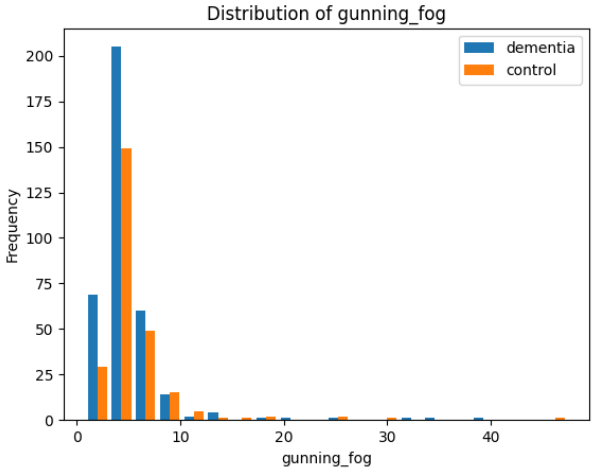


Figure 8.36: Gunning fog distribution.

Lexicon count

- Mann-Whitney result: [U statistic: 37570.0, p-value: 6.2215131123806 75e-05]

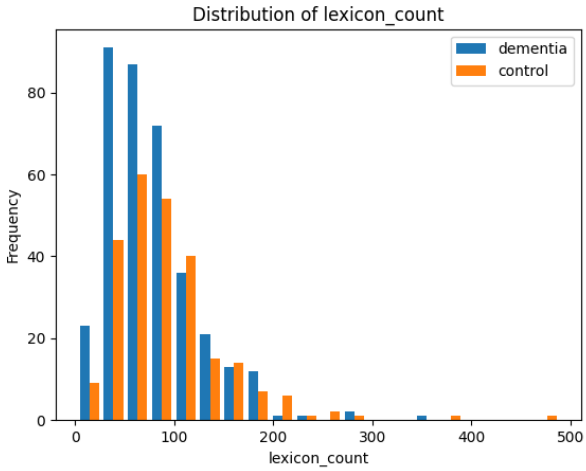


Figure 8.37: Lexicon count distribution.

Linsear write formula

- Mann-Whitney result: [U statistic: 42238.0, p-value: 0.045825203908256694]

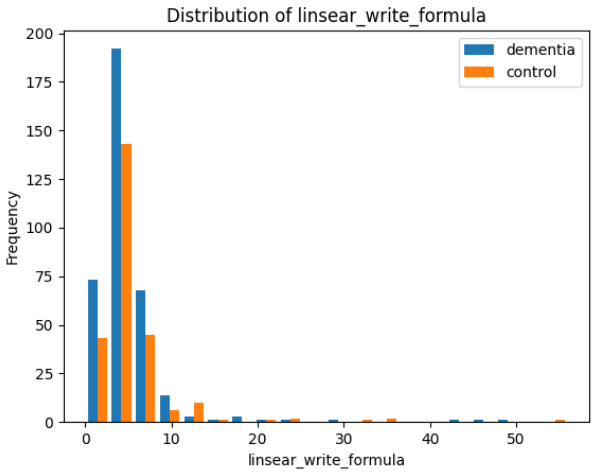


Figure 8.38: Linsear write formula distribution.

Mcalpine eflaw

- Mann-Whitney result: [U statistic: 41901.5, p-value: 0.03275364912141393]

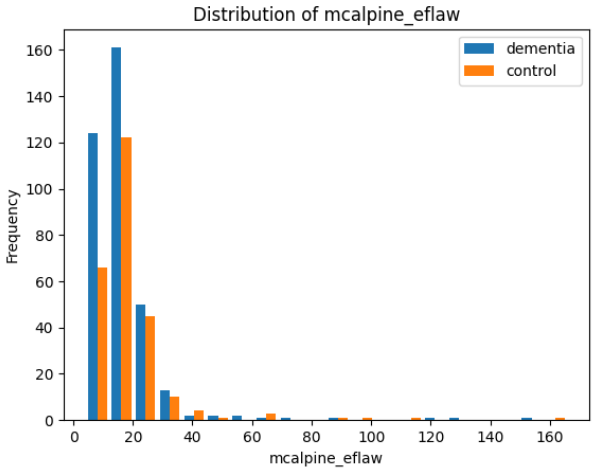


Figure 8.39: Mcalpine eflaw distribution.

Overall readability

- Mann-Whitney result: [U statistic: 42239.5, p-value: 0.04322547906057454]

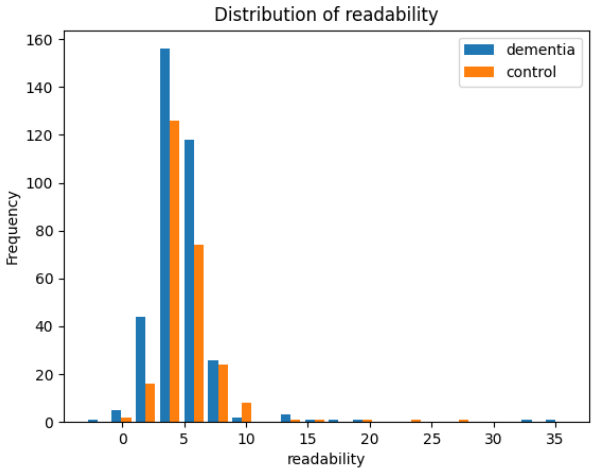


Figure 8.40: Overall readability distribution.

Smog index

- Mann-Whitney result: [U statistic: 40049.5, p-value: 0.003148447767625805 8]

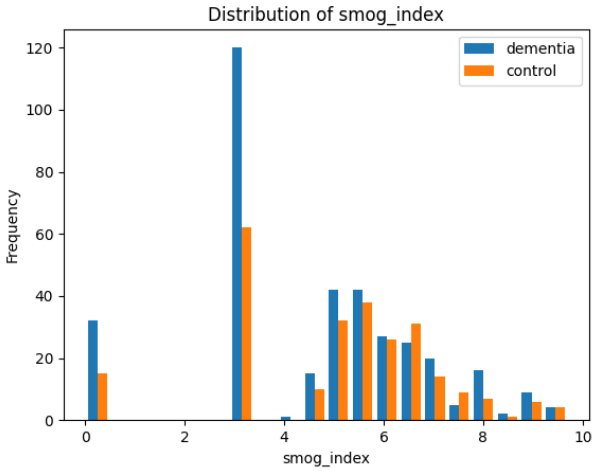


Figure 8.41: Smog index distribution.

Spache readability

- Mann-Whitney result: [U statistic: 32523.5, p-value: 3.593143626445065e-10]

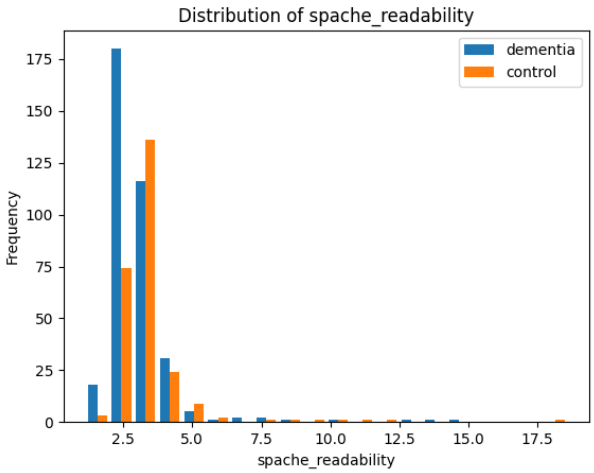


Figure 8.42: Spache readability distribution.

Text standard

- Mann-Whitney result: [U statistic: 42239.5, p-value: 0.043225479060574 54]

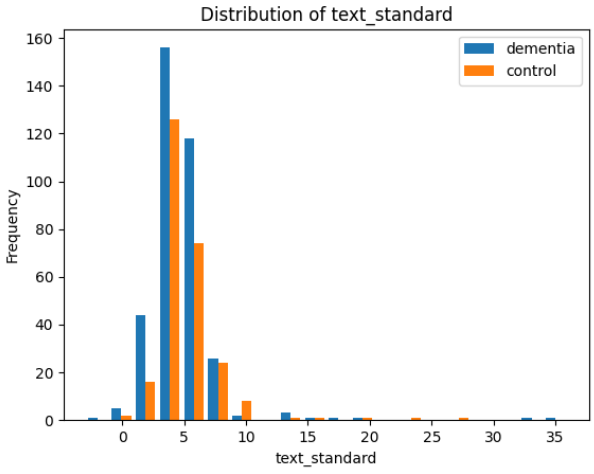


Figure 8.43: Text standard distribution.

Mean length of sentence

- Mann-Whitney result: [U statistic: 38912.0, p-value: 0.0006434329643023259]

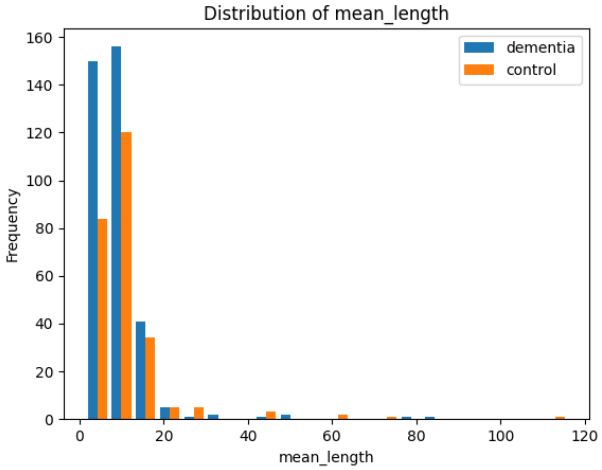


Figure 8.44: Mean length of sentence distribution.

8.3 Final Equation

A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the scoring factor. The following Table 8.1 shows the score weight of the p-value.

p-value range	weighting
0-0.01	5
0.01-0.02	4
0.02-0.03	3
0.03-0.04	2
0.04-0.05	1

Table 8.1: Score Weighting

The scoring factors with a p-value less than 0.05 will be included in the equation to distinguish people from dementia and forecast the level of dementia. The range of scores of each

scoring factor is 0-1. When the scoring factor is directly proportional to dementia risk, the final result of the scoring factor is equal to one minus the result before any further calculation. The details of scoring factors, their weighting, and directional proportion will be included in Table 8.2 below.

Scoring factor	p-value	weighting	Directional proportional to dementia
Polysyllabic, monosyllabic ratio	0.014	4	Inversely
Question ratio	1.44e-09	5	Inversely
Ratio of conjunction	0.00032	5	Inversely
Ratio of noun	2.46e-10	5	Inversely
Ratio of pronoun	8.04e-09	5	Directly
Ratio of verb	4.77e-07	5	Inversely
MLTD	0.049	1	Inversely
Rate of speech	0.0045	5	Directly
Mean of volume	0.012	4	Inversely
Skewness of volume	0.0029	5	Inversely
Automated readability index	0.0061	5	Inversely
Coleman Liau index	2.14e-05	5	Inversely
Dale chall readability	1.41e-10	5	Inversely
Flesch kincaid grade	2.14e-06	5	Inversely
Flesh Reading	2.08e-09	5	Inversely
Gunning Fog	0.0054	5	Inversely
Lexicon count	6.22e-05	5	Inversely
Linsear write formula	0.045	1	Inversely
Mcalpine eflaw	0.033	2	Inversely
Overall readability	0.043	1	Inversely
Smog index	0.0031	5	Inversely
Spache Readability	3.59e-10	5	Inversely
Text standard	0.043	1	Inversely
Mean length of sentence	0.00064	5	Inversely

Table 8.2: Summary of scoring factors

To prevent the double counting effect, the scoring factors measuring similar items will be taking an average of their weighted value. For example, ”Mean of volume” and ”Skewness of

volume" are also calculated as the volume of the audio. Thus, the volume of the audio contributed to the final equation equal to

$$\frac{4(\text{Mean of volume}) + 5(\text{Skewness of volume})}{2}$$

The final scoring equation equal to

$$\frac{(4PM + 5Q + 5C + 5N + 5(1-P) + 5V + MLTD + 5(1-S) + volume + readability + 5ML)}{11}$$

where PM is Polysyllabic, monosyllabic ratio, Q is Question ratio, C is Ratio of conjunction, N is Ratio of noun, P is Ratio of pronoun, V is Ratio of verb, S is Rate of speech, ML is Mean length of sentence, $volume$ is

$$\frac{4(\text{Mean of volume}) + 5(\text{Skewness of volume})}{2}$$

, $readability$ is

$$\frac{5AR + 5CL + 5DC + 5FK + 5FR + 5GF + 5LC + LW + 2ME + OR + 5SI + 5SR + TS}{13}$$

where AR is Automated readability index, CL is Coleman Liau index, DC is Dale chall readability, FK is Flesch kincaid grade, FR is Flesh Reading, GF is Gunning Fog, LC is Lexicon count, LW is Linsear write formula, ME is Mcalpine eflaw, OR is Overall readability, SI is Smog index, SR is Spache Readability and TS is Text standard

The speaker with the lowest score in the conversation will be assumed a speaker affected by dementia. The possible range of the final score is 0 to 4.40 and the average final score of the dementia group is 1.73 and the average final score of the control score is 2.86. For the testing sample lower than the average final score of the dementia group, the speaker will also be assumed as dementia. The lower the score of the sample, the higher level of the dementia.

9 Conclusion and Future work

9.0.1 Conclusion

Data preprocessing was a big task for us in this project, including resampling the audio frequency, extracting MFCC information from the audio, audio segmentation into parts, STT, and also removing samples without sufficient information. Nevertheless, the training process still took a long time, especially when doing the hyperparameter research for emotion detection in both SVM and CNN models.

In this project, we analyzed the spoken narratives of people with and without dementia using NLP-based techniques, lexical diversity, speech fluency, signal features, voice quality, and emotion to finalize an equation to distinguish people with dementia through their conversation. The results indicate that the POS features, readability of the text, and volume of the audio are more important factors to distinguish speakers who are affected by dementia. We were successful in building an equation for classification people from dementia and their level of dementia, and also implementing fast and accurate models for emotion classification.

9.0.2 Future Work

In the future, we should collect dementia audio with emotion labeling. Since the dementia bank dataset does not label the emotion of the speaker in the audio, we used the RAVDESS dataset instead of the dementia bank dataset to train the emotion detection model. We do not

have a suitable dataset to conclude whether the emotion is a factor to distinguish people from dementia. We can add emotion to the final scoring equation after we tested this factor with the suitable dataset.

We also plan to experiment on how the knowledge learned by our model could be transferred to detect people with dementia that speak languages different than English.

Moreover, we can increase the accuracy of our STT and speaker distinguish model to replace assembly AI API service.

Last but not least, after the company collects its users' data and finishes the labeling work, we can use that dataset as the training dataset instead of the data we collected from dementiaBank, etc data sources. Hence, the project would be more tailor-made for company usage. After having more samples of data, we can update the threshold of the final scoring for distinguishing people with dementia and their level of dementia.

References

[1] Deepvibes company website. [Online]. Available: https://deepvibes.ai/

[2] R'mani Haulcy and James Glass. Classifying alzheimer's disease using audio and text-based representations of speech. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.624137/full

[3] What is dementia? [Online]. Available: https://www.cdc.gov/aging/dementia/index.html

[4] assemblyai website. [Online]. Available: https://www.assemblyai.com/

[5] When people with dementia become withdrawn. [Online]. Available: https://www.scie.org.uk/dementia/living-with-dementia/difficult-situations/being-withdrawn.asp

[6] M. G. Daniel Kempler. Language and dementia: Neuropsychological aspects. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2976058/

[7] P. G. W. H. Kayla Chapina, Natasha Clarkeb. A finer-grained linguistic profile of alzheimer's disease and mild cognitive impairment. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0911604422000136

[8] R. J. Xuan Le, Ian Lancashire. Natural language processing methods for the detection of symptoms of alzheimer's disease in writing. [Online]. Available: http://www.cs.toronto.edu/pub/gh/Google-talk.pdf

[9] C. D. Ali Khodabakhsh Inmeta, Serhan Kusxuoglu.

[10] A. P. C. B. C. C. G. Z. Marco Zuin, Loris Roncon. Risk of dementia in patients with atrial fibrillation: Short versus long follow-up. a systematic review and meta-analysis. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8518611/

[11] J. H. M. P. Shamila Nasreen, Rohanian. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.640669/full

[12] P. A. S. L. Fasih Haider, Sofia de la Fuente Garcia. Affective speech for alzheimer's dementia recognition. [Online]. Available: https://www.researchgate.net/publication/341234029_Affective_Speech_for_Alzheimer's_Dementia_Recognition

[13] O. S. Silva Banovic, Lejla Junuzovic Zunic. Communication difficulties as a result of dementia. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6195406/#:~:text=The%20main%20characteristics%20of%20speech,a%20louder%20voice%20when%20speaking.

[14] M. Ratini. Voice problems and alzheimer's disease. [Online]. Available: https://www.webmd.com/alzheimers/voice-speaking-problems-alzheimers

[15] A. Robinson. Dementiabank website. [Online]. Available: https://dementia.talkbank.org/

[16] R. F. A. Livingstone, Steven R. The ryerson audio-visual database of emotional speech and song (ravdess). [Online]. Available: https://zenodo.org/record/1188976#.Yx9nsXZBy5c

[17] What is a spectrogram? [Online]. Available: https://pnsn.org/spectrograms/what-is-a-spectrogram

[18] Understanding the mel spectrogram. [Online]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[19] J. M. N. A. K. A. A. C. I Kamarulafizam, Sh Hussain Salleh. Heart sound analysis using mfcc and time frequency distribution. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-68017-8_102

[20] M. Courthoud. How to compare two or more distributions. [Online]. Available: https://towardsdatascience.com/how-to-compare-two-or-more-distributions-9b06ee4d30bf

Acknowledgments

I would like to acknowledge and give my warmest thanks to both my industrial supervisor Federico Ghiradelli and academic supervisor Prof. Lamberto Ballan who made this work possible. Their advice carried me through all the stages of writing my project. I would also like to thank Prof. Alberto Testolin helped me to request access to the DementiaBank dataset.