Chapter 3

Depression Dataset

In this work, the DAIC-WOZ depression database [9] compiled by the USC's Institute of Creative Technologies was utilized as the raw data source for our generated depression dataset. The DAIC-WOZ depression database consists of clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD) [18]. The length of interviews ranges from 7 to 33 minutes with an average of 16 minutes. DAIC stands for Distress Analysis Interview Corpus, a large corpus in which this database originates, and WOZ indicates the interview approach, namely the Wizard-of-Oz method, where participants interact with a virtual interviewer called "Ellie" that in their believe is autonomous, but in reality, is controlled by an unseen human interviewer in the next room. An illustration of the interview is shown in Figure 3.1 according to [18].

Figure 3.1: An illustration of the Wizard-of-Oz interview. The participant interacts with virtual interviewer, Ellie, during the interview.

The DAIC-WOZ depression database [9] contains visual as well as acoustic recordings and transcriptions of 189 participants (ID ranging from 300 to 492), split into a train set (107 participants), a development set (35 participants), and a test set (47 participants). **Important here to remember is that in all experiments in this thesis, the train set and the development set are merged to construct the training dataset (142 participants) and the original test set is kept as test dataset (47 participants) to test the performance of the developed model in this thesis.** For each session, an eight-item Patient Health Questionnaire depression scale (PHQ-8) is provided as Ground Truth (GT), which indicates the severity of depression, and a PHQ-8 Score ≥ 10 implies that the participant is undergoing a major depression (MD) [24].

3.1 Definition of PHQ-8 System

One of the standardized and validated methods for assessing and diagnosing the severity measure for depressive disorders in large clinical studies is the so-called eight-item Patient Health Questionnaire depression scale developed by Kroenke and Spitzer et al. [23]. The PHQ-8 Score consists of 8 of the 9 criteria (also known as PHQ-8 Subscores), on which the DSM-IV diagnosis of depressive disorders is based [13]. These 8 different aspects of depressive criteria are shown in the Table 3.1 in the section of PHQ-8 Subscores according to [24].

Table 3.1: A Patient Health Questionnaire eight-item depression measure (PHQ-8)

Over the last 2 weeks, how often have you been bothered by any of the following problems?	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
PHQ-8 Subscores				
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself - or that you are a failure	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
PHQ-8 Score				
Total score ___ = ___ + ... + ___ (sum of all PHQ-8 Subscores, 0 - 24)				
PHQ-8 Binary				
Final result ___ = 1 if PHQ-8 score $\geq$ 10 else 0				

Table 3.1 also demonstrates the definition and relationship between each score in the PHQ-8 system. To obtain the PHQ-8 Score, one would be inquired about the number of days in the past 2 weeks one had experienced a particular depressive symptom. Based on the response and the following conversion: 0 to 1 day means "not at all," 2 to 6 days means "several days," 7 to 11 days means "more than half the days," and 12 to 14 days means "nearly every day," the PHQ-8

Subscore for each criterion is acquired by assigning points (0 to 3) to each category, respectively. The results of PHQ-8 Subscores are then summed up to produce a total PHQ-8 Score between 0 to 24 points, from which a binary state of MD is further derived based on the PHQ-8 Score with 10 as threshold. If PHQ-8 Score $\geq$ 10, it results in an outcome of true classification of having MD, otherwise false. The representation of the depression severity at each numerical range in accord with the PHQ-8 Score is shown in the Table 3.2.

Table 3.2: A representation of PHQ-8 Score

PHQ-8 Score	Level of depressive symptoms	State of MD
0 - 4	not significant	No
5 - 9	mild	No
10 - 14	moderate	Yes
15 - 19	moderately severe	Yes
20 - 24	severe	Yes

So far the definition of the PHQ-8 system (GT of DAIC-WOZ database [9]) has been well explained in-depth. The corresponding underlying relationships among there 3 scores are also established, that is PHQ-8 Subscores, PHQ-8 Score, and PHQ-8 Binary, ranging between 0 to 3, 0 to 24, and 0 / 1, respectively. Hence, it is conspicuous that 3 different prediction scores can be chosen as the output format of the developed depression estimation architecture and either be considered as a classification predictive modeling problem or a regression predictive modeling problem. A classification head provides an advantage of exact prediction by predicting a discrete class label, which resembles the way PHQ-8 structures, whereas a regression head provides an advantage of minimizing the error in decimal places by predicting a continuous quantity. Therefore, several different variations of prediction for such supervised learning task based on DAIC-WOZ database [9] can be found in the previous automatic depression estimation works. Williamson et al. [38] and Gong et al. [16] train their model with regression head by minimizing the RMSE to successfully predict the PHQ-8 Score and further derive the final binary state of MD through cut-off point. Ma et al. [30] and Bailey et al.[4] regard depression detection as a classification problem and solely predict the binary result of MDD of a participant, which is also investigated in other studies [25, 31, 40]. Alhanai et al. [3] and Valstar et al. [37] design 2 models with 2 different output head, one with classification head to model PHQ-8 Binary outcomes and the other with regression head for multi-class outcomes of PHQ-8 Score. Similar to that, Dham et al. [12] also develop 2 models for classification and regression approach. However, instead of predicting PHQ-8 Score, PHQ-8 Subscores were predicted, and the results of final PHQ-8 Score as well as PHQ-8 Binary are calculated according to the definition. More recently, Haque et al. [19], Song et al. [33], and Lin et al. [26] deploy a specific criterion function during the training process to fuse the cross entropy loss and the loss of depression severity assessment since their designed model output with 2 branches, namely a depression classifier for PHQ-8 Binary and a PHQ regression model for PHQ-8 Score.

In this study, in accord with the way PHQ-8 system structures and the consideration of depres-

sion estimation as a classification task, **a classification head, predicting PHQ-8 Subscores, is predominantly exploited in all of the experiments.** PHQ-8 Score as well as PHQ-8 Binary are then derived through the definition, resembling the method in [12]. However, to prove that PHQ-8 Subscores are most informative and do provide the best performance, comparative research is also conducted in chapter 5, which indeed justify this thought. For more details and results about the comparative research, please refer to the Chapter 5.

3.2 Data Understanding

As mentioned above, the DAIC-WOZ [9] depression database contains 189 sessions of interview recordings in total, meaning that 189 interviewees have participated. During the interview, the virtual interviewer, Ellie, will inquire each participant with a subset of possible queries, included direct questions, e.g. "How have you been feeling lately", "Have you ever been diagnosed with PTSD", and response with dialogic feedback, e.g. "I see", "Cool", "That's great." For each session, the recorded data has been transcribed and annotated for a variety of features, which can be categorized into 3 main types, that is visual features, audio features, and text features. In the following parts these features will be discussed in detail and for ethical reasons, no raw video is made available according to [37].

Visual Features: Both low-level and high-level visual features extracted based on *OpenFace* framework [5] and *FACET* software [27] are provided in DAIC-WOZ depression database [9] and listed below. The sampling rate of all visual data is $30\,\mathrm{Hz}$:

- 2D facial landmarks: 68 facial key points in 2D pixel coordinates estimated from video.
- 3D facial landmarks: 3D coordinates of 68 facial key points estimated from video. The points are in millimeters (mm) in world coordinate space, with camera begin at (0, 0, 0).
- Gaze direction: Gaze direction of both eyes in both world coordinate space and head coordinate space
- Head pose: 3D head position and rotation
- HOG: Histogram of oriented gradients on the aligned 112×112 area of the face on the image [10, 28]
- AUs: 17 action units based on facial action coding system (FACS)

In this thesis, the 3D facial landmarks and the gaze direction of 189 individuals are combined and utilized as our visual features to provide micro-facial expression changes for depression estimation. Since the sampling rate of visual data is $30\,\mathrm{Hz}$, the model could be trained to have the capacity of observing microexpression at the millisecond (ms) level. An illustration of micro-facial expression changes based on 3D facial landmarks is shown in the Figure 3.2. It is worth noting that the serial facial landmarks shown here have been speeded up 5 times faster than

normal, meaning that in the visual dataset, the microexpression is much more subtle. Another interesting phenomenon, which could be noticed here, is that most of the gaze directions are pointing down, implying a certain degree of nervousness, sadness, and gloom of the participant. Indeed, this is the clip where the participant answers yes to the question whether he has ever been diagnosed with PTSD or not.

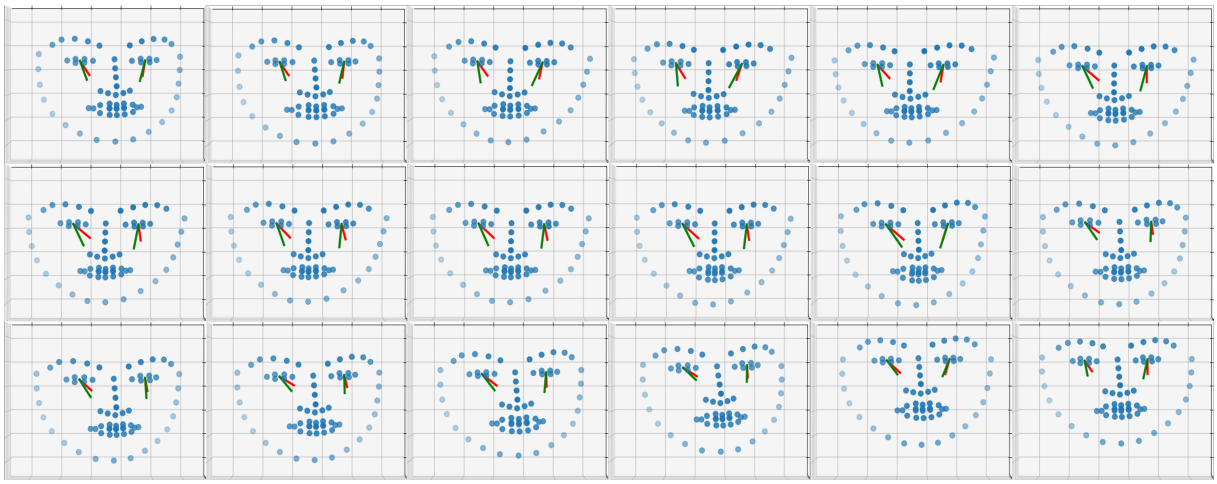


Figure 3.2: An example of Micro-facial expression. The changes of microexpression within 3 s is shown at here(top-left to bottom-right). The red and green marks extending from the eyes visualize the gaze direction in world coordinate and head coordinate, respectively.

Audio Features: The audio data provided in the DAIC-WOZ database [9] consists of a raw audio file, a COVAREP file with pre-extracted audio features, and a formant file, which are recorded over the entire interview. Details of each file are explained below:

- Raw audio file: Recording of the raw audio signal of the whole interview with a head-mounted microphone at a sampling rate of 16 kHz. It was processed so as to possibly only record the voice of the participant. Hence, the voice of the virtual interviewer was silence-suppressed. However, it still contains small amounts of bleed-over of the virtual interviewer.
- COVAREP file: Pre-extracted audio features utilizing the COVAREP(v1.3.2), a freely available open source Matlab and Octave toolbox for speech analyses [11], over the entire recording at every 10 ms, i.e. sampling rate 100 Hz. These involve prosodic features such as fundamental frequency (F0) and voicing (VUV), spectral features including mel cepstral coefficients (MCEP0-24) and harmonic model and phase distortion mean (HMPDM0-24), and energy- or voice quality-related features, for instance, normalized amplitude quotient (NAQ), quasi open quotient (QOQ), maxima dispersion quotient (MDQ), etc.
- Formant file: Containing the first 5 formants, which are the vocal tract resonance frequencies tracked throughout the interview.

In this work, instead of using pre-extracted features in the COVAREP file or formant file, the raw audio file of each individual was preferred and utilized to provide insights into acoustic features

as the model is expected to have the capability of estimating the severity of depression based on audio signal without too many feature engineering techniques, which makes the automatic multi-modal depression estimation feasible for the purpose of serving the public. Imagine how easy it would be for the user to just record their voice and the outcome of MDD will be analyzed directly without any prerequisite of grasping different concepts of higher-level audio features, i.e. MCEP0-24, HMPDM0-24, NAQ, MDQ, etc. A clip of raw audio signal waveform corresponding to the transcript sample shown in the Table 3.3 is illustrated in the Figure 3.3. High-value amplitude areas indicate that the participant is responding to the queries from Ellie, whereas low-value amplitude areas demonstrate the above-mentioned voice suppression while Ellie is speaking.

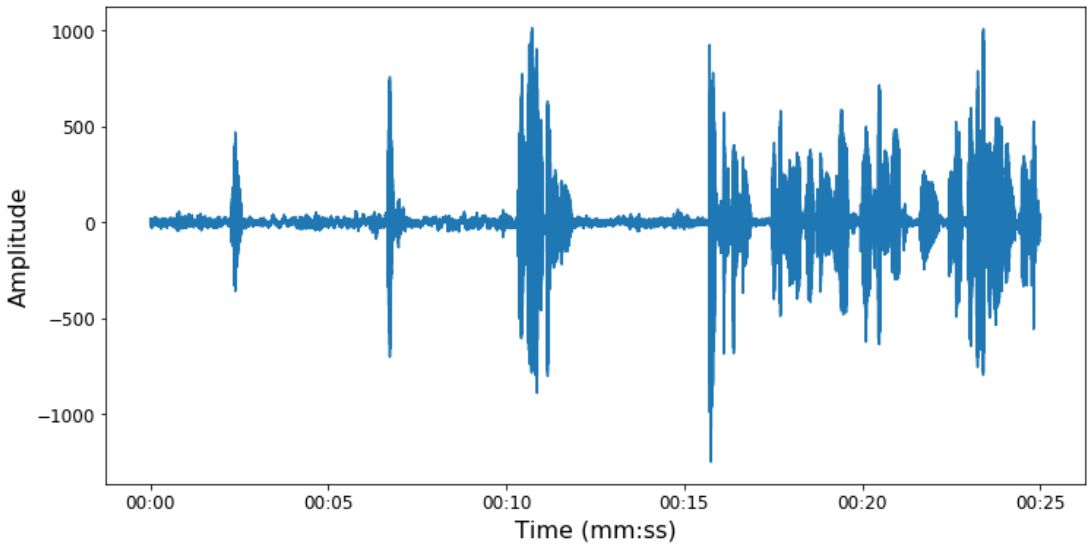


Figure 3.3: An example of Raw audio waveform. Audio signal of Participant 321 beginning from $6'06''$ to $6'31''$.

Text Features: Extensive questionnaire queries of the virtual interviewer as well as verbal responses of the participant are transcribed into text recorded in the transcript file and annotated with the name of the speaker. A sample transcript file is shown in the Table 3.3. In addition, the timestamps of each response are also recorded in a unit of second (s). The transcription conventions according to DAIC-WOZ official documentation are shown below:

- Every word is converted into lower case.
- Incomplete word is annotated with the intended word, followed by a comment with the part that was actually pronounced in angle brackets like: people <peop>.
- Unrecognizable words are indicated as 'xxx'.

In this thesis, the transcript file is not only exploited to contribute textual features to the model, but also the timestamps, which play a critical role in data preprocessing. However, further steps still need to be implemented in order to turn text into an applicable data source since computers can't process words but solely numbers. More details will be illustrated in the Section 3.4.

Table 3.3: An overview of selected samples from the transcript file which are leveraged as text data

Start time in s	Stop time in s	Speaker	Value
...	...	...	...
366.615	368.535	Ellie	have you ever been diagnosed with ptsd
369.375	369.905	Participant	yes
370.960	372.740	Ellie	how long ago were you diagnosed
373.160	374.750	Participant	um it was over five years ago
376.270	377.530	Ellie	what got you to seek help
378.750	380.010	Participant	i couldn't function
380.580	388.990	Participant	i couldn't sleep i couldn't ...
390.350	391.120	Participant	shut down
...	...	...	...

3.3 Potential Problems and Solutions for Dataset

Although the DAIC-WOZ depression database [9] abounds in various data types and data features, which, to a large extent, benefits numerous research for automatic depression estimation with different data-driven approaches, it has been well reported that the DAIC-WOZ depression database [9] contains assorted errors and noises, which will not only cause potential difficulties during the model training process but also sabotage the model performance, which, in the worst case, will mislead the model's attention, leading to completely irrelevant and inapplicable results. Furthermore, it can also be seen from the Section 3.2 that one of the potential problems of insufficient data exists as the DAIC-WOZ depression database [9] consists of only 189 participants. Another potential problem is the imbalanced dataset, involving the imbalance of MD as well as gender imbalance, particularly shown in acoustic features. These potential problems will cause not only failed to fully train a generalized model but also biases. Therefore, in the following subsections, these phenomena as well as solutions for all the above-mentioned problems will be discussed.

3.3.1 Error correction

A list of known errors existing in the DAIC-WOZ depression database [9] is shown below after carefully examining the database:

- GT labeling error: The PHQ-8 Score of participant 409 is 10 but the given PHQ-8 Binary was 0 instead of 1, which contradicts the definition of the PHQ-8 system.
- Missing value error: A value in PHQ-8 Subscores of participant 319 isn't present, which causes *'NaN'* (not a number) problem during the training.
- Data format error: In the development set, the data format of some features of particular participants is different than others. For example, instead of float type, the coordinate of

the facial key points of participant 367 was given in string type. Moreover, since some samplings failed during the interview, those values are assigned to *'-1.#IND'* in string format, meaning not a number. These assigned values have to be replaced in order to process.

- Missing transcript: The transcriptions of the virtual interviewer are missing in some interviews.

To solve the labeling error, missing value error, a pre-check function is created to automatically double-check all the PHQ-8 ground truth of each individual by inspecting the availability of all PHQ-8 Subscores and summarizing them together to form the expected PHQ-8 Score, which will be converted to the expected PHQ-8 Binary based on the definition. These two expected values will then be compared with the provided GT of PHQ-8 Score and PHQ-8 Binary. If only one of the eight values in the PHQ-8 Subscores of a participant is missing, this lost value will be automatically calculated according to the given PHQ-8 Score. However, if there are multiple missing values, then the process will be terminated and the ID of this problem participant will be printed out as there isn't any solution for this. Fortunately, this multiple missing values problem doesn't exist. Another task of this pre-check function is to ensure the correct data format for each loaded feature. All of the loaded data will be converted into a float format. For non-numeric values such as *'NaN'* or *'-1.#IND'* will be set to 0. As for the missing transcription of the interviewer, since only the features of the participant were focused on and extracted in this work, meaning that only the sentences and timestamps of the participant are desired, it doesn't lead to any complication and thus is ignored.

3.3.2 Noise reduction

Similar to errors, the acknowledged noises in the DAIC-WOZ depression database [9] are listed below:

- Acoustic noise: There are usually strong noises at the beginning and end of the audio recording because the agent is supporting the interviewee to settle down while the audio recording has already started.
- Bleed-over of the voice of virtual interviewer: Small amount of Ellie's voice is still recorded in spite of silence suppression shown in figure 3.4.
- Interruption during the interview: Some interviews involve long or short interruptions caused by technical issues or ringing of the cellphone, which should be removed.
- Irrelevant interaction: Each recording contains interactions between the participant and the researcher prior to the beginning of the interview which needs to be removed.

As disclosed above, most of the concern about noises relates to acoustic data. Hence, in this subsection, an example of the process of audio data cleaning will be well explained in-depth and an illustration is demonstrated in the Figure 3.4. Three different audio signal waveforms

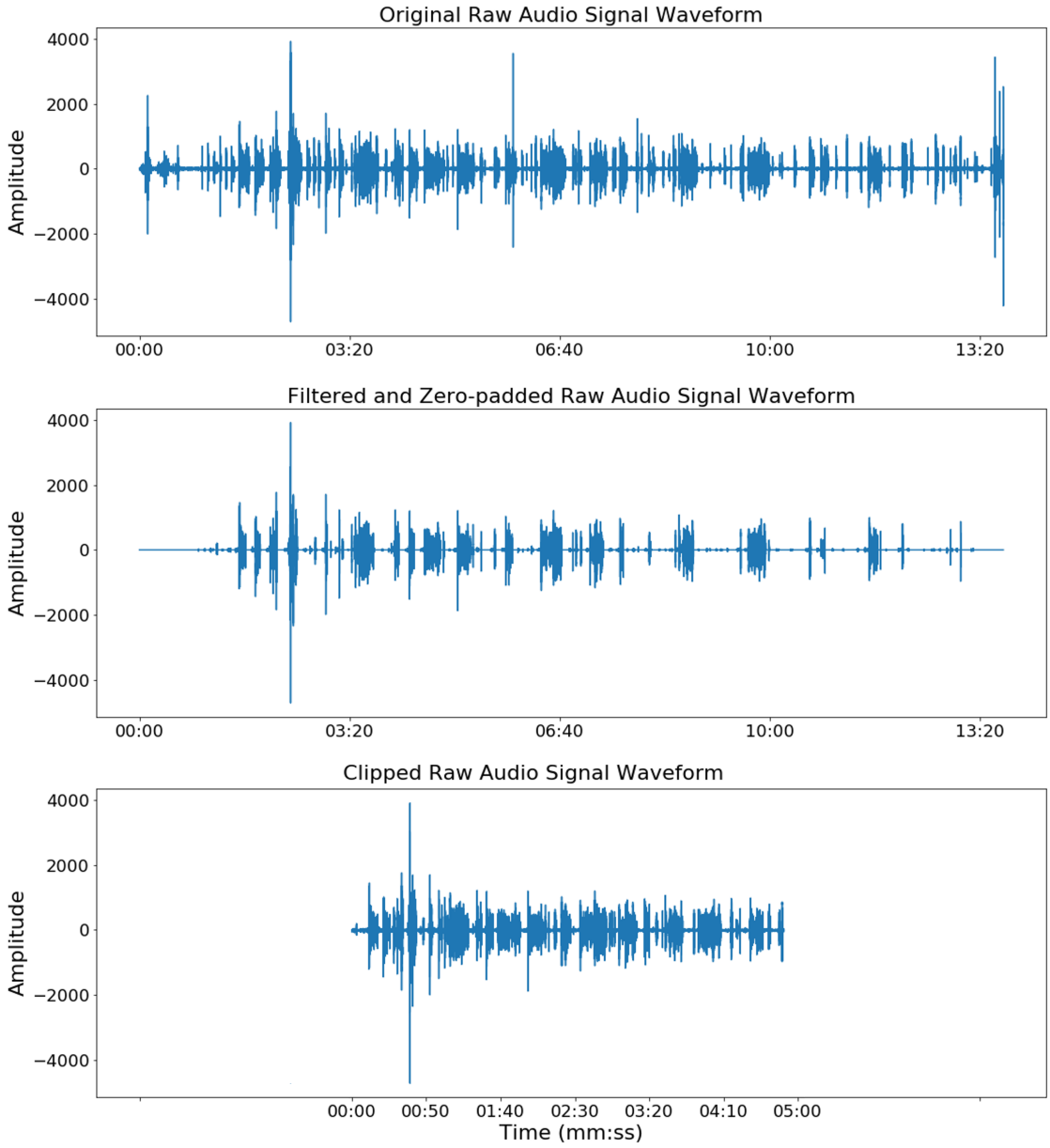


Figure 3.4: An illustration of noise reduction of audio signal, which includes the illustration of the original audio signal, interim result after filtering, and the final clipped audio signal exploited in this work.

are shown here, namely "Original Raw Audio Signal Waveform", "Filtered and Zero-padded Raw Audio Signal Waveform", and "Clipped Raw Audio Signal Waveform", ordering from top to bottom. As their name implies, distinct methods have been applied respectively to clean the data. The raw signal waveform in the first plot is the initial and unprocessed audio data and the filtered signal waveform in the second plot is the interim result after filtering out all of the noises as well as voices besides the participant in alignment with the audio length of the raw signal waveform. By comparing these two, it is conspicuous that the raw signal consists of a huge amount of acoustic noises, the bleed-over voice of the virtual interviewer, etc. Therefore, it

is then cropped and clipped based on the timestamps provided in the transcription and includes only the acoustic features of the participant. This final outcome is illustrated in the third plot of the Figure 3.4, which is also the ultimate audio signal utilized in this thesis later in the Section 3.4, "Preprocessing".

3.3.3 Handling of an imbalanced dataset

A major challenge in training a shallow or deep depression estimation model with the DAIC-WOZ database [9] lies in the unequal distribution of the dataset, including an uneven sample of depressed and non-depressed participants as well as gender imbalance, particularly appeared in acoustic features. It has been widely reported that imbalanced classes in a dataset will greatly affect the performance of the ML model. Moreover, many current benchmarks [4, 16, 30, 3] have shown great adversity of undergoing data imbalance among different levels of depression, which incurs a large bias in the predicted results. Hence, several techniques have been developed to solve this problem. Ma et al. [30] applied the random sampling technique to the non-depressed class to randomly crop for each subject to match the number of clips in the depressed class, on which no particular operation has been conducted. Gong et al. [16] performed random-oversampling by simply duplicating samples to make the number of samples for each PHQ-8 Score roughly equal.

In this subsection, the illustrations of imbalanced phenomena and the approaches to coping with such problems in this thesis will be shown and discussed in detail.

Imbalance between depressed and non-depressed participant: DAIC-WOZ depression database [9] involves the recording of 189 participants, split into train set, development set, and test set. As aforementioned, in this thesis, our generated training dataset (142 participants included) is based on the combination of the train set and development set for the training process and the the test set is only composed of the provided test set in this dataset (47 participants included). Since only the training dataset is related to the model training itself, only the distribution in the training dataset will be discussed here. However, the distribution in the test dataset also resembles this result in the training dataset. An overview of the distribution of depressed (D) and non-depressed (ND) participants, as well as female (F) and male (M) participants, are shown in the Table 3.4.

Table 3.4: An overview of the distributions of PHQ-8 Binary Score regarding gender and major depression

Training	Female (F)	Male (M)	Total F+M
ND	39 (28%)	60 (42%)	99 (70%)
D	24 (17%)	19 (13%)	43 (30%)
Total ND+D	63 (45%)	79 (55%)	142 (100%)

On first look at the last column, it can be noticed that the proportion of depressed to non-depressed participants is about 3 to 7, meaning that only 30% of the participants are being

classified as depressed, whereas non-depressed participants constitute about 70% of the participant, which is about 2.3 times larger than that of depressed ones. Furthermore, between each gender, the distribution of PHQ-8 Binary in male participants shows a greater imbalanced phenomenon than in the female. To further dive deep into PHQ-8 Score as well as PHQ-8 Subscores, bar graphs of both are shown in the Figure 3.5. One can tell that no matter which subclass in PHQ-8 Subscores or PHQ-8 Score, the tendency slants to the right side, meaning a strong bias toward non-depressed class, thus making this database not reliable by directly adopting it for model learning.

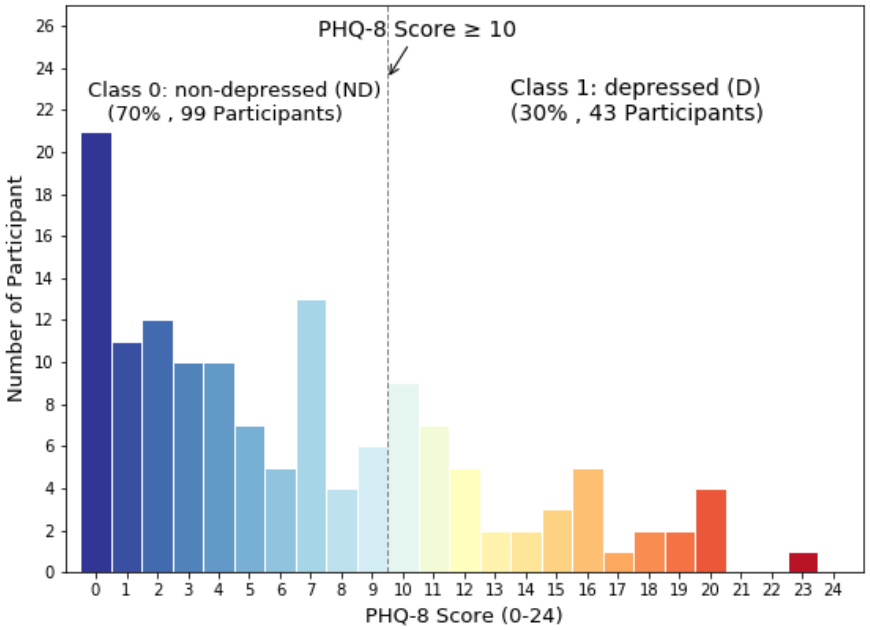
To solve this, a "weighted random sampler" in PyTorch is utilized for the data loader to equally load the data from each class. However, instead of loading each PHQ-8 subclass (0-4) equally, which is the predicted score for our 8 classification heads mentioned previously, the PHQ-8 classes corresponding to the PHQ-8 Score, ranging from 0 to 24, are loaded equally since PHQ-8 Subscores are fixed to each participant and there isn't any other way to equalize the number of subclasses while loading the batches based on either the participants or clips of the participant. Moreover, a dynamic weighted loss function is also applied to minimize this imbalanced phenomenon rather than a loss function without any weight or a static weighted loss function. Due to the fact that the subclasses can't be loaded equally between each batch, only the distribution of each subclass per batch can be derived. Therefore, the weight for the weighted loss function is dynamically calculated according to the number of each subclass of each batch throughout the training.

Gender imbalance, particularly in acoustic feature: Another imbalance one can observe from the table 3.4 is the gender imbalance. In total, female constitutes about 42% of the participants (equal to 68 people), where 39 female participants are non-depressed and 24 female participants are having MD, while male constitutes about 55% of the participants (equal to 79 people), where 60 male participants are non-depressed and only 19 male participants are having MD. There is a 10% difference shown in the number of female and male participants. This imbalance affects acoustic features particularly strong as one could imagine that it is hard to tell just based on the facial key points shown in the Figure 3.2 or text shown in the Table 3.3 whether this person is a man or a woman, whereas it is conspicuous just by listening to the voice of the recording or observing the fundamental frequency (F0) in the spectrogram.

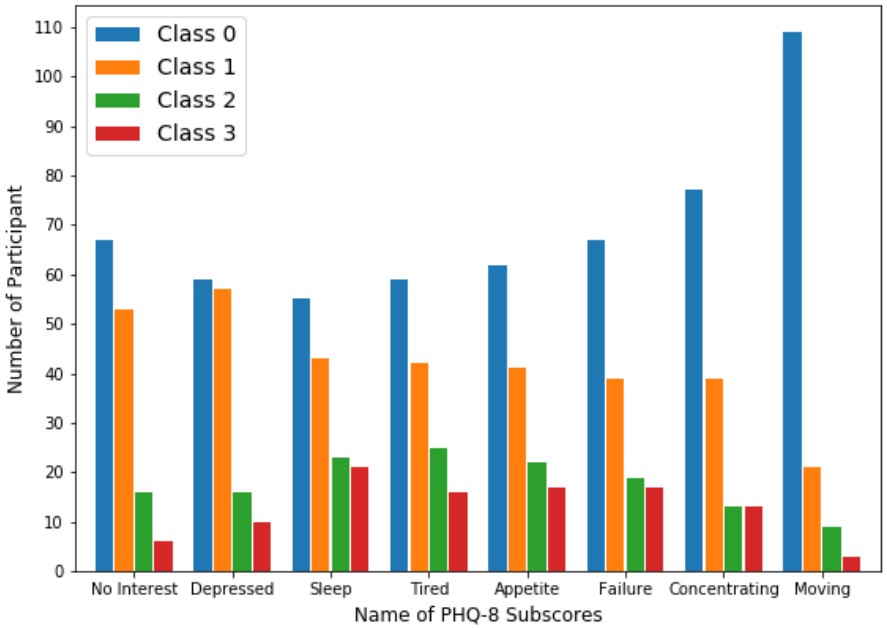
Therefore, to solve this specific problem for audio and ameliorate the phenomenon of acoustic bias or so-called gender bias, an online software tool [1] is exploited to convert each voice of the participants in the recording to the contrary gender for gender balancing, and a new audio dataset is then generated, specifically to train the backbone of the audio branch in our multi-modal model to have the better as well as non-biased capability of extracting depressive characteristic of MD.

3.3.4 Overcoming of a small-scale dataset

The other potential problem of the DAIC-WOZ depression database [9] is the scale of the database. There are only 142 sessions in the training dataset, i.e. interviews of 142 participants, which is a relatively little number of samples compared with the complexity of the depression



(a) Distribution of PHQ-8 Score



(b) Distribution of PHQ-8 Subscores

Figure 3.5: An overview of the PHQ-8 distribution. Distribution of PHQ-8 Score (a) and PHQ-8 Subscores (b).

estimation task. This small-scale dataset not only leads to a hard time to train a representative model but also incurs failures of the generalization ability of the model. This means that the model can encounter at least the following issues: overfitting, underfitting, outliers, sampling bias, missing values, etc. Hence, the sliding window technique is applied to segment the interview

into N overlapped clips and increase the dataset size, with a window size of $60\,\mathrm{s}$ and an overlap size of $10\,\mathrm{s}$.

The final outcome is illustrated in the figure 3.6. Before applying the sliding window technique, there are only 142 and 47 participants in the training dataset and test dataset respectively. After that, the scale of the dataset has been expanded tenfold, that is there are around 13 hundred clips to train the model and around 5 hundred clips to test the model's accuracy. With the help of this technique, our model shows a significant performance improvement and stability, which will be demonstrated more in detail in the Chapter 5.

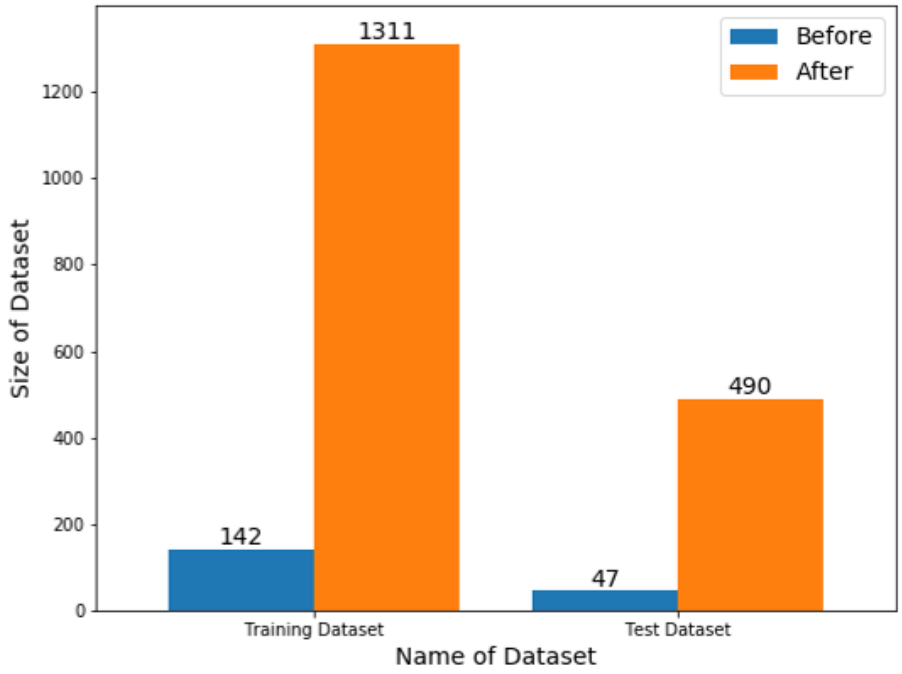


Figure 3.6: Augmentation of the dataset. The result before and after applying sliding window technique to clip and increase the scale of the dataset.

3.4 Preprocessing

So far the DAIC-WOZ database [9] has been introduced in-depth, along with potential problems such as errors and noises as well as our proposed solutions. In this section, all these techniques are going to be combined into a preprocessing framework for each data type designed to remove these errors to provide a cleaner dataset, which is also in a uniform data format, to utilize for the model training and testing.

3.4.1 Text Data

As illustrated in the Table 3.3, the contents of the transcription record the transcribed conversation between the virtual interviewer and the participant together with the timestamps as well as the speaker's name. However, since only the features of the participant are interested in this thesis,

only the features from each interviewee are extracted, namely the sentences of each participant. Furthermore, the start-stop time pair of each sentence of the participant has also been extracted as it is essential to filter and ensure that the timeline between each input data domain is roughly aligned with each other so that the late fusion will not lose its meaning. An illustration of text data preprocessing is demonstrated in the figure 3.7. In addition to sentence extraction, a further step has to be conducted to transform and represent text in a language that a computer could be able to process and understand, i.e. sentence embedding. Instead of utilizing word embedding, which deals with individual words, sentence embedding is chosen here as it not only retrieves information more efficient compared with word embedding by converting the whole sentence into a vector, but also helps the computer in understanding the context, intention, and other nuances in the entire text.

In this work, the pre-trained model of universal sentence encoder large from Google [17, 6] is exploited as our sentence embedding model. The model has a Transformer encoder-like architecture and is trained on a variety of data sources and a variety of tasks to dynamically accommodate a wide variety of natural language understanding tasks according to [17]. It encodes variable-length English text as well as sentences and outputs a 512-dimensional vector that can be used for text classification, semantic similarity, clustering, and other natural language tasks. Of course, there is also a possibility of training our sentence embedding model, which could even boost the performance. However, in our opinion, this would lead to a problem that the model would not have a generalization ability to understand different texts besides from the DAIC-WOZ depression database [9] or overfit with the text as the problem mentioned before. That is why the universal sentence encoder from Google has been chosen.

3.4.2 Visual Data

As illustrated in the Figure 3.2, the visual data consists of the micro-facial expression of facial key points and gaze direction. However, the raw data of both in the DAIC-WOZ database [9] are in fact separate, unnormalized, and need to be reformatted and cropped out the irrelevant parts. Therefore, this preprocessing for the micro-facial expression has been developed. An example of raw visual data is demonstrated in the table 3.5, including 2 sample frames of facial key points in the table 3.5a and gaze direction vectors in the table 3.5b.

As you could tell from the the Table 3.5, either facial key points or gaze direction are given in 2D format instead of 3D format, even though there are both 3D data features. 68 3D facial key points are annotated from 0 to 67 with X, Y, Z corresponding to its axis and 4 3D gaze direction contains the gaze direction vectors of both eyes in both world and head coordinates, which is marked with "h". 0 represents the right eye, whereas 1 represents the left eye. Hence, the first step of the preprocessing is to reformat them into the standard 3D coordinate format, where the last dimension represents the three axes, namely the x-axis, y-axis, and z-axis. For 68 3D facial key points, they are transformed from $(T \times 204) \longrightarrow (T \times 68 \times 3)$; for 4 gaze direction vectors, they are transformed from $(T \times 12) \longrightarrow (T \times 4 \times 3)$, where T indicates the number of the frame. Secondly, one can notice that the range of the axes in facial key points varies a lot, i.e. the x-axis from -67 to -142 but the z-axis from 505 to 565, which could potentially lead to a problem of

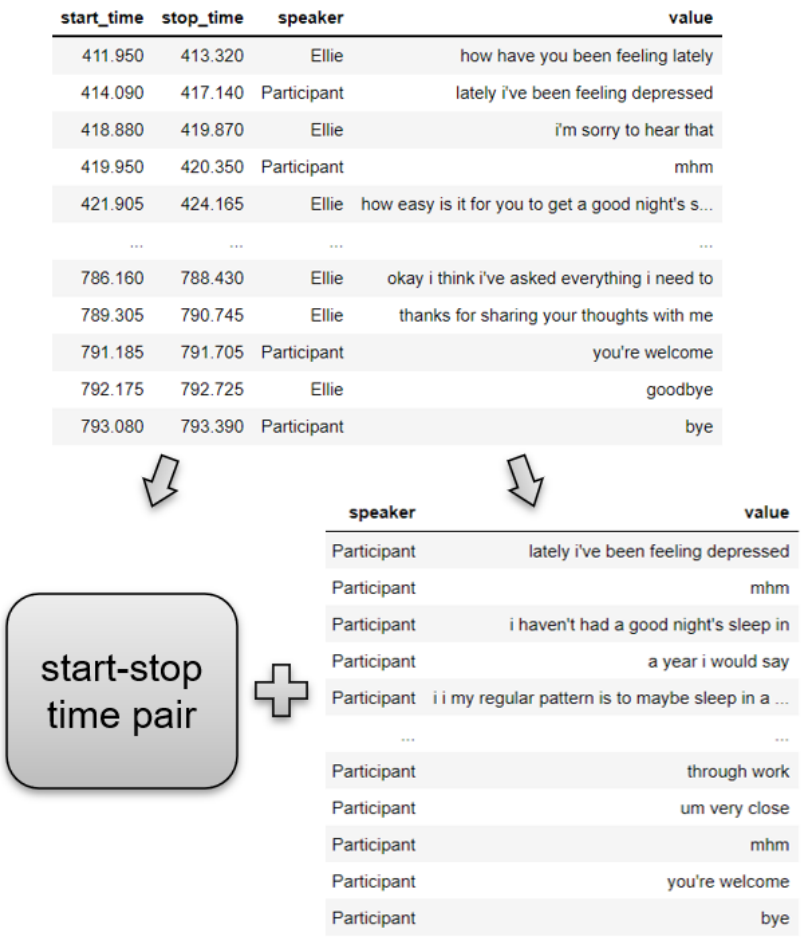


Figure 3.7: Preprocessing of text data. Sentences and timestamps of the participants are being extracted.

Table 3.5: Example of raw visual data. Coordinate of 68 3D facial key points in 3.5a and 3D gaze direction vectors in 3.5b

Frame	X0	X1	…	X67	Y0	Y1	…	Y67	Z0	Z1	…	Z67
11730	-142.2	-142.5	…	-67.6	-72.1	-51.5	…	4.2	565.9	569.6	…	505.5
11731	-141.4	-141.9	…	-67.5	-71.1	-50.7	…	4.6	565.4	568.7	…	504.6

a . 68 3D facial key points

Frame	x_0	y_0	z_0	x_1	y_1	z_1	x_h0	y_h0	z_h0	x_h1	y_h1	z_h1
11730	0.17	0.28	-0.95	-0.03	0.38	-0.92	0.08	0.38	-0.92	-0.11	0.49	-0.86
11731	0.27	0.32	-0.90	-0.05	0.39	-0.91	0.19	0.42	-0.88	-0.12	0.50	-0.85

b . 4 3D gaze direction vectors

bias between axes. Therefore, the facial key points are normalized with the following equation:

$$\mathbf{X}' = a + \frac{(\mathbf{X} - \mathbf{X}_{min})(b - a)}{(\mathbf{X}_{max} - \mathbf{X}_{min})}, \tag{3.1}$$

where the input and the output: $\mathbf{x}, \mathbf{x}' \mapsto \mathbb{R}^3$ and $a = 0, b = 1$ as the facial key points are normalized to range: 0 - 1. As for the gaze direction, no normalization is required since the given vectors either in world coordinate or head coordinate are unit vectors already. After the normalization, both of them are combined into a larger matrix with a dimension of $(T \times 72 \times 3)$. Finally, the parts where Ellie speaks as well as irrelevant interactions are cropped out based on the start-stop time pair extracted from the text preprocessing part shown in the Figure 3.7. Now, a clean, as well as normalized visual data of micro-facial expression as illustrated in the Figure 3.2, has been prepared for direct utilization of the model learning.

3.4.3 Audio Data

In the previous Subsection 3.3.2, it has been well discussed why and how the original raw audio signal is processed to achieve the final clipped raw audio signal shown in the third diagram of the Figure 3.4. However, instead of utilizing the clipped raw audio signal, a more informative and effective audio data format is preferred because the raw audio signal itself still consists of plenty of redundant segments.

The first idea is to transform the clipped raw audio signal into the log-spectrogram by applying STFT and nonlinear transformation. The outcome is illustrated in the Figure 3.8 with time (s) in x-axis and frequency (Hz) in y-axis. To observe the difference between both gender and having major depression, the following spectrograms are included: the woman without MD, the woman with MD, the man with MD, and the man without MD. At first glance, one could notice that for each spectrogram, there is always a particular light band crossing the whole spectrogram. For the

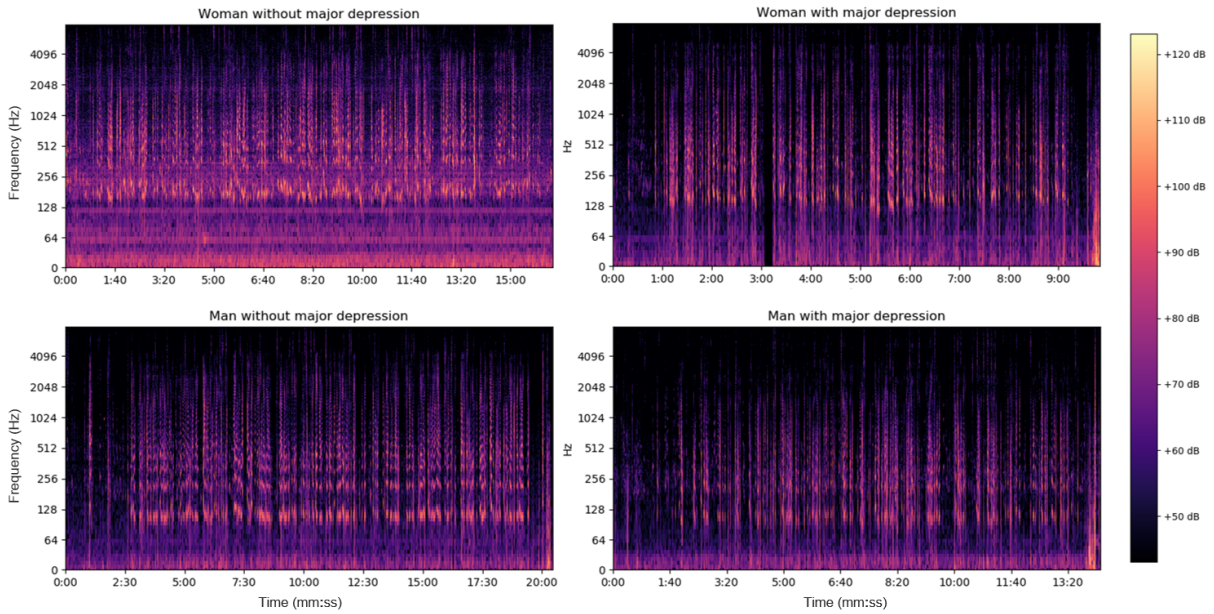


Figure 3.8: A comparison of log-spectrogram.

female spectrogram, it is around 210 Hz and for the male spectrogram around 120 Hz. In fact, these particular frequencies differed in gender are the so-called "fundamental frequency (F0)", whose values seen here are in accord with the research from Traunmüller et al. [36]. Moreover, it is observable that the spectrogram from the female with MD resembles the spectrogram from the male without MD, which, in our opinion, could potentially cause confusion to the model and thus harm the performance. Therefore, in order to increase the difference and avoid this problem, the second approach is come up, i.e. log-mel spectrogram.

Log-mel spectrogram differs from spectrogram by converting the frequency scale in spectrogram to the so-called mel-scale, which is a more human perception-like frequency scale designed by Stevens et al. [34]. As you can imagine, it is not a huge challenge for us humans to distinguish the voice of both genders. Hence, by exploiting this technique, there is no doubt that the subtle differences in Figure 3.8, especially between the female with MD and the male without MD, will rise and the complication will be ameliorated to the greatest extent possible. Illustrations of the converted log-mel spectrograms of identical individuals as in spectrograms are shown in the Figure 3.9. This time, by perceiving the apparent difference of F0 from both genders and comparing the log-mel spectrogram between the female with MD and the male without MD, it can be concluded that this approach brings great success in solving this issue. Moreover, based on the observation, the color of the log-mel spectrogram of the participants with MD is darker than the participants without MD, indicating that they usually speak quietly and less energetically, which could be a potential characteristic of having major depression.

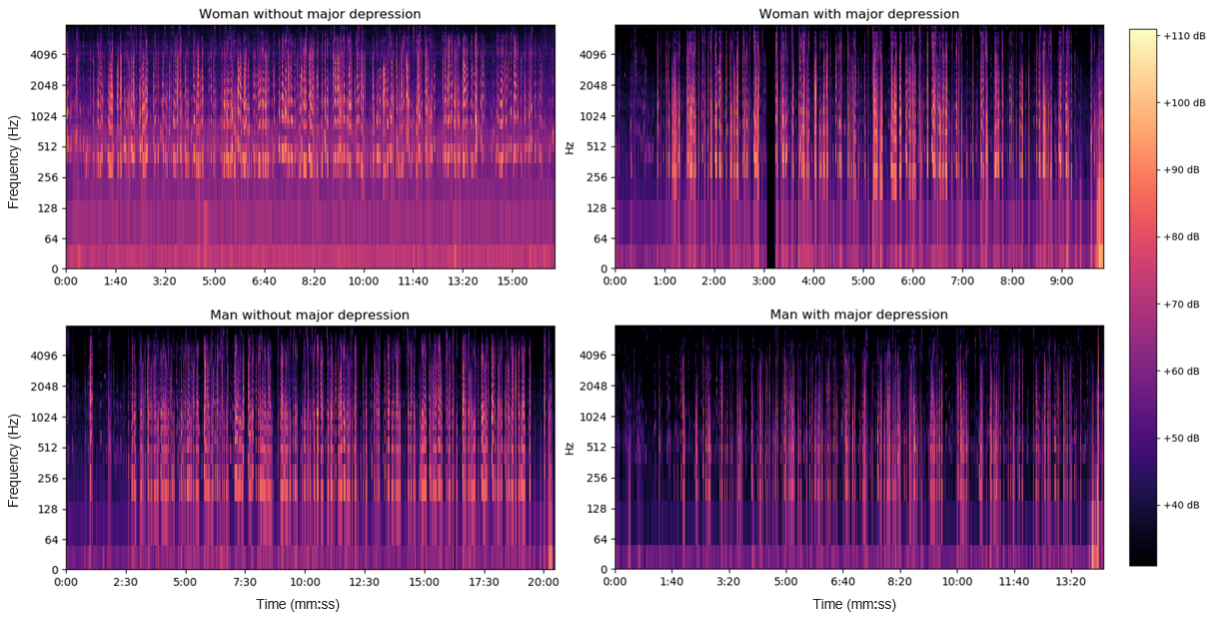


Figure 3.9: A comparison of log-mel spectrogram.

Finally, after taking all the above-mentioned perspectives into consideration, the preprocessing framework for the acoustic features could be summarized in the following 4 steps:

1. Cropping out all the parts from Ellie and noises form the original raw audio signal to form the clipped raw audio signal based on start-stop time pair extracted from the transcription.

2. Applying STFT to extract spectrogram from the clipped raw audio signal, and transforming it to mel spectrogram by converting the frequency scale to mel-scale. Here for STFT, Hann function was used with a window size of 2048 and hop size of 533, which is derived from the sampling rate of audio data and visual data, namely $\frac{\text{acoustic sampling rate}}{\text{visual sampling rate}} = \frac{16\,\text{kHz}}{30\,\text{Hz}}$. The number of the mel band for mel-scale is 80.

3. Converting the mel-spectrogram with linear transformation to log-mel spectrogram.

4. Standardizing the log-mel spectrogram with the following equation:

$$\mathbf{X}' = \frac{\mathbf{X} - \mu}{\sigma}, \tag{3.2}$$

where μ: mean of input $\mathbf{X}$ and σ: standard deviation of input $\mathbf{X}$.

3.5 Dataset Generation

Throughout this work, three different datasets have been generated for the various purposes in the experiment, i.e. raw dataset, clipped dataset, and audio dataset. By summing up the points as well as processes mentioned in the Section 3.3 and 3.4, three different scripts are created to automatically generate these datasets, which also indicates that three data loaders are created to load each dataset. In this section, details, as well as the objectives of each dataset, will be explained.

3.5.1 Raw Dataset

The raw dataset is the original DAIC-WOZ depression database [9] without any above-mentioned feature engineering techniques such as data clipping, gender balancing, sampler, etc. For this dataset, the data loader loads the data based on the participant, meaning that for a batch size of 10, the whole recording of 10 different individuals including visual data of micro-facial expression, audio data of log-mel spectrogram, and sentence embedding of transcription will be loaded into the model for model learning and the result will then be predicted. At first glance, this seems to be logical because this is exactly how the interviewers perceive and diagnose whether a patient has MD or not. They interrogate the participant with several questions, meanwhile, observing his or her reaction as well as the expression. After the whole interview is finished, the decision is then made. However, for computers, one problem that could directly be realized is that this loaded data size is too enormous, leading to several difficulties, e.g. inefficient training process, infinite time to train, running out of memory problems, etc. The interview length is ranging from 7 minutes to 33 minutes with an average of 16 minutes and it costs at least $25\,\text{GB}$ to load each data type. Moreover, since no techniques have been applied to solve and handle the imbalanced and small-scale dataset problems. It is obvious and to anticipate that training the model, either multi-modal or single-modal, based on this raw dataset will fail in all experiments. Indeed, proven by the results from the failure of the experiments in the Section 5.2, in which the model only predicts class 0 (non-depressed) as the result no matter which participant is being

processed, it can be concluded that the raw dataset fails and thus is being abandoned in the work. That is why our focus is turned to the second generated dataset, namely the clipped dataset.

3.5.2 Clipped Dataset

As opposed to the raw dataset, the clipped dataset is the edited original DAIC-WOZ depression database [9] by applying various feature engineering techniques mentioned above in the Section 3.3 and 3.4. Moreover, in order to reduce the computation load while model training, the sliding window technique is utilized to clip the whole interview into several segments with a window size of $60\,\mathrm{s}$ and an overlap size of $10\,\mathrm{s}$. The data loader, in this case, loads the data based on clips instead of participants, implying that the model predicts the depression status of each clip rather than each participant. If an interview of a participant has been split into 10 clips, these 10 clips will be loaded into the model, possibly in different batches since the weighted random sampler is exploited to randomly as well as equally load the clips from each different class, which will then outcome 10 groups of predicted PHQ-8 Subscores for each clip even though there are all from the same participant. The pseudocode of the script to automatically generate the clipped dataset is shown in algorithm 1.

Algorithm 1: Generation of clipped dataset

Input : $dataset\_path$, $output\_path$
Output : A new generated dataset

Initialize new_GT, $window\_size$, $overlap\_size$
$participants$ = ***LoadPaticipants***($dataset\_path$)
for $participant$ *in* $participants$ **do**
  $GT \leftarrow$ extracting the GT of current participant
  $visual\_path$, $audio\_path$, $text\_path \leftarrow$ getting all file paths of each data type
  % Loading all data type
  $visual\_features$, $visual\_sr$ = ***LoadVisual***($visual\_path$)
  $audio\_features$, $audio\_sr$ = ***LoadAudio***($audio\_path$)
  $text\_features$, $time\_pair$ = ***LoadText***($text\_path$)
  % Clipping the data
  $clipped\_visual$ = ***VisualClipping***($visual\_features$, $visual\_sr$, $time\_pair$)
  $clipped\_audio$ = ***AudioClipping***($audio\_features$, $audio\_sr$, $time\_pair$)
  $mel\_spectrogram$ = ***Normalize***(***ConvertMel***($clipped\_audio$, $audio\_sr$, $frame\_size = 2048$, $hop\_size = 533$, $num\_mel\_bands = 80$))
  % Applying sliding window technique and storing the segments
  $num\_segment$ = ***SlidingWindow***($clipped\_visual$, $mel\_spectrogram$, $text\_features$, $visual\_sr$, $window\_size$, $overlap\_size$, $output\_path$)
  $new\_GT \leftarrow$ storing the $num\_segment$ times duplicated GT in $output\_path$
end

After turning to the clipped dataset instead of the raw dataset, great success has been shown in

all experiments. Therefore, the clipped dataset is predominantly utilized in the experiment.

3.5.3 Audio Dataset

The third dataset that is utilized throughout this thesis is the audio dataset. As mentioned before, this audio dataset is specifically generated for solving the gender bias phenomenon in acoustic features. To generate it, the gender balancing technique pointed out in the Subsection 3.3.3 has been deployed by converting the audio recording of a participant to the contrary gender voice. Hence, for each participant, there are now 2 raw audio signals available, namely, the original one and the transgender one. These 2 raw audio signals are then passed through the preprocessing steps listed in the Subsection 3.4.3 to extract the final, normalized log-mel spectrograms, and additionally, similar to the clipped dataset, the sliding window technique is exploited to cut the whole log-mel spectrograms into several clips to reduce the computational load later in the training process.

With such a deployment of the gender balancing technique to address this problem, the impact of the gender bias could be maximally alleviated and the model could have the true competence of estimating the depression severity solely according to the acoustic depression characteristic and ignoring the gender differences. Therefore, this newly generated audio dataset is used particularly to train the audio backbone either in the single modality audio model or in the multimodality model. In fact, proving by the result of the experiments, a significant performance improvement can be observed after training with this dataset.