Master's Thesis of Cognitive Science

Automatic classification of major depressive disorder using the optimal speech feature combination

최적의 음성특징 조합을 활용한 주요우울장애 자동분류

February 2024

Graduate School of Humanities
Seoul National University
Interdisciplinary Program of Cognitive Science

Hyunsun Ham

Automatic classification of major depressive disorder using the optimal speech feature combination

Examiner Jun-Young Lee

Submitting a master's thesis of
Cognitive Science

February 2024

Graduate School of Humanities
Seoul National University
Interdisciplinary Program of Cognitive Science

Hyunsun Ham

Confirming the master's thesis written by
Hyunsun Ham
February 2024

Chair _____(Seal)
Vice Chair _____(Seal)
Examiner _____(Seal)

Abstract

This study focuses on finding the optimal combination of speech features to detect major depressive disorder(MDD). Many works have already shown the utility of voice biomarkers for automatic detection of MDD using various speech tasks. Despite the utility of spectral features for MDD detection, there is no consensus of which speech subsets hold relevant characteristics to explain the clinical symptoms of MDD. In this study, we examine the classification performance of the different speech dimensions to verify the most discriminative speech indicators and find the optimal combination using several speech feature subsets and validate their predictive capability using a BDI prediction model.

Voice of reading out pre-defined paragraphs was extracted from both 72 depressed adults and 70 healthy controls. 210 speech features were extracted from each audio recording and grouped into four speech subsets: spectral features, prosodic features, voice quality, and formants. Extracted features were selected based on Recursive Feature Elimination(RFE). The criteria for feature selection were based on importance scores calculated using the Extreme Gradient Boosting(XGboost). We then evaluated the classification performance of each subset individually and assessed

the classification performance of all possible combinations of speech feature subsets.

In the analysis of individual speech subsets, the spectral features demonstrated high performance in the classification of the two distinct groups. The spectral features were then used as the baseline features to examine all possible feature combination including the baseline. Among the seven possible combinations, the combination of spectral and prosodic features outperformed all other subset combination with an F1 score of 0.83. This addresses the combined synergy of spectral and prosodic features could be reliable combination to identify MDD. To evaluate the diagnostic utility of the optimal combination, we built a Beck Depression Inventory (BDI-II) prediction model, which obtained a Mean Absolute Error of 7.19. The further investigation into the correlation between selected speech features and BDI-II subscale scores, based on the BDI-II two-factor model, revealed a notable association between somatic factors and several speech indicators in spectral and prosody subsets. This suggests that depressive speech may potentially linked to various clinical symptoms and subtypes of depression, particularly those associated with the somatic factors. The overall findings highlight the significance of feature-based analysis in clinical speech research. Future studies

should explore the psychological and neural foundations associated with these relevant speech features.

Keyword : major depressive disorder, optimal speech feature combination, speech subsets, automatic classification, speech biomarkers

Student Number : 2022-24910

Table of Contents

Chapter 1. Introduction 1

1.1. Background 1

1.2. Structure of the thesis 4

Chapter 2. Literature Review 5

2.1. The importance of early diagnosis of MDD 5

2.2. Speech biomarkers 6

2.3. Depressive speech 8

2.4. Speech-based automatic detection of MDD 11

Chapter 3. Methods 13

3.1. Systematic Workflow 13

3.2. Participants 14

3.3. Speech task 15

3.4. Speech subsets and feature extraction 16

3.5. Feature selection 19

3.5. Model training 22

Chapter 4. Experiments and Results 24

4.1. Single speech feature subset 24

4.2. Speech feature subset combination 26

4.3. BDI Prediction 30

4.4. BDI two-factor analysis 31

Chapter 5. General Discussion 36

Chapter 6. Conclusion 41

References 43

Appendix 48

Appendix 1 48

Appendix 2 50

Abstract in Korean 51

List of Tables

Table 1. Demographics of participants 15

Table 2. Extracted features: components and statistics 19

Table 3. Selected speech features 22

Table 4. Classification results for single speech subset 25

Table 5. Correlations between speech features and cognitive-affective factor scores 31

Table 6. Correlations between speech features and somatic factor scores 34

List of Figures

Figure 1. Systematic workflow.....13

Figure 2. Number of speech features selected.....20

Figure 3. Features selected by importance scores 21

Figure 4. ROC curve of SVM classifier.....29

Figure 5. Scatter plot of actual and predicted BDI scores ...30

Figure 6. BDI-II two-factor model.....33

Figure 7. Scatter plot of correlation between two-factor scores and speech features.....35

Chapter 1. Introduction

1.1. Background

Major depressive disorder (MDD) is a complex mental health condition marked by a spectrum of primary symptoms such as persistent low mood, anxiety, restlessness, and a deep emotional numbness. When these fundamental symptoms co-occur with additional cognitive symptoms like negative attitude and cognitive distortions, the condition is categorized as MDD. These cognitive distortions typically involve detrimental self-assessments and pessimistic expectations for one's future. Consequently, it aggravates the quality of one's life and severely affects the socio-economic burden in terms of diagnosis, early intervention, and rehabilitation (Cummins et al., 2015). The COVID-19 pandemic has significantly intensified existing mental health challenges, leading to a marked increase in depression and suicidal ideation. The widespread impact of the pandemic has forced people into isolation due to the essential lockdowns and social distancing protocols. In South Korea, the experience of COVID-19 quarantine has been associated with a higher risk of depression and an increase in depressive symptoms, in comparison to those not subjected to quarantine. Kim et al.(2022) suggests that the period of enforced

isolation during the pandemic has had negative psychological consequences, including a spike in depression rates among the population.

The current clinical diagnostic process is generally time-consuming and subjective in nature as there is no single clinical characterization of depressive symptoms, and many of the assessments are based on self-inventory evaluation (Cummins et al., 2015; Robin et al., 2020; Silva et al., 2021). It is significant to acquire a golden time of early intervention and treatment which could also assist lessening the risk of suicidal behaviors (Cummins et al., 2015; Stasak et al., 2021). Thus, there is a pressing need to find cost-effective and scalable markers that can be easily measured and operated by automated models (Tasnim & Novikova, 2022).

Multiple studies have identified some potential biomarkers for depression detection. Speech signal is one such marker that appears to be a reliable modality. Specific acoustic markers from human speech make a significant gap between normal and depressive speech. Spectral information is one of the most widely and frequently used for depressed speech discrimination(Cummins et al., 2011). Detailed spectral information can be captured through various features, including Mel frequency cepstral coefficients

(MFCCs). This information generally offers considerable insight into speaker's identity and is easily quantifiable and cost-effective for computational analysis. Recent methodologies of classifying clinical speech are much focused on end-to-end learning (Robin et al., 2020; Tasnim & Novikova, 2022), in which the model is fed up with raw speech audio and the features are automatically extracted and selected through black-box system. These pre-trained models have consistently achieved higher performance, however, there are insufficient findings of which speech features are key contributors to explain depressive symptoms. As further investigation on feature-based approach is still needed particularly in clinical speech, where speech could not be explained alone by a few parameters, we address the following research questions:

1) Do spectral features hold the most discriminant power in depressive speech?
2) What is the optimal combination of speech subsets for depression detection?

1.2. Structure of the thesis

In the following sections of the thesis, Chapter 2 reviews previous findings in the domain of detecting depression through speech analysis. Chapter 3 demonstrates the proposed methods, including details on data collection and feature engineering. Chapter 4 details the experiments and their outcomes, which includes an evaluation of the classification performance of each single speech feature subset, as well as the combined effectiveness of all possible combinations of speech feature subsets. In this chapter, the optimal combination and its corresponding numerical performance are presented. We also present a BDI-II prediction model, using the feature subset combination proven to be optimal through the experiments. At the end of this chapter, we further investigate the correlations between BDI-II subscale scores and selected speech features based on the two-factor structure of BDI-II. Chapter 5 demonstrates the general discussion based on the results and findings from the experiments, and the conclusions are summarized in Chapter 6, along with a proposal for further research that can be explored in the field of clinical speech.

Chapter 2. Literature Review

2.1. The importance of early diagnosis of MDD

Like other mental illnesses, MDD benefits from early diagnosis and intervention. However, a significant challenge lies in the lack of awareness among both affected individuals and those around them, leading to a potential delay in appropriate intervention. The analysis of the prevalence of MDD across different life stages reveals a U-shaped pattern, with the highest rates observed between the ages of 18 and 29, decreasing through middle age, and then increasing again in the later life (Lee et al., 2017). This highlights the importance of ongoing societal attention to enable early screening and intervention, especially during the stages of youth and old age. The exact pathology of MDD remains uncertain, with discussions suggesting a diverse range of factors such as biological, genetic, hormonal, and environmental influences (Nemade et al., 2019). Given the absence of a single pathology, the heterogenous nature of MDD makes the clinical decisions in screening, diagnosis, and intervention even more difficult (Bembnowska & Jośko-Ochojska, 2015). Early diagnosis and intervention for psychotic disorders are crucial to prevent significant functional decline and the high risk of mortality during a critical period (Davey & McGorry, 2019). MDD,

which often begins in early life and commonly experienced throughout adulthood, may offer a strategic window for interventions with lasting effects. The primary objective of many early interventions in psychiatry is to provide treatments at the earliest possible stage (Davey & McGorry, 2019). Failing to intervene at an appropriate period may result in increased severity and substantial aggravation of the symptoms, making it even more challenging to enter the effective treatment. To avoid such risks, it is essential to establish an efficient system and framework for the overall procedure of diagnosis and intervention of MDD. This involves implementing timely interventions based on the severity level, supported by effective diagnostic tools.

2.2. Speech biomarkers

A biomarker refers to objectively measurable characteristics utilized to assess biological responses to normal and pathological processes as well as therapeutic interventions (Cummins et al., 2015; Robin et al., 2020). In clinical trials, biomarkers play a vital role in the rational development of medical diagnostics and therapeutics (Califf, 2019). According to the report by Califf(2019), it is essential to understand the difference between the concept of biomarkers and clinical outcome assessment(COA). COAs typically

measure the way people feel and function, representing outcomes directly relevant to them. This distinction is important as the COAs are designed to meet specific criteria for the regulatory approval of therapeutics. On the other hand, biomarkers serve various purposes, notably having a potential to predict COA measurements through physiological signals. Some non-invasive biomarkers including voice, facial expressions, and gaze tracking are readily accessible and obtainable in daily life. Consequently, the use of biomarkers can be advantageous in overcoming the limitations of COAs.

In recent years, digital biomarkers have emerged, leveraging various digital devices such as smartphones and wearables to collect information. This technological advancement has expanded the boundaries of traditional measurements for diagnosis and prescription. Human voice is well-known biomarkers that clinicians have long used as diagnostic bases. Speech provides valuable insights into cognitive and motor functions, which are often affected in various psychiatric and neurodegenerative diseases (Robin et al., 2020). The complexity of speech, involving multiple cognitive and motor processes, makes even a brief speech sample a sensitive indicator of cognitive health and functioning across various illnesses. For instance, the speech characteristics of neurodegenerative disorders such as Parkinson's disease and other conditions

characterized by motor impairment, have already been extensively investigated in prosody research. Prosody, as a unique yet essential component of spoken language, is closely linked to neural networks that underpin language functions (Nevler et al., 2019). It plays a crucial role in the phonological representation for specific words in the auditory–aural system and primarily contributes to suprasegmental aspects of sentence processing (Ash et al., 2013; Neveler et al., 2019). Despite the ongoing challenges in standardization and quantification due to linguistic variations and the absence of universally accepted metrics, utilizing speech markers has the potential to redefine the framework of diagnosis in the clinical field, which have traditionally relied on subjective criteria for diagnosis.

2.3. Depressive speech

For many years, extensive research has focused on speech and its connection to psychomotor disturbances linked to mental disorders. Depressed speech has often been described as lacking liveness, being monotonous, and having a flat tone. These perceptual characteristics have been associated with acoustic variations related to several metrics, such as fundamental frequency, formant structure, power distribution, or amplitude modulation (France et al.,

2000). In recent years, the concept of depressive speech has been widely investigated through interdisciplinary research in speech science and speech technology. Many attempts to define the depressed speech and quantifying its speech signal at a precise level through vector-based measures have therefore allowed for a more concrete understanding within various speech subsets. There are several speech features typically extracted and analyzed from raw speech sample in a short-term time scale, considered as relevant markers to capture commonly occurring vocal effect (Cummins et al., 2015; Robin et al., 2020), as well as associated symptoms such as increased anxiety, intense affective states, and low mood (Cummins, et al., 2015).

Spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) are well-known features for depressive speech analysis. MFCCs were shown to reflect vocal tract changes and have been widely introduced as useful features in speech recognition (Wang et al., 2019; Williamson et al., 2014). Previous studies in the speech engineering field have revealed the spectral features as informative factors particularly in detecting paralinguistic and emotional information of human speech (Wang et al., 2019). It was found that prosodic features are another relevant speech characteristics to support certain behavioral cognitive

symptoms of MDD (Scherer et al., 2013). The basic frequency and loudness range of depressed individuals were found to be decreased than that of normal group (Yang et al., 2012). As the severity of depression increases, the fundamental frequency(F0) range declines, resulting in monotonous speech (Cummins, et al., 2015; Yang et al., 2012).

Voice quality features are also reported as useful measures in depressive speech analysis. Voice quality measures include jitter, shimmer, and harmonic-to-noise ratio(HNR), which are found to be influenced by vocal fold tension and subglottal pressure (Afshan et al., 2018; Silva et al., 2021). The effect of depression on formants is also widely supported by prior works. Formant features are reported as distinguishable features for depression which reflect psycho-motor retardation as a representative symptom with tightening of the vocal tract (Cummins et al., 2017). Recent findings have provided insights into clinical speech research using the above speech features with physiological and psychological evidence. Among these features, spectral features are commonly used to distinguish the depressed group with a sensitivity of 77.8% and succeeded in predicting Hamilton Depression Rating Scale (HAMD) scores (Cummins et al., 2015).

2.4. Speech-based automatic detection of MDD

The use of speech biomarkers marks a significant leap forward in predicting and distinguishing mental disorders. This innovative approach holds the potential to improve diagnostic strategies by leveraging the extensive accessibility of smart devices to address the inherent variability in individual cases (Brietzke et al., 2019). Leveraging digital biomarkers allows us to transform the way we diagnose mental health conditions. By implementing a fully automated process that utilizes easily digitized data such as voice recordings, we can develop a novel diagnostic framework that is both more accurate and cost-effective.

Many studies have assessed depressed and non-depressed speech using two common approaches: the feature-based approach and transfer learning((Balagopalan & Novikova, 2021). The feature-based approach explores clinically relevant acoustic and linguistic features from both audio recordings and transcripts of recorded speech. This domain-knowledge based approach has been widely used to investigate novel feature sets, which offers benefits of interpretable model decisions, representation of speech in various modalities, and reduced computational resources. With transfer learning, on the other hand, features are no longer manually extracted and selected. It primarily employs powerful structural

mechanisms and uses a pre-trained model, which potentially achieves in higher performance without the need for extensive feature engineering. However, the transfer-learning approach stills holds the opaque nature in its explanation and interpretation.

This study primarily focuses on exploring how clinical characteristics of MDD manifest within specific speech features over the pursuit of end-to-end model superiority. While acknowledging the potential of comprehensive models, our focus is rooted in a feature-based approach that seeks to delineate the acoustic signatures of MDD. This method enhances the interpretability of our findings, ensuring that the decision-making process of our model is not just a byproduct of a black-box algorithm, but is instead transparent and directly linked to clinically observable phenomena. In light of the need for a solution that is both effective and feasible for real-world application, this study emphasizes the need for a model that balances precision with computational efficiency. By selecting and identifying key features, we aim to craft a model that is both cost-effective and suitable for widespread deployment, particularly in the context of clinical settings.

Chapter 3. Methods

3.1. Systematic Workflow

The experimental procedure consists of two major parts: feature engineering and classification. The proposed method is aimed at building the classification model, in which the input data is acoustic feature vectors from read speech data and the output is a binary result of the subjects' depression status. In this chapter, we first describe how the clinical speech data was collected, then we demonstrate the procedure of feature engineering including feature extraction and selection. Finally, we propose several classifiers trained for binary classification to detect depressive speech. The overall workflow is shown in Figure 1.

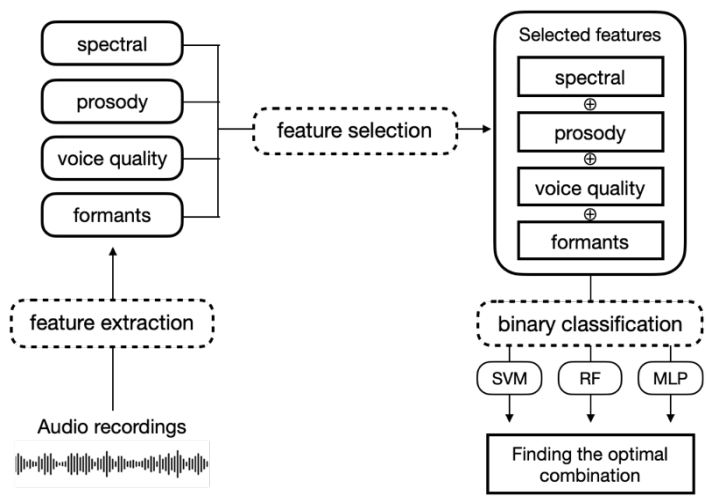

Figure 1: Systematic workflow

3.2. Participants

Seventy-two native Korean speakers(34 male, M_{age} = 25.1; SD = 2.8, range = 20-30) diagnosed with major depressive disorder and seventy paired controls(29 male, M_{age} = 24.5; SD = 3.3, range = 20-30) without any history of mental illnesses participated in the experiment. All subjects enrolled at Seoul National University Health Service Center located in Seoul, South Korea provided informed consent before participating in the experiment. Participants with MDD had an average of 16.9 years of education (SD=2.1), while the control group averaged with 16.3 years of education (SD=2.1). To measure the severity of symptomatology, both depressed and control groups responded to the Korean Beck Depression Inventory II(BDI-II) questionnaires, which is a 4-point self-rated measure for depressive symptoms including 21 questions. The average score of MDD group was 26.0 (SD=6.9), typically falls within the mild to moderate range of depression on average, whereas the control group showed an average score of 9.4 (SD=3.5). The demographics of participants is shown in Table 1.

Participants were required to meet specific criteria to be included in the experiment. Those with the following criteria were excluded: individuals with severe physical or cognitive conditions that could significantly affect the assessment process, patients with

profound sensory impairments, including significant hearing or vision loss, and those with substantial language disorders. Additionally, non-native individual with insufficient Korean language proficiency to understand the guideline of the speech tasks were not included in the study.

Table1. Demographics of participants

	Group		*p*-value
	CON(n=70)	MDD(n=72)	
Male/female	29/41	34/38	
Age	24.5±3.3	25.1±2.8	0.64
Years of education	16.3±2.1	16.9±2.1	0.58
BDI-II total scores	24.5±3.3	24.5±3.3	0.03

CON: healthy controls, MDD: major depressive disorder, BDI-II: Beck Depression Inventory-II

3.3. Speech task

All participants were to read three different paragraphs. In many clinical speech studies, spontaneous speech tasks are commonly used for assessing actual speech capabilities, but for the observation of target phonemes and standardized speech features, the study utilizes read speech tasks. Read speech tasks offer the

advantage of being highly convenient for phonation, acoustic characteristics, naturalness, and syntactic boundaries.

As the current study aims to find acoustic patterns rather than a higher-level of linguistic features(e.g. semantic or pragmatic level) which could be effectively analyzed in spontaneous speech, read speech tasks were used for standardized conditions while speaking aloud. The paragraphs were Korean Standardized Passage(Kim, 1996) balanced Korean vowels and consonants. These three pre-defined paragraphs had different topics(autumn, travel, the wind, and the sun) and each of the paragraph consists of 5 to 9 sentences, 56 to 129 syllables (The pre-defined paragraphs are attached in Appendix 1). The speech collection was consistently recorded at 15cm. The paragraphs were shown at an ordered sequence and every five seconds were given between each paragraph. The collected speech dataset consists of 426 audio samples.

3.4. Speech subsets and feature extraction

To identify the general patterns of depressed speech, the current study utilizes four defined speech feature subsets according to the prior research: spectral features, prosodic features, voice quality,

and formants. For feature extraction, we use Surfboard (Lenain et al., 2020), an open-source Python package for audio feature extraction in clinical conditions. It has a significant overlap with conventional libraries such as OpenSMILE (Eyben et al., 2013) and Praat (Boersma & Van Heuven, 2001), but it provides simple Python interface. Furthermore, the features can be selectively extracted in both low-level descriptors (LLD) and statistical functional level. LLDs are computed on frame-by-frame basis, and statistical functionals such as mean, median, maximum, and standard variation are computed on the low-level descriptors. A total 210 acoustic features were extracted from each audio sample and the feature set is categorized into four speech subsets. The table 2 shows extracted components grouped by four conventional subsets.

The spectral features involve general components of time series such as MFCC 1-12, which capture the spectrum of sound signals. As mentioned above, MFCCs are extensively used in speech and audio processing and have been reported as relevant information related to depressive voice (Rejaibi et al., 2022). Spectral kurtosis provides insights into the shape of the spectral distribution, while entropy quantifies disorder in the spectral components. Other measures such as spread, rolloff, skewness, centroid, and flux further characterize the spectral distribution,

providing comprehensive information about its width, frequency content, asymmetry, center of mass, and dynamic changes. In the prosodic features, the study mainly focuses on features related to speech expression. F0(pitch), duration, and intensity(energy) provide information about intonation, temporal characteristics, and vocal strength, respectively. Prosodic components such as pitch period entropy, log energy, and loudness contribute to the understanding of pitch variability, overall signal strength, and sustained phonations (Lenain et al., 2020). The voice quality features include jitter, shimmer and HNR. The level of jitter is primarily affected by a lack of control of vibration of the vocal cords. The voices of individuals with certain pathologies generally have an increased percentage of jitter (Teixeira et al., 2013). Shimmer measures the variations in amplitude between consecutive vocal cycles. HNR is another frequently used measure that signifies the balance between harmonics and non-harmonic noise in a voice signal. The formant subset includes resonant frequencies(F1-F4) in the vocal tract that contribute to the perception of vowel sounds. Changes in these formants(ΔF1-4) over time offer information about the dynamic nature of speech. Sliding-window formants capture formant frequencies within specific time windows, providing a temporal analysis of resonant frequency changes. Statistical

measures, including mean, standard deviation, maximum, minimum, first-derivative mean, and first-derivative standard deviation values were applied to each component in each speech subset.

Table 2. Extracted features: *components and statistics*

Subset	Components	Statistics
Spectral	MFCC 1-12, Spectral kurtosis, entropy, spread, rolloff, skewness, centroid, flux	mean, standard deviation, max, min, first-derivative mean, first-derivative standard deviation
Prosody	F0(pitch), duration, intensity, pitch period entropy, log energy, loudness	mean, standard deviation, max, min, first-derivative mean, first-derivative standard deviation
Voice quality	jitter, shimmer, HNR	mean, standard deviation, max, min
Formants	F1- F4, ΔF1-4, sliding-window formants	mean, standard deviation, max, min

3.5. Feature selection

Feature selection is significant to select the most relevant features with respect to clinical symptoms, minimizing redundancy in the

selected set of features (Tasnim & Novikova, 2022). We applied Recursive Feature Elimination (RFE) algorithm for feature selection. RFE is an algorithm that starts by incorporating all features, then iteratively eliminates less important features one by one while retraining, thus selecting significant features (Theerthagiri & Vidya, 2022). The eliminated features were selected by the importance score calculated by the Extreme Gradient boosting(XGboost), which finally left 3-5 important speech features within each speech feature subset. XGboost, an advanced machine learning algorithm, enhances RFE by providing a robust method for ranking the importance of features. The number of features was trained to be a minimum of two or more.

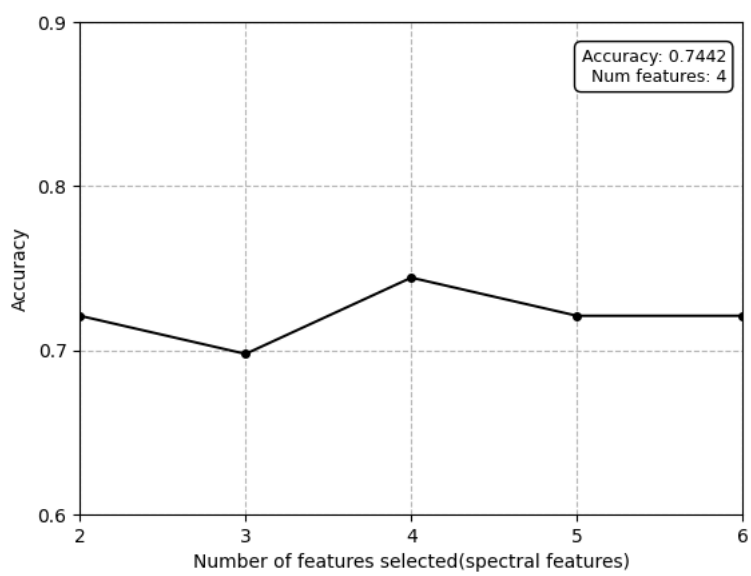

Figure 2. Number of speech features selected

It employs a sequence of decision tress where each tree is built to correct the errors of its predecessor, thereby enhancing the model performance iteratively (Zhang et al., 2022). This process uses a quantitative measure of 'importance' of each feature, determined by how much each feature contributes to the predictive accuracy across the tress. Figure 2 and 3 illustrates the spectral features selected through RFE-XGboost, with Figure 2 showing the number of features selected and Figure 3 detailing the specific features and their importance scores. The selected features from each subset are listed in Table 3.

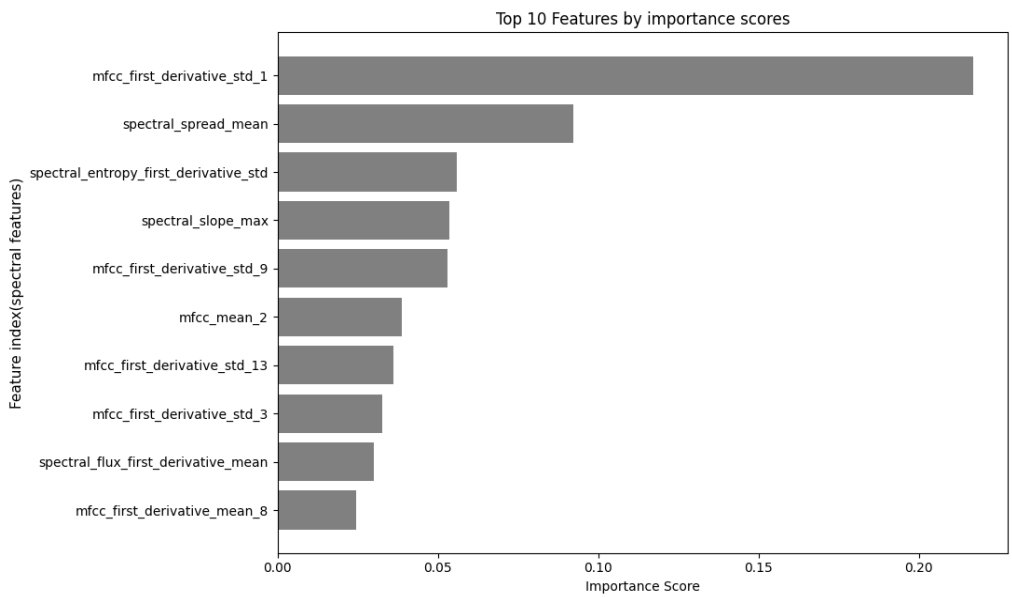

Figure 3. Features selected by importance scores

Table 3. Selected speech features

Subset(#number)	Selected features
Spectral(4)	mfcc_first_derivative_std_1
	spectral_spread_mean
	spectral_entropy_first_derivative_std
	spectral_slope_max
Prosody(3)	loudness_slidingwindow_max
	log_energy
	intensity_first_derivative_mean
Voice quality(5)	localabsoluteJitter
	apq11Shimmer
	localdbShimmer
	apq3Shimmer
	HNR
Formants(3)	F2_mean
	F1_first_derivative_std
	F2_max

3.6. Model Training

We train three machine learning models, partly following (Tasnim & Novikova, 2022), support vector machine(SVM), random forest(RF), and multi-layer perceptron(MLP). These models have been proposed as robust and cost-effective solutions for processing conventional acoustic features. In pursuit of optimal performance,

the hyperparameter values of SVM and RF are tuned following Balagopalan & Novikova(2021) and Tasnim & Novikova(2022). Instead of using the deep neural model(FNN) presented by Tasnim & Novikova(2022), we use MLP given the non-linear attributes of speech signals. The MLP consists of two hidden layers, both trained by ReLU activation function. The model is trained for 50 epochs and binary cross entropy is used for loss calculation. The output layer is optimized by sigmoid function (Sun et al., 2022). To ensure a valid evaluation and mitigate the effect of limited amount of data, a 5-fold cross validation is used, without any speaker overlap between training and testing data. The model performance is evaluated by several metrics: accuracy, recall, and F1 score.

Chapter 4. Experiments and Results

4.1. Single speech feature subset

Under the binary speaker-independent scenario, namely depressed or non-depressed, each of the four single speech subset and the combinations of speech subset was fed up with three classifiers. The classification results for single speech subset are presented in Table 4. Out of four subsets, the spectral features showed the most discriminant performance as expected. The SVM classifier achieved a recall of 0.73, an F1 score of 0.75, and an accuracy of 0.74 in spectral features. The prosodic features followed, with a recall of 0.69, an F1 score of 0.64, and an accuracy of 0.71 in the same classifier. Voice quality and formants demonstrated lower performance, with voice quality achieving a recall of 0.61 in SVM, 0.63 in RF, and 0.70 in MLP. Among all classifiers, the MLP demonstrated the highest efficacy, particularly with the spectral feature subset, which achieved a recall of 0.82, an F1 score of 0.80, and an accuracy of 0.84. In the same classifier, the prosody subset also demonstrated robust results with recall, F1, and accuracy scores of 0.80, 0.78, and 0.81, respectively. The spectral features consistently outperformed other subsets across all classifiers, indicating its potential as a reliable indicator in speech-

based detection models, in alignment with findings from prior studies.

Table 4. Classification results for single speech subset

Classifier	Speech Subset	Recall	F1 score	Accuracy
SVM	Spectral	**0.73**	**0.75**	**0.74**
	Prosody	0.69	0.64	0.71
	Voice quality	0.61	0.63	0.63
	Formants	0.59	0.60	0.61
RF	Spectral	**0.77**	**0.72**	**0.75**
	Prosody	0.71	0.66	0.69
	Voice quality	0.63	0.64	0.60
	Formants	0.58	0.61	0.59
MLP	Spectral	**0.82**	**0.80**	**0.84**
	Prosody	0.80	0.78	0.81
	Voice quality	0.70	0.68	0.74
	Formants	0.69	0.75	0.71

4.2. Speech feature subset combination

In this study, we strategically combined subsets of speech features, each comprising 3–5 key features selected for their diagnostic potential. Spectral features, which emerged as the most predictive in initial single subset tests, were used as a foundational baseline. We then systematically tested combinations of this baseline with other feature subsets to determine which fusion would enhance depression classification performance. Employing SVM, RF, and MLP classifiers, and validating through 5-fold cross-validation, the dataset was divided into training, testing, and validation segments of 70%, 15%, 15%. The synthesis of results, detailed in Appendix 2, demonstrate the classifiers' performance across various feature combinations on the test dataset.

In the results for the combination of spectral and prosodic features, the SVM classifier achieved a recall of 0.81, an F1 score of 0.78, and an accuracy of 0.85. The RF classifier showed slightly lower results with a recall of 0.80, an F1 score of 0.82, and an accuracy of 0.79. The MLP classifier outperformed the other two with a recall of 0.85, an F1 score of 0.83, and an accuracy of 0.86. In the outcomes where spectral features are combined with voice quality, the performance slightly decreases. The SVM classifier achieved a recall of 0.70, an F1 score of 0.73, and an accuracy of

0.69. The RF classifier scores a recall of 0.73, an F1 score of 0.68, and an accuracy of 0.75. The MLP maintained a robust performance with a recall of 0.72, an F1 score of 0.77, and an accuracy of 0.80. The combination of spectral features with formants shows further variation in classifier performance. A recall in the SVM classifier is 0.72, with an F1 score of 0.66 and an accuracy of 0.70. The RF classifier scores a recall of 0.67, an F1 score of 0.65, and an accuracy of 0.73. The MLP classifier demonstrates a recall of 0.69, an F1 score of 0.71. and an accuracy of 0.74. Among the two-feature subset combinations that include the spectral feature subset as a baseline, the combination of spectral and prosodic features consistently shows the highest classification performance for all three classifiers.

When spectral features, prosodic features, and voice quality are combined, there is a slight improvement in performance compared to adding a single subset to the baseline(spectral features). The SVM classifier achieves a recall of 0.73, an F1 score of 0.77, and an accuracy of 0.76. RF shows a recall of 0.79, an F1 score of 0.75, and an accuracy of 0.74. The MLP presents a recall of 0.81, an F1 score of 0.74, and an accuracy of 0.79. For the combination of spectral, prosodic features, and formants, the SVM achieves a recall of 0.74, an F1 score of 0.73, and an accuracy of

0.77. The RF classifier scores a recall of 0.79, an F1 score of 0.77, and an accuracy of 0.81. The MLP provides a recall of 0.76, an F1 score of 0.80, and an accuracy of 0.82. It is noteworthy that the inclusion of formants in addition to spectral and prosodic features enhances the discriminatory power for detection depression compared to adding voice quality. Particularly, the RF classifier showed the highest accuracy when combining spectral, prosodic, and formant features.

For the combination of three subsets without prosodic features, the performance of SVM includes a recall of 0.71, an F1 score of 0.63, and an accuracy of 0.73. RF classifier scores a recall of 0.68, an F1 score of 0.61, and an accuracy of 0.64. The MLP classifier shows a recall of 0.77, an F1 score of 0.71, and an accuracy of 0.75, indicating that it may not be as potent as when prosodic features are included in the combination. Upon comparing the combinations of spectral, prosodic, and voice quality features with those of spectral, prosodic, and formant features, it was observed that the latter combination yielded higher accuracy. However, given the overall metrics, the classification performance of both sets of combinations was generally similar. Finally, in the comprehensive examination of all speech feature subsets combined, the MLP classifier stands out with the highest performance,

achieving a recall of 0.72, an F1 score of 0.78 and an accuracy of 0.80. The overall findings suggest that the combination of spectral and prosodic features emerged as the most optimal combination. To assess the model’s performance based on the optimal combination of speech features, ROC curves were generated. Figure 4 illustrates the ROC curve of the SVM classifier using the optimal combination of spectral and prosodic features. The corresponding Area Under the Curve(AUC) value was determined to be 0.81.

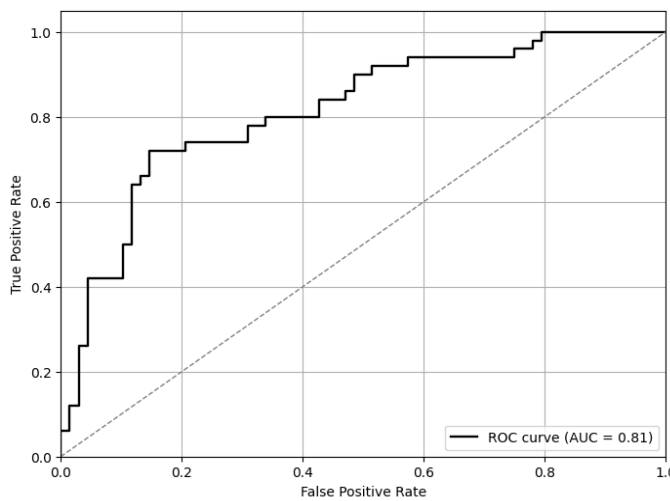

Figure 4. ROC curve of SVM classifier

4.3. BDI prediction

Based on the classification results from all possible combinations, we also developed a prediction model for the BDI-II scores based on SVM. This model leveraged the optimal combination of speech feature subsets, which includes both spectral and prosodic features. Figure 5 presents a scatter plot comparing the actual BDI scores to predicted scores specifically for the depressed group. The evaluation of model accuracy revealed a Mean Absolute Error (MAE) of 7.19 and an R-squared value of 0.42.

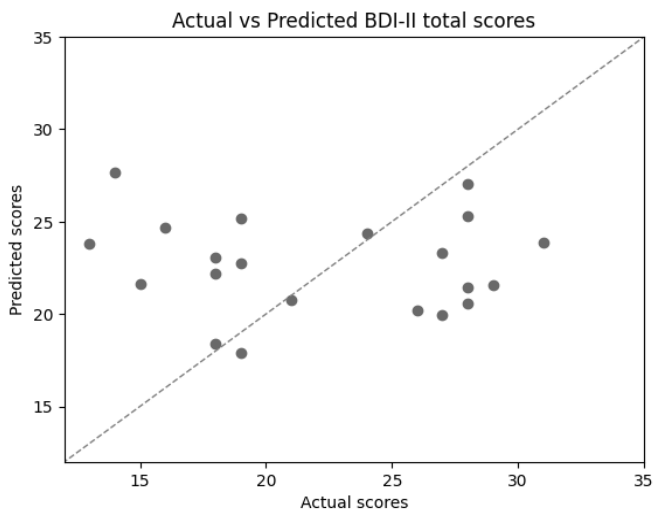

Figure 5. Scatter plot of actual and predicted BDI total score

4.4. BDI two-factor analysis

The study further investigated the correlations between BDI-II subscale scores and each of the selected speech features employed for classification. These subscale scores aligned with the two-factor model proposed by Whisman et al.(2000) and Al-Turkait et al.(2010), namely Cognitive-Affective and Somatic factors(shown in Figure 6), were validated through confirmatory factor analysis to assess the structure validity of BDI-II. By examining the correlations between individualized scores and selected speech features, the study aims to delineate a potential link between major clinical symptoms of depression and speech characteristics.

Table 5. Correlations between speech features and cognitive-affective factor scores

Subset (#number)	Selected features	r	*p*-value
Spectral(4)	mfcc_first_derivative_std_1	0.32	0.001
	spectral_spread_mean	0.29	0.004
	spectral_entropy_first_derivative_std	0.28	0.014
	spectral_slope_max	0.18	0.005
Prosody(3)	loudness_slidingwindow_max	0.22	0.007
	log_energy	0.23	0.046

	intensity_first_derivative_mean	0.15	0.033
Voice quality(5)	localabsoluteJitter	0.20	0.009
	apq11Shimmer	0.15	0.029
	localdbShimmer	0.08	0.047
	apq3Shimmer	0.13	0.578
	HNR	0.20	0.012
Formant(3)	F2_mean	0.13	0.002
	F1_first_derivative_std	0.09	0.564
	F2_max	−0.15	0.047

Initial steps involved categorizing the 21 BDI-II items into cognitive-affective or somatic factors, yielding total scores for each sub-factor across 142 participants. Table 6 and 7 shows the correlations between selected speech features and two-factor scores, presented with coefficient values and corresponding p-values. Figure 7 presents the scatter plot illustrating the distribution of correlations between two-factor scores and speech features. Results suggest that within the four subsets for final binary classification, correlations and significance of selected speech features were more pronounced in somatic factor scores than cognitive-affective factor scores. This observation highlights a potential association between somatic symptoms and specific speech features(Table 5), predominantly in spectral and prosodic

domains, surpassing the correlations observed with cognitive-affective factor scores(Table 6).

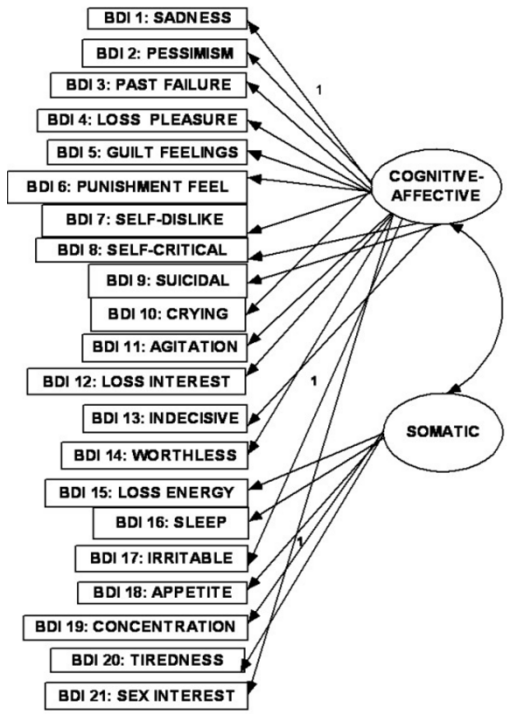

Figure 6. BDI-II two-factor model(Al-Turkait et al., 2010)

Our finding implies that there is a noteworthy correlation between somatic factor scores and specific vocal components, particularly within the spectral and prosodic domains. These vocal components, marked by their discriminative power observed in the previous section, motivate the question of a potential link between the physiological symptoms and certain characteristics embedded within vocal expressions. This suggests vocal patterns of

depression may potentially capture somatic symptoms in depression. Several notable features such as spectral entropy, loudness, and log energy measures stand out as relevant indicators of the complex interaction of emotional and physical dimensions in depressive voice. The results also suggest the potential utility of using speech features for capturing the somatic symptoms of depression such as fatigue and lack of energy.

Table 6. Correlations between speech features and somatic factor scores

Subset (#number)	Selected features	r	*p*-value
Spectral(4)	mfcc_first_derivative_std_1	0.37	0.005
	spectral_spread_mean	0.35	0.000
	spectral_entropy_first_derivative_std	0.37	0.002
	spectral_slope_max	0.28	0.035
Prosody(3)	loudness_slidingwindow_max	0.28	0.011
	log_energy	0.27	0.005
	intensity_first_derivative_mean	0.19	0.001
Voice quality(5)	localabsoluteJitter	0.11	0.037
	apq11Shimmer	0.13	0.002
	localdbShimmer	0.15	0.001

	apq3Shimmer	0.08	0.053
	HNR	0.02	0.461
Formant(3)	F2_mean	−0.03	0.216
	F1_first_derivative_std	0.12	0.051
	F2_max	−0.09	0.613

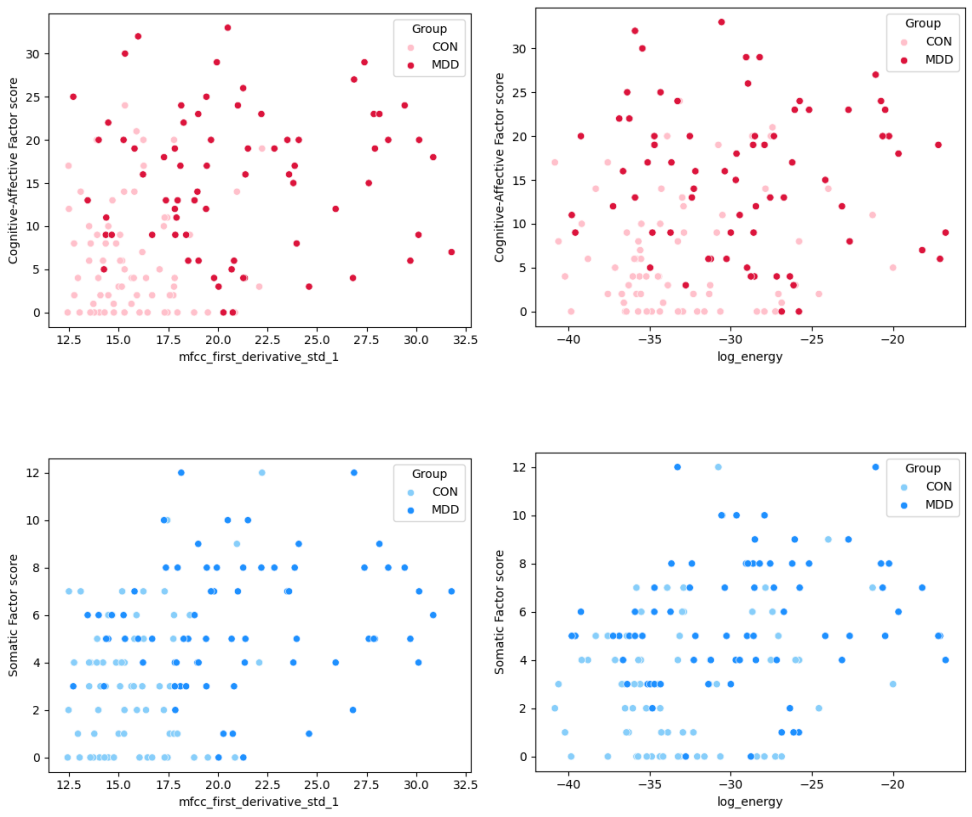

Figure 7. Scatter plot of correlation between two-factor scores and speech features

Chapter 5. General Discussion

In the initial phase of our assessment, we examined the classification results of each single speech subset, and the spectral features were identified as the most discriminative features for detecting MDD among the four speech dimensions. This thus confirmed our first hypothesis, which posited that spectral features would outperform other speech subsets. As shown in Table 4, the spectral features consistently yielded better classification results than all other speech subsets for every classifier. This suggests that the spectral features contain relevant information for emotional speech, as supported by previous findings (Lee et al., 2021).

Based on the results shown in Table 4, we included spectral features as a baseline in all feature combinations, resulting in a classification experiment for detecting depressive speech across different speech subsets, presented in Appendix 2. Among all possible combinations of speech subsets, the combination of spectral and prosodic features demonstrated the most superior performance across nearly all evaluation metrics. This performance was either on par with or better than the results achieved by using all four speech subsets. In terms of classifiers, MLP consistently outperformed the other two classifiers. The cross-validated F1

scores(shown in Table 7) highlight the overall excellence of the chosen combination in detecting depressive speech.

A notable finding is that when prosodic features were added to combinations where the baseline feature was enhanced with only one additional speech subset, it resulted in a substantial improvement in classification performance. This suggests that prosodic features, like spectral features, may capture essential characteristics of depressive speech. As previous research has also indicated, prosodic features play a significant role in assessing psychomotor disturbances in depressive speech (Cummins et al., 2015; Yang et al., 2013). In depressive speech, we typically observe reduced energy variability, pitch variability, and speech rate, which are central prosodic indicators. This underscores the importance of studying depressive prosody alongside spectral features.

While voice quality and formants did not contribute to improving classification performance as much as prosodic features, formant features had a greater impact on performance enhancement compared to voice quality. This may indicate that the influence of depressive speech characteristics commonly observed in vowel formants played a significant role. Another remarkable observation is that employing only two influential speech subsets for

classification yielded higher performance compared to using all speech features, implying that increasing the number of features may not necessarily improve the classification performance.

Table 7. 5-fold cross-validated F1 score on the optimal combination validation set

Classifier	#Fold					5-fold Mean
	1	2	3	4	5	
SVM	0.74	0.74	0.76	0.85	0.81	**0.78**
RF	0.79	0.84	0.80	0.83	0.84	**0.82**
MLP	0.85	0.81	0.81	0.81	0.87	**0.83**

In light of our further experiments, it becomes more apparent that a potential correlation exists between somatic factor scores and specific vocal components, particularly within the spectral and prosodic dimensions. These vocal components, distinguished by their discriminatory capacity as observed, suggest a further exploration into the potential interconnection between distinctive vocal characteristics and somatic symptoms. Somatic symptoms often have the potential to affect behavioral patterns. Such changes, then, may also impact vocal characteristics. Hence, the correlation between somatic symptoms and speech features may reflect an interaction between behavior and voice. This

discussion also leads to another follow-up thought. When somatic symptoms of depression manifest, there can be alterations in emotional expression. These changes may be reflected in vocal characteristics, particularly in prosodic features. In summary, the relevant correlation between the somatic factor and speech features suggests the possibility that physical symptoms or behavioral changes associated with depression are reflected in voice.

In this study, the primary goal was to enhance our understanding of the clinical explanatory power of depression-related speech features, particularly in the context of detecting clinical symptoms of depressive disorders and emphasize the clinical interpretability of depressive speech by leveraging the synergy of specific speech feature combinations. We used the read speech task; however, it may have inherent limitations in capturing the emotional dynamics which can be easily observed in spontaneous speech. Furthermore, it is important to note that data scarcity may introduce performance biases among various speech subsets. Future work should explore more effective elicitation methods of depressive vocal markers from spontaneous speech through a larger dataset. Another limitation of this study arises from the benchmarking of most classifiers against those previously utilized in clinical speech research. This introduces challenges in

mitigating potential performance differences and distortions, given that the classifiers used in prior studies were trained on specific datasets in certain context. Furthermore, as is often the case in the clinical speech domain, it is unavoidable to encounter variability due to data scarcity. Some classifiers may have been optimized to train on large dataset, so achieving a more effective performance comparison may require a substantially large sample size to ensure robustness. Given these limitations, future work should prioritize exploring novel speech tasks that can effectively capture emotional dynamics of depressive speech. Finally, there is a pressing need to investigate the generalizability of vocal markers and their elicitation methods such as recording systems across larger dataset to enhance the reliability and applicability of the findings.

Chapter 6. Conclusion

This study aims to explore various speech features with varying classification capabilities to effectively distinguish clinical symptoms of depression. We have presented our work with the goal of providing an objective diagnostic tool to support clinicians in their diagnosis of depression. The results have validated our hypotheses by examining and comparing participants' acoustic features through a read speech task. We found that the combined speech subsets of spectral and prosodic features outperformed other speech subset combinations as well as a single spectral subset. These findings may indicate that depressive speech patterns are more comprehensively explained by the combined influence of spectral and prosodic features within the overall speech pattern. It further underscores the need for future analyses to delve into the psychological and neural underpinnings that contribute to the discriminative power of each speech subset. Our findings also highlight the significance of feature-based analysis, particularly in the domain of clinical speech research. In conclusion, the development of a highly reliable and cost-effective markers for automated assessment is pivotal, with a focus on minimizing computational demands. This requires meticulous training on

specific speech patterns closely associated with clinical symptoms. Such efforts hold utmost importance in real-world clinical applications, where the demand for accurate and efficient diagnostic tools is crucial.

References

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech communication*, *71*, 10-49.

Kim, Y., Kwon, H. Y., Lee, S., & Kim, C. B. (2022). Depression During COVID-19 Quarantine in South Korea: A Propensity Score-Matched Analysis. *Frontiers in public health*, *9*, 743625.

Robin, J., Harrison, J. E., Kaufman, L. D., Rudzicz, F., Simpson, W., & Yancheva, M. (2020). Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. *Digital biomarkers*, *4*(3), 99–108.

Silva, W. J., Lopes, L., Galdino, M. K. C., & Almeida, A. A. (2021). Voice Acoustic Parameters as Predictors of Depression. *Journal of voice : official journal of the Voice Foundation*, S0892-1997(21)00205-8.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, *47*(7), 829-837.

Califf, R. M. (2018). Biomarker definitions and their applications. *Experimental Biology and Medicine*, *243*(3), 213-221.

Stasak, B., Epps, J., Schatten, H. T., Miller, I. W., Provost, E. M., & Armey, M. F. (2021). Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Communication*, *132*, 10-20.

Tasnim, M., & Novikova, J. (2022, December). Cost-effective

Models for Detecting Depression from Speech. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1687-1694). IEEE.

Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.

Lee, J. H., Park, S. K., Ryoo, J. H., Oh, C. M., Choi, J. M., McIntyre, R. S., Mansur, R. B., Kim, H., Hales, S., & Jung, J. Y. (2017). U-shaped relationship between depression and body mass index in the Korean adults. *European psychiatry : the journal of the Association of European Psychiatrists*, *45*, 72–80.

Nemade, R., Reiss, N., & Dombeck, M. (2019). Biology of depression—neurotransmitters.

Bembnowska, M., & Jośko-Ochojska, J. (2015). What causes depression in adults?. *Polish Journal of Public Health*, *125*(2).

Davey, C. G., & McGorry, P. D. (2019). Early intervention for depression in young people: a blind spot in mental health care. *The Lancet Psychiatry*, *6*(3), 267-272.

Nevler, N., Ash, S., Irwin, D. J., Liberman, M., & Grossman, M. (2019). Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology*, *6*(1), 4-14.

Ash, S., Evans, E., O'Shea, J., Powers, J., Boller, A., Weinberg, D., Haley, J., McMillan, C., Irwin, D. J., Rascovsky, K., & Grossman, M. (2013). Differentiating primary progressive aphasias in a brief sample of connected speech. *Neurology*, *81*(4), 329–336.

Wang, J., Zhang, L., Liu, T., Pan, W., Hu, B., & Zhu, T. (2019).

Acoustic differences between healthy and depressed people: a cross-situation study. *BMC psychiatry*, *19*, 1-12.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2014, November). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (pp. 65-72).

Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., & Morency, L. P. (2013, April). Automatic behavior descriptors for psychological disorder analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1-8). IEEE.

Yang, Y., Fairbairn, C., & Cohn, J. F. (2012). Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, *4*(2), 142-150.

Afshan, A., Guo, J., Park, S. J., Ravi, V., Flint, J., & Alwan, A. (2018). Effectiveness of voice quality features in detecting depression. *Interspeech 2018*.

Cummins, N., Vlasenko, B., Sagha, H., & Schuller, B. (2017). Enhancing speech-based depression detection through gender dependent vowel-level formant features. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16* (pp. 209-214).

Brietzke, E., Hawken, E. R., Idzikowski, M., Pong, J., Kennedy, S. H., & Soares, C. N. (2019). Integrating digital phenotyping in clinical characterization of individuals with mood disorders. *Neuroscience & Biobehavioral Reviews*, *104*, 223-230.

Kim, H. (2012). Neurologic speech-language disorders.

Lenain, R., Weston, J., Shivkumar, A., & Fristed, E. (2020). Surfboard: Audio feature extraction for modern machine learning. *arXiv preprint arXiv:2005.08848*.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the munich open-source multimedia feature extractor. *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, 835–838.

Boersma, P., & Van Heuven, V. (2001). *Speak and unSpeak with PRAAT RAAT*.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, *71*, 103107.

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, *9*, 1112-1122.

Theerthagiri, P., & Vidya, J. (2022). Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. *Expert Systems*, *39*(9), e13064.

Zhang, B., Zhang, Y., & Jiang, X. (2022). Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm. *Scientific Reports*, *12*(1), 9244.

Balagopalan, A., & Novikova, J. (2021). Comparing acoustic-based approaches for Alzheimer's disease detection. *arXiv preprint arXiv:2106.01555*.

Sun, H., Wang, H., Liu, J., Chen, Y. W., & Lin, L. (2022, October). CubeMLP: An MLP-based model for multimodal sentiment

analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 3722–3729).

Lee, S., Suh, S. W., Kim, T., Kim, K., Lee, K. H., Lee, J. R., Han, G., Hong, J. W., Han, J. W., Lee, K., & Kim, K. W. (2021). Screening major depressive disorder using vocal acoustic features in the elderly by sex. *Journal of Affective Disorders*, *291*, 15–23.

Whisman, M. A., Perez, J. E., & Ramel, W. (2000). Factor structure of the Beck Depression Inventory—Second Edition (BDI-ii) in a student sample. *Journal of clinical psychology*, *56*(4), 545–551.

Al-Turkait, F. A., & Ohaeri, J. U. (2010). Dimensional and hierarchical models of depression using the Beck Depression Inventory-II in an Arab college student sample. *BMC psychiatry*, *10*(1), 1-14.

Appendix

Appendix 1: Pre-defined paragraphs(Kim, 2005)

During the speech collection, each paragraph was presented using visual materials displayed on a 24-inch monitor, with a font size of 18 points, zero letter spacing, 160% line spacing, and justified alignment.

<가을>

우리나라의 가을은 참으로 아름답다. 무엇보다도 산에 오를 땐 더욱더 그 빼어난 아름다움이 느껴진다. 쓰다듬어진 듯한 완만함과 깎아 놓은 듯한 뾰족함이 어우러진 산등성이를 오르다 보면, 절로 감탄을 금할 수가 없게 된다. 붉은색, 푸른색, 노란색 등의 여러 가지 색깔들이 어우러져, 타는 듯한 감동을 주며 나아가 신비롭기까지 하다. 숲 속에 누워서 하늘을 바라보라. 쌍쌍이 짝지어져 있는 듯한 흰 구름, 높고 파란 하늘을 쳐다보고 있노라면 과연 예부터 가을을 천고마비의 계절이라 일컫는 이유를 알게 될 것만 같다. 가을에는 또한 오곡백과 등 먹거리가 풍성하기 때문에 결실의 계절이라고도 한다. 햅쌀, 밤, 호두뿐만 아니라 대추, 여러 가지 떡, 크고 작은 과일들을 맛볼 수 있는데, 가을의 대표적인 명절인 추석에 우리는 이것들을 쌓아 놓고 조상님들께 차례를 지내기도 한다. 또한 가을은 독서의 계절이라고도 하여 책을 읽으며 시시때때로 명상에 잠기기도 하는데, 독서는 우리에게 마음을

살찌우고 아름답게 하는 힘을 주기 때문이다.

<여행>

일상이 문득 너무 무덤덤할 땐, 여행 같은 특효약이 또 있을까. 갑갑하고 빡빡한 생활의 흔적을 잊고 떠나자. 몸도 마음도, 자신감 충만한 느낌이 가득해질 것이다. 지도 따라 자전거로, 쌍쌍이 전국일주를 해보자. 캔 커피를 담뿍 챙겨, 자동차로 신나게 달려보자. 교외 고속도로를 쭉 달리면서, 샘솟는 해방감 만끽해보자. 그러나 참여행의 백미는 맛난 계란을 야금야금 까먹는, 멋이 좋은 기차 여행이 아닐까.

<바람과 햇님>

바람과 햇님이 서로의 힘이 더 세다고 다투고 있을 때, 한 나그네가 여유롭게 따뜻한 외투를 입고 걸어왔습니다. 그들은 누구든지 나그네의 외투를 먼저 벗겨야 힘이 더 세다고 하기로 결정했습니다. 북풍은 위에서 힘껏 불었으나 불면 불수록 나그네는 콧물을 흘리며 외투를 단단히 여몄습니다. 이 때 햇님이 계획대로 워낙 뜨거운 햇빛을 가만히 내려 쬐니 나그네는 외투를 얼른 벗었습니다. 이리하여 북풍은 햇님이 힘이 더 세다고 인정하지 않을 수 없었지요.

Appendix 2: Synthetic results of experiments

1: spectral features, 2: prosodic features, 3: voice quality, 4: formants

Classifier	SVM			RF			MLP		
# Speech subset	Recall	F1 score	Accuracy	Recall	F1 score	Accuracy	Recall	F1 score	Accuracy
(1,2)	**0.81**	**0.78**	**0.85**	**0.80**	**0.82**	0.79	**0.85**	**0.83**	**0.86**
(1,3)	0.70	0.73	0.69	0.73	0.68	0.75	0.72	0.77	0.80
(1,4)	0.72	0.66	0.70	0.67	0.65	0.73	0.69	0.71	0.74
(1,2,3)	0.73	0.77	0.76	0.79	0.75	0.74	0.81	0.74	0.79
(1,2,4)	0.74	0.73	0.77	0.79	0.77	**0.81**	0.76	0.80	0.82
(1,3,4)	0.71	0.63	0.73	0.68	0.61	0.64	0.77	0.71	0.75
(1,2,3,4)	0.70	0.66	0.67	0.75	0.71	0.71	0.72	0.78	0.80

국문 초록

최적의 음성특징 조합을 활용한 주요우울장애 자동분류

서울대학교 인문대학
협동과정 인지과학 전공

함현선

본 논문은 주요우울장애를 진단하기 위한 최적의 음성특징 조합을 찾는 것을 목적으로 한다. 다양한 선행연구에서 주요우울장애 자동진단을 위한 음성 바이오마커의 유용성을 이미 입증하였으나, 주요우울장애의 임상적 증상 및 아형을 설명하는 데 가장 관련성이 높은 음성특징이 무엇인지에 대한 공통된 합의는 여전히 부족하다. 본 연구에서는 기존 선행연구에서 사용된 다양한 음성특징 하위집합을 사용하여 각 음성특징의 분류 성능을 검토하고 음성기반 진단의 정확도를 향상시킬 수 있도록 하는 최적의 음성특징 조합을 제안한다.

본 논문에서는 개별 음성지표 및 음성 하위집합들의 분류성능 향상 기여도를 검증하여 최적의 음성특징 조합을 제안한다. 제안된 최적의 음성특징 조합은 예측모형을 통해 기존 우울증 진단을 위한 신경심리평

가도구로 사용되는 벡 우울척도(BDI-II) 점수에 대한 예측능력을 검증하는 데 사용된다. 본 연구에서는 주요우울장애 진단을 받은 72명의 환자와 70명의 대조군을 대상으로 표준문단읽기를 실시하였으며, 총 세 문단을 발화한 모든 개별 오디오 파일로부터 210개의 음성특징이 추출되었다. 추출된 음성특징은 스펙트럼 특징, 운율 특징, 음질 특징, 그리고 포먼트 특징의 네 가지 음성 하위집합으로 분류되었다.

특징 선택은 Recursive Feature Elimination(RFE) 알고리즘을 통해 최적화되었으며, 특징선택의 기준은 Extreme Gradient Boositng(XGboost)을 사용하여 계산된 중요도 점수에 기반하였다. 선택된 개별 하위집합 내 음성특징들은 Support Vector Machine(SVM), Random Forest(RF), Multi Layer Perceptron(MLP) 분류기의 입력값으로 사용되었고, 개별 하위집합의 분류성능을 1차로 확인하여 베이스라인을 설정하였고, 베이스라인을 포함한 나머지 하위집합을 모두 활용하여 가능한 모든 조합의 분류성능을 2차로 검증하였다.

개별 하위집합 수준의 분석에서 스펙트럼 특징은 우울군과 정상군을 구분하는 데 가장 뛰어난 성능을 보이며 첫 번째 연구가설과 부합하였다. 음성 하위집합의 조합 중에서는 스펙트럼 특징과 운율 특징의 조합이 MLP 분류기에서 F1 score 0.83으로 다른 모든 조합보다 뛰어난 분류성능을 보였다. 이 결과를 바탕으로 스펙트럼 특징과 운율 특징 조합을 주요우울장애 자동분류를 위한 최적의 조합으로 제안하였고, 해당 조합의 진단적 유용성을 평가하기 위해 BDI-II 예측모델을 구축하였으며, 평균 절대 오차(MAE)는 7.19로 나타났다. 또한 본 연구에서는

최적의 조합으로 제안된 음성 특징과 주요우울장애의 아형 및 임상적 증상의 관계를 보다 탐색적으로 규명하기 위해 다수의 BDI-II 구성 타당도 연구에서 입증된 인지-정서요인(Cognitive-Affective)과 신체요인(Somatic) 구조에 근거하여 모든 참가자들의 인지-정서요인 점수와 신체요인 점수를 산출하였고, 요인별 점수와 음성특징이 어떤 상관을 보이는지 분석하였다. 분석 결과, 최적의 조합으로 제안되었던 스펙트럼 특징과 운율 특징 중 일부 음성특징이 인지-정서요인 점수보다 신체요인 점수와 더 높은 상관을 보였고, 이를 통해 음성이 주요우울장애에서 드러나는 신체적 상태 및 패턴을 반영할 수 있을 것으로 해석되었다.

본 연구에서는 우울 음성을 식별하는 데 핵심적인 스펙트럼 특징과 운율 특징 조합의 시너지 효과를 확인할 수 있었으며, 두 음성특징이 주요우울장애 진단에 중요한 음성정보임을 확인할 수 있었다. 향후 연구에서는 이 두 음성특징의 임상적 타당성을 확보할 수 있도록 하는 신경심리학적 후속연구가 뒷받침되어야 하며, 언어학적 이해를 토대로 한 도메인 기반 접근 방식을 통해 음성기반 자동진단모델의 정확도를 높이려는 시도가 지속되어야 할 것이다.

주요어 : 주요우울장애, 최적의 음성특징 조합, 음성 하위집합, 자동분류, 음성 바이오마커

학번 : 2022-24910