공학박사 학위논문

주요 우울 장애의 음성 기반 분석: 연속적인 발화의 음향적 변화를 중심으로

Speech-Based Analysis of Major Depressive Disorder: Focusing on Acoustic Variations in Consecutive Utterances

2023 년 2 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

이 수 빈

주요 우울 장애의 음성 기반 분석: 연속적인 발화의 음향적 변화를 중심으로

Speech-Based Analysis of Major Depressive Disorder: Focusing on Acoustic Variations in Consecutive Utterances

지도교수 이 교 구

이 논문을 공학박사 학위논문으로 제출함

2023 년 1 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

이 수 빈

이수빈의 공학박사 학위논문을 인준함

2023 년 1 월

위 원 장	이 원 종	(인)
부위원장	이 교 구	(인)
위 원	곽 노 준	(인)
위 원	서 봉 원	(인)
위 원	남 주 한	(인)

Abstract

Major depressive disorder (commonly referred to as depression) is a common disorder that affects 3.8% of the world's population. Depression stems from various causes, such as genetics, aging, social factors, and abnormalities in the neurotransmitter system; thus, early detection and monitoring are essential. The human voice is considered a representative biomarker for observing depression; accordingly, several studies have developed an automatic depression diagnosis system based on speech. However, constructing a speech corpus is a challenge, studies focus on adults under 60 years of age, and there are insufficient medical hypotheses based on the clinical findings of psychiatrists, limiting the evolution of the medical diagnostic tool. Moreover, the effect of taking antipsychotic drugs on speech characteristics during the treatment phase is overlooked.

Thus, this thesis studies a speech-based automatic depression diagnosis system at the semantic level (sentence). First, to analyze depression among the elderly whose emotional changes do not adequately reflect speech characteristics, it developed the mood-induced sentence to build the elderly depression speech corpus and designed an automatic depression diagnosis system for the elderly. Second, it constructed an extrapyramidal symptom speech corpus to investigate the extrapyramidal symptoms, a typical side effect that can appear from an antipsychotic drug overdose. Accordingly, there is a strong correlation between the antipsychotic dose and speech characteristics. The study paved the way for a comprehensive examination of the automatic diagnosis system for depression.

Keywords: major depressive disorder, automatic depression diagnosis system, depression speech corpus, mood-inducing sentence, speech characteristics, extrapyramidal symptoms

Student Number: 2013-22411

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Major depressive disorder (MDD) (commonly called depression) is a serious disorder that depletes motivation, induces various cognitive and psychogenic symptoms, and impairs daily functioning. Depression affects all sorts of people and is prevalent worldwide. It is estimated that 3.8% (280 million) of the world's population suffers from depression. It affects 5.0% of all adults and 5.7% of people over 60 years of age; indeed, it can occur at any age [7]. Additionally, per a survey conducted in the United States, 8.1% of American adults over the age of 20 experience depression, and women (10.4%) experience depression nearly twice as often as men (5.5%) [8]. Although depression is highly prevalent, it is often underdiagnosed [9], inducing delayed treatment with serious adverse outcomes, including cognitive impairments, worsening physical illnesses, and suicide [10]. The main symptoms of depression are diverse and mainly include a depressed mood, decreased motivation and interest, decreased appetite, unstable sleep patterns, and anxiety.

Engineering research seeks to manage and monitor such clinical symptoms. Several biomarkers (e.g., image, video, speech, and text) for diagnosing and monitoring depression have long been studied. Studies based on visual cues (images, videos) have mainly employed facial analysis to analyze interview video data of subjects [11, 12]. Text information is also a good biomarker that can reflect the human cognitive level. Thus, many studies seek to detect depression in social media messages or texts expressing one's thoughts [13, 14]. Given that people with depression are characterized by disproportionate alpha oscillations, electroencephalogram (EEG) data have also been consistently suggested and validated as biomarkers for detecting depression [15]. Moreover, among the most commonly proposed general biomarkers for depression diagnosis research is speech data [16], which is the main subject of this paper. Regarding visual cues, databases may reflect other emotions and produce misleading results, and text information is not suitable for subjects not versed in expressing emotions in text, hiding their true emotion. Regarding EEG data, data accessibility is low, and its real-world application is limited. However, audio is a widely used powerful biomarker because it is easy to acquire audio data. Further, it can capture the natural emotions of a subject through voice characteristics during the speech process.

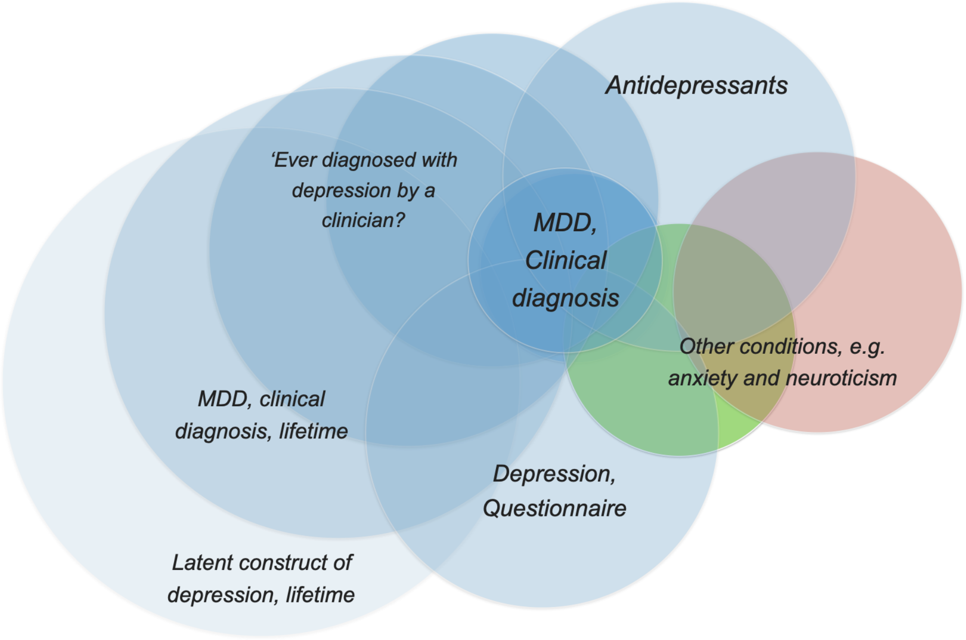
Even so, speech data has many limitations. It is challenging to build a generalized database that reflects many factors (linguistic, demographic) that affect speech characteristics, and it is challenging to obtain relevant real-world data. This thesis aims to bridge the gap in the depression analysis literature using speech data by introducing a new approach to the voice-based depression analysis method.

1.1 Research Motivations

The study established several research motivations based on past speech-based depression detection studies to set the goals of this thesis. Thus, this section explains the thesis motivation and novelty relative to previous studies.

1.1.1 Bridging the Gap Between Clinical View and Engineering

Fig. 1.1 A major point of criticism of combining different depression diagnosis phenotypes [1]


Antidepressants
'Ever diagnosed with depression by a clinician?
MDD, Clinical diagnosis
Other conditions, e.g. anxiety and neuroticism
MDD, clinical diagnosis, lifetime
Depression, Questionnaire
Latent construct of depression, lifetime

First, this thesis aims to bridge engineering and MDD. Most previous studies on detecting depression have modeled a simple relationship between speech data and the diagnostic depression scale.

From Figure 1.1, the phenotype of depression is expressed vaguely, and its range is not precisely defined. MDD is diagnosed when depressed mood, loss of interest, changes in appetite and weight, sleep disturbance, sense of worthlessness, fatigue, and suicidal thoughts are present for at least two weeks or more for most of the day. It is diagnosed by comprehensively considering multiple interviews with clinicians and the subject's depressive scale. Thus, clinical opinions are crucial to the analysis of MDD, and research beyond simple correlation analysis of the depression scale is required. This thesis conducts a depression analysis study through clinical observation.

1.1.2 Limitations of Conventional Depressed Speech Corpora

For speech-based depression diagnosis analysis, building a depression speech corpus is a vital step. Speech corpora for various research attempts have been proposed, and studies have employed popular databases such as Daic-Woz [17] or AVid dataset [18]. However, such databases have a fixed population age group of people in their 20s and 30s, criteria for determining depression are limited to question-based evaluation tools, and the databases are not evaluated by psychiatrists. However, depression occurs more frequently among the elderly or people suffering from certain diseases. Moreover, given that it is a mood disorder that must be diagnosed and managed clinically, speech corpus is necessary.

1.1.3 Lack of Studies on Depression Among the Elderly

In old age, people experience severe psychological stress from deteriorating physical function, various degenerative diseases, death of close relatives or

Table 1.1 Different depressive symptoms of adult and elderly [3, 4]

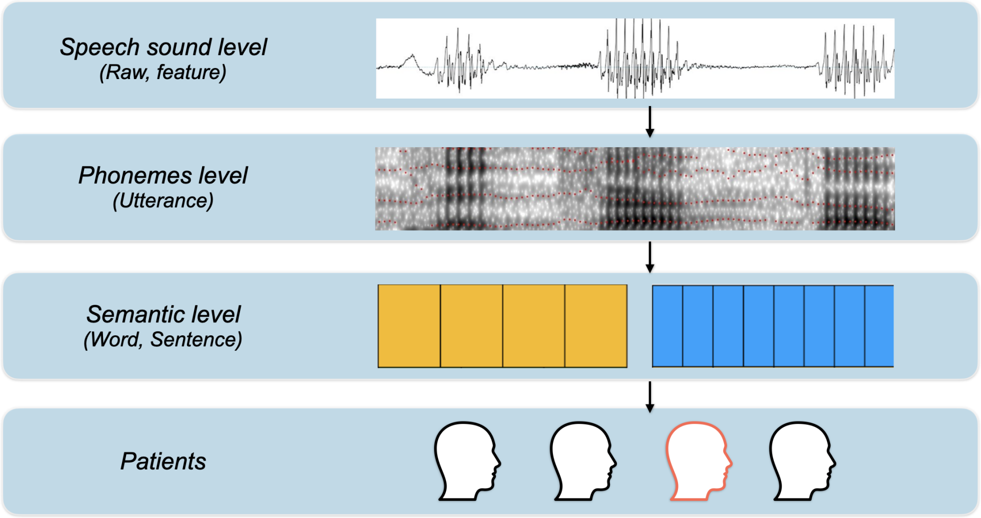
Symptom Domain	Adult (< 60 yo)	Elderly (> 60 yo)
Mood	Depressed, Anhedonic, Suicidal thoughts	Weary, Hopeless, Angry, Anxious, Thoughts of death
Somatic	Changes of sleep, appetite, psychomotor , Increased pain	Pain, Somatic symptoms with effects of medications
Cognitive	Decrease of concentration, Indecisiveness	Decrease of selective attention, working memory/retrieval, new learning, processing speed, executive function

spouses, economic loss because of retirement from work or social life, alienation from family and society, and loss of daily-life roles. They can occur at once or sequentially and are not easily experienced in other ages. Depression is among the most notable mental disorders, along with dementia [19]. Thus, given that the social and physical causes of depression in old age differ from those of other age groups, the symptoms of depression often also differ. From Table 1.1, the elderly suffering from depression exhibit different mood, somatic, and cognitive patterns from adults. In adults under 60 years of age, suicidal thoughts are dominant, whereas in elderly people over 60 years of age, fear of death prevails. A decline in cognitive ability from a decline in overall physical ability is also prevalent. Although the change in speech characteristics is expected given the change in symptoms per age group, previous studies mainly focus on the age group in their 20s and 30s. This thesis examines the characteristics of depression per age by modeling depression in the elderly.

1.1.4 Depression Analysis on Semantic Level

The setting of the observation unit is vital to determine the presence of depression in a subject through the voice-based automatic depression diagnosis system. In particular, to determine the presence of depression in a patient, the subject must be the final judgment unit, going through many intermediate steps from the raw audio signal to the mid-level feature expression or pronunciation level analysis. Prior studies discover new mid-level features and conduct many methods of analyzing features related to pronunciation such as formants. However, analysis at the semantic level has been very limited. Most studies on the semantic level analysis used in text-based automatic depression diagnosis systems determine the degree of depression of depressed subjects through the emotional scale of words. However, this approach is based on text information that hinges on free will, which is difficult to apply to a speech-based system. Of course, it is possible to compose natural data through a free format interview, but for semantic level analyses, it is common to convert what the speaker said into text information through speech recognition and analyze it. This study proposes a semantic-level analysis of speech features by aggregating them to a higher semantic level (word, sentence) beyond the conventional mid-level feature representation method.

Fig. 1.2 Analysis unit of automatic depression diagnosis system

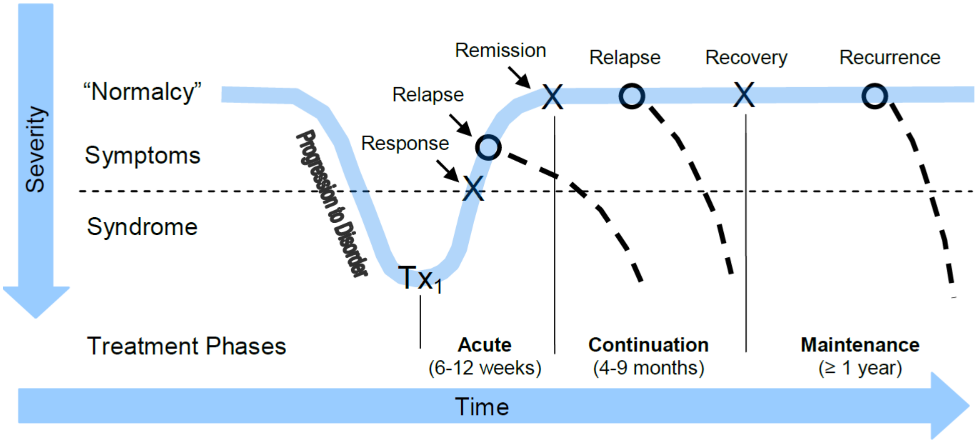


1.1.5 How Antipsychotic Drug Affects the Human Voice?

Previous studies of automatic diagnostic systems for MDD focus on monitoring after the onset of depression. However, the status of MDD is constantly changing from its initial detection through treatment and recovery. There are diverse treatments for MDD, including drug therapy using antidepressants, psychotherapy, phototherapy, electroconvulsive therapy, and transcranial magnetic stimulation. Of these treatments, drug therapy is the most widely used. Successful treatment is possible in 80-90% of depression, and people can return to their former life. Early detection and active treatment of depression can effectively enhance the therapeutic effect and prevent a recurrence.

From Fig. 1.3, the severity of depression continuously changes through the sub-processes (acute, continuation, maintenance) of MDD treatment. One clinical question emerges:

Fig. 1.3 Phases of treatment for major depressive disorder [2]



Q: Do changes in severity in the course of treatment for MDD induce changes in speech characteristics?

Accordingly, this thesis focuses on analyzing extrapyramidal symptoms (EPS), which are among the symptoms that may occur when taking antipsychotic drugs. EPS stems from the blockade of dopamine D2 receptors in the substantia nigra in the dopamine pathway, resulting in motor function abnormalities, drug-induced parkinsonism, acute dystonia, Akathisia, and delayed tardive dyskinesia. Hypothetically, these motor nervous system abnormalities also affect the vocalization mechanism comprising respiration, phonation, articulation, and resonance.

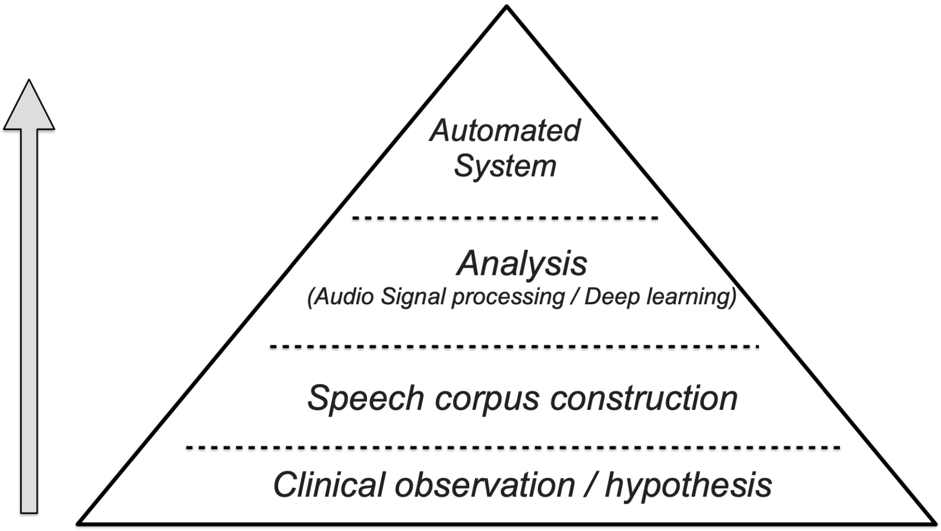
This thesis aims to probe the clinical assumption that EPS, a side effect of antipsychotic drugs, can also have a significant effect on speech production.

1.2 Thesis objectives

Given the motivations in the previous section, this thesis has several research objectives that are indispensable for the diagnosis of MDD. The interesting objectives below can help provide a new perspective to study MDD while bridging the gaps of existing studies.

From Figure 3, all studies in this thesis began were organized as follows: a clinical hypothesis, constructing reliable depressive speech corpora, and finally constructing an automatic depression diagnosis system.

Fig. 1.4 Research methodology of this thesis


Automated System
Analysis
(Audio Signal processing / Deep learning)
Speech corpus construction
Clinical observation / hypothesis

(1) Constructing a depression speech corpus through a new collection method This thesis proposes a new methodology that is different from the conventional depression speech corpus construction technique and develops easy-to-read sentences used for diagnosing depression. Hence, the thesis constructs a novel elderly depressive speech corpus.

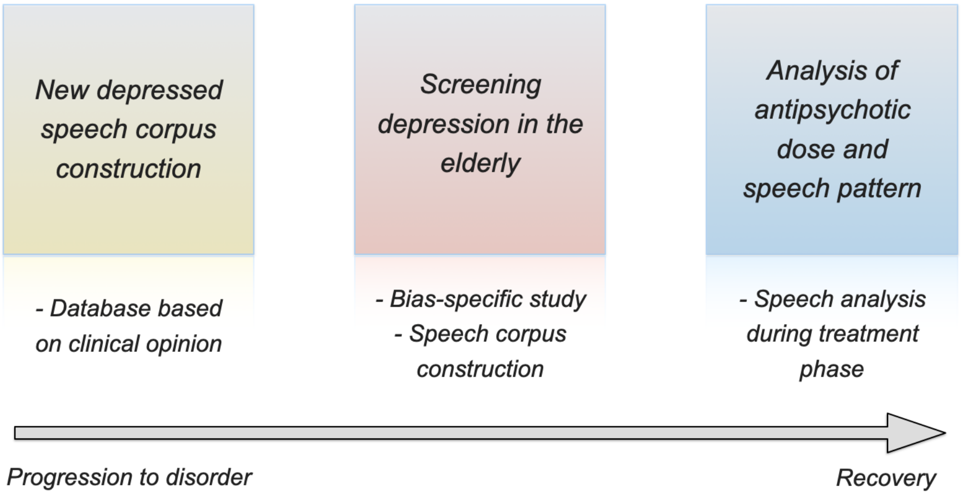
(2) Suggest sentence-level analysis of MDD based on clinical hypothesis Based on the sentence set for diagnosing depression and the constructed speech corpus developed, this thesis analyzes the diagnosis of depression at the semantic level. Depression analysis at the sentence level using speech characteristics is based on the clinical assumption that when a sentence containing a specific emotional word is read, the speaker's emotion can be induced into a specific state. Moreover, when the speech corpus is interview-type data, the answer to the question can be at the semantic level, and an answer unit is an observation unit for specifying the speaker's emotional state.

(3) Investigate the correlation of antipsychotic dose and speech characteristics This thesis investigates changes in speech characteristics that may occur during the treatment of MDD, thereby pioneering a new field of diagnosis and analysis of MDD. The study constructs an extrapyramidal speech corpus containing EPS (EPS corpus) and develops a corresponding sentence set. The construction methodology of the sentence set is the same as the clinical assumptions for the previous goal.

1.3 Outline of the thesis

The main contribution of the thesis is to define various tasks that must be complemented to make a diagnostic tool for MDD, as illustrated in Fig. 1.5. They are covered in chapters 3, 4, and 5, respectively, and their detailed contributions are as follows.

Fig. 1.5 Overview of the thesis


New depressed speech corpus construction
- Database based on clinical opinion
Screening depression in the elderly
- Bias-specific study
- Speech corpus construction
Analysis of antipsychotic dose and speech pattern
- Speech analysis during treatment phase
Progression to disorder
Recovery

Chapter 3 This chapter notes the limitations of conventional depressed speech corpora and introduces the elements to be observed in the speech corpus construction process. Thus, it develops sentence sets with two purposes and introduces the construction process in-depth.

Chapter 4 This chapter introduces the construction of the Korean language-based elderly depression speech corpus based on the mood-inducing sentences (MIS) designed in the previous chapter. The study proposes an experiment designed through a clinical assumption, based on which the experimental process of VoiSAD, an automatic elderly depression diagnosis system, is introduced, and a clinical disruption is conducted.

Chapter 5 This chapter conducts the correlation analysis between speech characteristics and the dose of antipsychotic drugs. It constructs a Korean-based

EPS speech corpus through sentence development to investigate the pattern of speech characteristics of EPS-related voice change. Through this, speech patterns of EPS and non-EPS groups were investigated.

The chapter briefly lists the limitations of the previous MDD diagnosis studies and explains the motivation for this thesis, highlighting the uniqueness and main goals of this thesis. The next chapter presents the clinical definition and symptoms of MDD and other clinical backgrounds and characteristics of speech analysis.

Chapter 2

Theoretical Background

2.1 Clinical View of Major Depressive Disorder

Depression is among the common mood disorders and refers to a condition where overall mental functions, such as thinking content, thinking process, motivation, interest, behavior, sleep, and physical activity continue to deteriorate. These symptoms generate various problems over time and induce MDDs. MDDs refer to cases where symptoms of the noted diagnostic criteria appear for more than two weeks. Often, when someone says they have depression, they note such symptoms which may recur. Depression represents a pathological condition, not a chronic disease or disability. Repeated and recurrent depressive disorders appear more than once continuously. Long-term outpatient treatment and medication can prevent a recurrence, and neighbors can reduce social dysfunction and prevent short-term drugs or suicide. Even if depression is not continuous, most of them go back and forth between remission and depression. That is, it is chronic depression. Current approaches to monitoring MDDs mainly rely on intermittent reports from affected individuals, family members, and friends. These reports are often subjective and include cognitive limitations and social stigma [20] and disregard and overestimate the patient's condition. In psychiatry, a state of depression does not mean a state of temporary degradation.

2.1.1 Types of Depression

Depression can occur in various forms. There are various clinical definitions, including major depression and bipolar disorder within a broad, generalized category called depression. Moreover, there are various types per the manifestation pattern within a major depression. Melancholia-type depression describes a situation where a person wakes up early in the morning with a clear sense of loss of pleasure, appetite, and weight and is severely sensitive to trivial things, even feeling pain. A person with atypical depression has no problem with sleep, but is very sensitive to criticism or rejection from others. They are greatly influenced by surrounding situations and social relationships. If depression is accompanied by psychosis, it signals a dangerous condition where one may hear voices and think they are in danger; emergency treatment is required. Thus, MDDs have a diverse range of definitions. Although the symptoms are not worse than major depression, it is called mood failure when the mood decline continues for more than two years. Generally, it lasts for most of one's life, and one often remains alone and fails to adapt to social life. With bipolar disorder, there is a manic and repeated state of feeling excited, active, and full of confidence, culminating in depression, calming down, and losing confidence. If depression starts at the time of the first outbreak, it can be challenging to distinguish it from major depression. It has a stronger genetic impact than any other type of depression; thus, similar depression often occurs in children when parents have bipolar disorder.

MDD has a detailed disease definition system that shows a wide variety of patterns. Moreover, it is a state of severe depression that requires immediate treatment along with bipolar and mood disorders. However, many types of depression are challenging to identify. As one proceeds from adolescence through adulthood to old age, depression may stem from different external or internal factors at each period. Moreover, seasonal depression, depression accompanied by alcoholism and drug abuse, postpartum depression, and menopause depression also stem from physical disorders caused by various diseases.

This thesis mainly addresses MDDs, which are very severe stages of depression that require immediate treatment, and focuses on patients with severe depression rather than those suffering from mild depression.

2.1.2 Major Causes of Depression

The exact cause of depression is not yet clearly known; it is thought to stem from various causes rather than a single cause. As with other mental disorders, it is generally believed that a variety of biochemical, genetic, and environmental factors contribute to depression. Below are some factors noted as leading causes of depression.

Abnormalities in the neurotransmitter system The three main neurotransmitters that affect the onset of depression are serotonin, noradrenaline, and dopamine. Before depression sets in, the signal diminishes or becomes confused. Medications for depression increase the neurotransmitters serotonin, noradrenaline, and dopamine at synapses between nerves and nerves in the brain [21, 22].

Genetic factors Depression is not hereditary; however, many people with depression have a family history of it. Depression is closely related to genetic predisposition. Bipolar disorder (manic depression), one of the types of depression, has a greater genetic predisposition and is 24.5 times more common than the normal group. People with major depression are three times more likely to observe major depression in their siblings and children than in the control group [23].

Aging and disease As we age, our ability to acquire new knowledge, adapt to change, and recall memories declines. Neurotransmitters in the brain also decrease with age. Some older adults who start with dementia experience severe depression in the early stages, sometimes with psychotic depression. Depression can also be caused when a person suffers from a life-threatening disease such as cancer, a temporary heart attack, or a terminal illness.

Stress in daily life Depression can also occur after personal traumatic events such as the death or divorce of a close friend or family member. Studies like Holmes-Rahe Life Stress Inventory quantifies the effects of personal events on depression [24].

Table 2.1 Symptoms of depression [5]

Psychological symptoms	Physical symptoms
- Continuous low mood or sadness - Feeling hopeless, helpless, irritable and intolerant of others - Having no motivation and not getting any enjoyment out of life - Feeling anxious or worried - Having suicidal thoughts or thoughts of harming yourself	- Moving or speaking more slowly than usual - Changes in appetite or weight - Unexplained aches and pains - Lack of energy - Changes to your menstrual cycle and disturbed sleep

2.1.3 Symptoms of Depression

The symptoms of depression are complex and vary among people. For example, two people with completely different symptoms could be diagnosed with the same major depressive disorder. The symptoms persist for weeks or months and are bad enough to interfere with your work and social and family life [5]. From Table 2.1, such symptoms of depression may appear as physical, psychological, and social symptoms, where people avoid social activities or lose interest in daily life. Depression in children and teenagers appears as a sense of helplessness in daily life, such as refusing to go to school or losing weight [25].

2.1.4 Diagnosis of Depression

Depressive disorders induce challenges in the functioning of patients and maintaining daily activities and increase the risk of suicide. Thus, appropriate treatment through early detection is very important. Given that depression often recurs and becomes chronic, it is crucial to detect it early and accurately measure and diagnose the severity of the symptoms. Therefore, there is steady

Table 2.2 Depression assessment instruments [6]

Scale	Type	Number of items
Beck Depression Inventory (BDI)	Self-report	21
Center for Epidemiologic Studies Depression Scale (CES-D)	Self-report	20
Hamilton Depression Rating Scale (HAM-D)	Clinician-led	21
Montgomery-Åsberg Depression Rating Scale (MADRS)	Clinician-led	10
Patient Health Questionnaire (PHQ-9)	Self-report	9
Geriatric Depression Scale (GDS)	Clinician-led	30

progress in developing evaluation tools for measuring depression and diagnosing depressive disorders. However, there are so many assessment tools that it is challenging for clinicians or researchers to select the most appropriate tool for evaluation. From 2.2, self-report measurement is a method of evaluating participants' answers to specific questions on a questionnaire or test paper and can be performed within about 10 minutes. It has the advantage of allowing for periodic checks of the treatment's effectiveness. However, it is difficult to accurately detect the facts if the patient does not respond frankly and denies the truth [26]. A higher level of depression diagnosis is also possible through structured interviews and clinical evaluation. It has the advantage of a standardized format, furnishing insight into the overall picture of depression by considering biometrics social factors, and allowing for filtering out errors in incomplete or inconsistent evaluations through scrutiny.

2.2 Objective Diagnostic Markers of Depression

The depression assessment instruments described in Table 2.2, a diagnostic criterion for mood disorders, including depression, maintains a practical position. However, there is no practical objective indicator, as the diagnosis includes values. Further, there is a limitation in that only unified judgment is guaranteed and the function of setting the boundary between the summit and the ideal cannot be properly performed. Clearly, symptoms of mood disorders induce biological, social, and behavioral responses that cannot be reflected in the results of the depression evaluation tool. The early diagnosis of depression and the development of practical and quantitative methods for determining treatment options are critical, considering medical, economic, and social costs. Moreover, biological indicators can help with fewer side effects and more effective antidepressant selection.

2.3 Speech in Mental Disorder

Mental disorder refers to the impairment of a wide range of mental functions with various symptoms. It is usually a combination of abnormal thoughts, emotions, behaviors, and abnormal relationships with others; it is characterized by significant clinical impairment in an individual's cognitive and emotional control behavior. Further, it is often associated with pain or damage to important functional areas [27]. There are many types of mental disorders, such as schizophrenia, depression, and intellectual disabilities caused by drug abuse. Speech is a set of recorded utterances used as a basis for language descriptive analysis. Speech corpus, a database of subjects' voices, provides many suggestions for the diagnosis and treatment of mental disorders in patient studies such as neuroscience, sociology, and psychopathology. Spoken language research on patients is an essential approach to understanding the mental activity of the human brain and can be learned by artificial intelligence for early screening and diagnosis. By combining corpus with mental illness, researchers can extract linguistic features from many common facts, and use linguistic, psychological, medical, and other interdisciplinary means to reveal the expression, behavior, and brain processing of pathological groups [28].

Schizophrenia Schizophrenia is characterized by significant perceptual impairment and behavioral changes. Symptoms may include persistent delusions, hallucinations, disordered thinking, very disordered behavior, or extreme agitation. People with schizophrenia can experience persistent challenges in cognitive function [29]. Attempts to analyze the speech characteristics of patients suffering from schizophrenia have been intermittent, and classification performance shows 93% accuracy [30]. When schizophrenia develops and reaches the language proficiency division stage, text analysis is also possible. This task is mainly studied as a combination of speech and speech-transcribed text.

Bipolar Disorder Bipolar disorder is a type of mood disorder that exhibits mania and depression illustrations. Illustration means the symptoms do not continue, but appear for a certain period and show a repeated pattern toward improvement. In general, mania refers to a condition where one feels very good and elevated more than usual. People with bipolar disorder experience alternating bipolar symptoms and depression. During a depressing episode, the person experiences a depressing (sad, irritating, and empty) mood, pleasure, or interest in a particular activity almost every day. The speech analysis for bipolar disorder accords with the voice analysis of depression because depression is also included in the area of bipolar disorder. A representative speech corpus is captured in the AMoSS Interview Dataset [31], and several studies have attempted to analyze the correlation between speech characteristics and bipolar patterns [31, 32, 33].

Attention deficit hyperactivity disorder / Autism Attention deficit hyperactivity disorder is a major neuroactivity disorder in childhood, often with motor and sensory symptoms that persist into adulthood. Autism spectrum disorder is a category of neurodevelopmental disorders characterized by limited and repetitive ranges of behavioral patterns, interests, and activities while exhibiting sustained damage to interchangeable social communication and social interactions from early childhood. The relationship between Speech and such Neurodevelopmental disorders induces movement control disorders such as speech motor control, and several studies show that pronunciation accuracy and speech rate decline [34, 35]

2.4 Speech Production and Depression

When the speaker wants to speak, he first exhales through the lungs, and the vocal cords vibrate by the exhaled breath (exhalation) to produce sound (voice). The voice produced is refined into various speech sounds we can perceive according to the activities of organs in the oral cavity, such as the tongue, lips, and soft palate. The voice is amplified per the shape of the vocal cords extending through the mouth, producing the speech sound. The production of language begins with the underlying cognitive process. It plans the content and structure of the speech, activates the movement, and drives the speech-generating system to command the voice responsible for generating the appropriate sound. Human speech stems from a very complex anatomy, providing the ability to utter a variety of acoustic signals in a harmonious and meaningful way. Given that such human vocalizations have very complex patterns, they are biomarkers for various human health conditions. Moreover, what the speaker is trying to say, the storage of working memory, and the planning and execution of utterances are closely related to human cognitive aspects. For the most part, emotional states are associated with physiological responsive responses (e.g., changes in the autonomic and somatic nervous systems), which affect many aspects of the speech-generating process [36]. The sympathetic excitation associated with an anger state often causes changes in breathing and increased muscle tone, which affects the vibrations and shape of the vocal cords and the acoustic characteristics of speech [37]. Further, the effect of emotion on speech production works in the same way as the emotion of depression.

Based on this assumption, several prior studies observe the voices of depressed patients and find common speech characteristics. Depression-induced neurotransmitter system abnormalities manifested in alterations in muscle tone and control affect the rhyme and quality of the produced speech. Disturbances in muscle tone affect vocal cord behavior, and changes in respiratory muscles alter subglottic pressure. Prosody traits have also been shown to be affected by the speaker's level of depression [38, 39, 40, 41].

2.5 Automatic Depression Diagnostic System

Many methodologies have been devised to construct an automatic depression diagnostic system based on a conventional depressed speech corpus. Constructing an automatic depression diagnostic system can be largely divided into the process of feature representation for mathematically capturing speech characteristics and the design and experimentation of machine learning and deep learning (DL) algorithms to model it. Researchers have attempted to ascertain various objective multimodal measures for depression [42]. Among them, vocal acoustic features were of particular interest because they are generated by a complex neuromuscular system that allows for assessing psychomotor disturbances distinctively seen in depression [43]. These features included (1) prosodic effects such as a decrease in pitch variability [44], energy variability [45], and overall speech rate [46]; (2) changes in the voice quality such as an increased aspiration [40] and spirantization [47]; (3) alterations in formant features [48]; and (4) the effect on the speech spectrum such as a decrease in the subband energy variance and changes in energy distribution [16]. Using such changes in vocal acoustic features, several automatic classification systems for depression have been developed of which classification accuracy ranges between 60 and 80% [42]. However, they have several methodological drawbacks this study addresses. First, some tests were developed using unstandardized speech samples, such as naturalistic interactions with family members [49] and autobiographical stories [50] that are subject to interference from environmental stimuli, interviewer bias, or willingness of the interviewee relative to the standardized mood induction procedures [51]. Second, none considered the different emotional reactivity to the speech contents. Apparently, people with depression show blunted reactivity to positive and negative emotional stimuli from the environment [52]. Third, acoustic features are inherently different between the elderly and young and middle-aged adults and between males and females. For example, a higher shimmer and jitter and a lower harmonics-to-noise ratio were reported in the elderly relative to young and middle-aged adults [53]. Moreover, lower fundamental frequency (F0), lower F0 standard deviation, higher absolute jitter, and higher soft phonation index appear in elderly males relative to elderly females [54]. However, most prior studies analyze the relationship between depression and acoustic features without considering age or sex differences but incorporating them into a single model [55, 50, 56]. Further, few addressed the diagnostic performance of voice-based tests for depression in the elderly for males and females, separately.

This study introduces the detailed process of building an automatic depression diagnostic system from the extraction of acoustic features to modeling speech data.

2.5.1 Acoustic Feature Representation

Speech parameterization or feature extraction is the conversion of raw speech signals into more abstract representations of less redundant signals. A commonly used method constructs raw audio into short frames of tens of milliseconds, computes it through mathematical formulas, and vectorizes it through superposition. It is generally extracted from voice samples on a short time scale but may also include extraction on a much longer time scale. Frames extracted in this chronological order can be aggregated into delta and delta-delta coefficients and observation units the experimenter wants to observe. Thus, it is very suitable for observation. This new structured representation is more suitable for use in classification or prediction systems. Next, we will explain in more detail the acoustic features used for common depressed speech tasks.

Prosodic Features The prosodic characteristics are a composite of hyper-segmental acoustic features of speech (i.e., beyond the vocabulary, syntax, and semantic content of the signal). Its main features are the fundamental frequency (F0) (recognized as pitch), the intensity (recognized as loudness), the speed spoken in normal conversations, the rhythm, and the timing recognized as pattern formation. Related features include jitter and shimmer (change between periods of frequency and intensity), energy distribution between forms, and cepstral feature. In fact, the fundamental frequency (F0, vocal cord vibration ratio) and energy are the most widely used prosodic features because they are related to the perceptual properties of pitch and volume. Speech in depressed patients is generally described as dull, monotonous, and lifeless, showing a prosodically consistent tendency, such as pitch reduction, pitch range reduction, slow speech rate, and pronunciation error. Many linguistic studies examine such characteristics [57, 58, 59]. In particular, F0 is affected by personality characteristics related to a person's basic mood change, level of agitation and anxiety, and depression [60]. Another likely prosodic feature for recognizing depression is language speed; many studies report that patients suffering from depression speak at a slower rate than normal people [61]. Furthermore, recent results indicate that the change in speech velocity is potentially stronger when extracted at the phoneme level of speech production [62].

Voice Quality Features Voice quality is a characteristic auditory color of an individual's voice derived from various laryngeal features and continuously executed through the individual's speech. The unique way of speaking is reflected in how a particular person makes a particular voice [63]. It can be an important feature because clinically depressed conditions can affect laryngeal control, which captures information related to airflow through the portal from the lungs, a source of voice production. Speech quality measurements include jitter, a small period-to-period change in speech pulse timing, amplitude in planetary regions, and harmonic-to-noise ratio (HNR) of harmonic contrast (spectrum). Given the assumption that motion delay in depression reduces laryngeal tension results, aspiration, jitter, and shimmer, defined as the reciprocal of HNR, are reported to have a significant correlation with depression severity (HAMD) [40]. Scherrer's studies [64] also show a high association between speech quality characteristics and depression severity.

Formant Features The vowel is a voiced sound where the airflow caused by vocal cord vibration passes through the pharynx or oral cavity. Depending on the front and rear positions of the tongue, it is divided into high and low and round and flat vowels per the height of the tongue. The resonance frequency varies per the location and degree of the point narrowed by the shape of the vocal cords. Resonance in the vocal tract is referred to as a function, and the corresponding frequency is referred to as a function frequency. The pause frequency is sequentially displayed as F1, F2, F3, F4, and Fn from a low frequency [65]. Formant features are widely used in speech-based depression classification systems. When the first three formation frequencies and bandwidths were grouped, significant differences were found between patients with depression and control [49].

Spectral Features Spectral features are characterized by speech spectra and frequency distribution of speech signals in high-dimensional representations at specific time instance information. Commonly used spectral features include Power Spectrum Density and Mel Frequency Cepstral coefficient (MFCC). In particular, MFCC is among the most widely used spectral features used for speech parameterization. In a recent study, Cummins [66] found a significant negative correlation of depression levels with time derivatives with added MFCC in feature and acoustic space dispersion, where subband energy variability decreased with increasing depression levels [40]. Most studies of speaker recognition (SR) tasks and other common speech analyses employ such spectral features, which are also widely used in the field of depression recognition model formation.

2.5.2 Classification / Prediction

Many attempts have been made to diagnose depression using the speech characteristics listed in the previous section. The development of an automatic depression diagnosis system based on speech can be divided by the problem definition. First, the task of checking the existence of depression is considered binary classification, which is a problem of distinguishing between subjects with and without depression. Moreover, it is possible to define a problem by classifying the severity of depression based on the severity of depression measured by the depression assessment tool. Zimmerman's study [67] defined classification by defining the severity of depression in four stages (no depression (0–7); mild depression (8–16); moderate depression (17-23); and severe depression (>24)) based on the HAMD scale. The problem definition of the depression diagnostic model is also defined by the method of predicting the score of the scale of the depression assessment tool. These two problems are also defined in a mixed format. That is, depression exists if the predicted scale is greater than or equal to a certain value through depression scale prediction. Score-level prediction is the assignment of unknown speech samples to successive value mental state evaluation scale scores. The performance of a score-level prediction system is reported nominally as a measure of the difference between predicted and observed values, such as root mean square error (RMSE) and mean absolute error.

Various attempts have been proposed for speech-based automatic depression classification systems. Moore's work [68] used statistical measurements (comparison of pairwise ANOVA) to construct classifiers using second-order discriminant analysis and reported up to 91% accuracy for male speakers and 96% accuracy for female speakers using leave-one between cross-validations. Moreover, the classification accuracy reported by Low's study [49] is 50% to 75% for class 2 gender-independent Gaussian Mixture Model (GMM) classifiers. There are also multi-feature attempts, and several studies investigate the suitability of individual prosody, voice quality, spectrum, and speech features. Cummins' work [66] shows 79% performance with a combination of MFCCs and formant features among a wide range of acoustic and spectral features tested using GMM. In combination with Principal Component Analysis, Helfer [69] classified the Mundt dataset into upper and lower classes, comprising samples of speakers with 17 or more HAMD scores or seven or less, using features based on the first three morphological trajectories and associated dynamic (velocity and acceleration) information. Using the GMM backend yields ROC 0.7. The audio/visual emotion task competition events in 2013 (AVEC 2013) [18] and 2014 (AVEC 2014) [70] included participants predicting individual self-reported levels of depression (Beck Depression Index, BDI, Score) in a given multimedia file (using multimedia signal processing techniques). From the ranking of the competition, various methods of score prediction techniques have been proposed.

Fortunately, the rapid development of DL has motivated DL approaches for depression recognition and has been free from the design of mid-level feature extraction methods. Given that the DL-based approach has developed based on the vision community, many studies have proposed an algorithm that combines audio and visual information. In the case of the Daic-Woz database, a representative depression database, it is suitable to use visual and audio cues as multi-inputs because it contains visual and speech information on the facial expression of the interviewer. However, given database limitations, DL-based methodologies do not account for 100% of all approaches, and various methodologies close to existing methods have been steadily proposed [71] as deep neural network architectures. For depression score estimation, He [72] proposes a 1D-CNN and 2D-CNN-based methodology to estimate the severity of depression through attempts to fuse manually crafted and deeply learned functions in speech and shows the performance of RMSE of 10.00 and 9.98 in AVEC2013 and AVEC2014 databases, respectively. Further, Dong [73] shows RMSE performance of 8.73 and 8.82 on AVEC 2013 and AVEC 2014 databases, respectively, using deep SR and speech emotion recognition features learned with pre-trained networks. However, since most of the studies are multi-modality studies such as visual, text, and audio, DL methodologies using only audio information have rarely been proposed.

Chapter 3

Developing Sentences for New Depressed Speech Corpus

3.1 Introduction

Collecting the voices of people suffering from depression is an essential starting point for automatically identifying depression. Moreover, using a person's voice is a reliable tool for judging depression. Given that the constructed speech corpus is subject to observation, modeling, and analysis, databases of non-generalized distorted distributions cannot provide accurate insights to experimenters. The corpus built for speech-based research can be evaluated as representation and balance. Representation of the speech corpus is obtained per population or speech selection and overall quantity and size, and balance is obtained per the balance and weight between speech chunks that make up the corpus. Various voice corporations are being built in various ways per their respective purposes in various research institutes, but it is challenging to find one with both representation and balance. It is essential to establish a voice corpus with representation and balance in research examining various aspects of the voice of depression subjects, such as social linguistics, and data to ascertain the voice characteristics of patients with depression for phonetic purposes. Beyond the establishment of an automatic diagnostic system, speech corpus is essential as basic data for research on other speech engineering and language pathology.

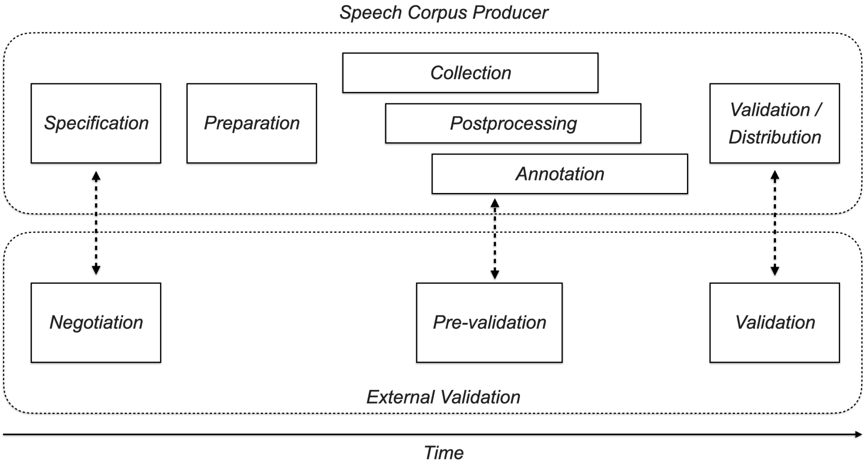
Accordingly, this chapter addresses the process of developing sentence sets that will be the centerpiece of the new depressed speech corpus construction. After discussing the limitations of conventional depressed speech corpora through the construction process of a general speech corpus, we introduce why we should develop a sentence set for the depression detection task. Further, we introduce the process of designing utterance sentences.

3.2 Building Depressed Speech Corpus

3.2.1 Elements of Speech Corpus Production

This section describes the process of speech corpus production in chronological order. Figure 3.1 shows the relationship between the main steps of the process and the time axis from left to right. The specification of the task is preferentially set to produce a speech corpus. This step assumes control of the overall design of the speech corpus, and most of the intention of the speech corpus producer is contained per the purpose and object of the task and observation unit. From the preparation through the specification stages, the steps depend on the results or data generated in the previous stage. Thus, the order must be strictly followed. The full-scale creation stage can be conducted in parallel. Post-processing and

Fig. 3.1 Schedule of a speech corpus production



annotation run in parallel on many corpus production collections, saving time. Moreover, an independent validation agency should conduct external validation. In most cases, such a design is not feasible given a lack of funds. However, at least, internal verification must be performed. The detailed description of all the tasks in Figure 2 is as follows.

Speaker Specification The most important element of setting the specification of the speech corpus is to set and recruit the speaker's profile. Regarding constructing a speech corpus for general speech recognition, demographic factors (gender, age) are evenly distributed. Controlling the speaker's linguistic characteristics (dialect, mother tongue) is the most important factor, but the characteristics of the speaker for this study are pathological factors. Tasks for diagnosing mood disorders set speaker profiles based on the prevalence of mood disorders, and recruitment of normal groups without mood disorders is crucial to model the difference between normal groups and patients.

Speaking Contents / Style The speech corpus producer sets the task by directing the speaker to a specific utterance task (along with a ”virtual machine” in one or more human conversation partners or wizard of Oz [WoZ] experiments). Usually, the content of speech corpus for speech recognition should be stated in phonological units such as phonemes, syllables, and morphemes, but tasks such as emotion recognition or depression diagnostic modeling should be specified in higher-level units (vocabulary, sentences) than phonological units. Moreover, the content of the speech corpus collection can be controlled by setting a specific conversation topic. The style setting of speech instructed to the speaker is an element that determines the format of the speech corpus, such as read, question, answer, command/control speech, descriptive speech, and spontaneous speech formats.

Recording Setup Basically, the recording settings define the acoustic properties of the resulting corpus, thus defining the usefulness of the data for a particular application or investigation. A good way to elicit a very natural and spontaneous way of speaking is to do tasks that require some cognitive activity for the speaker. Depending on the location definition, there are telephone, on-site, and field recordings and WoZ methods. Furthermore, a technical specification that specifies the sampling rate, number of channels, and file formats of audio files is also an essential step.

Collection In this step, the details set in the previous step of recording the speech signal, which is the core of all speech corpus production steps, should be performed. High-quality assurance must be achieved through continuous documentation, pre-validation, quality control, and data pipeline creation.

Annotation / Validation In general, there is no time relationship information about the script content unless the speech chunk is associated with the script chunk. If the corpus comprises paragraphs of read text, each signal file stores one paragraph of speech, and one paragraph is mapped with information corresponding to the purpose of the task. For example, regarding the depression diagnostic model formation that corresponds to the subject of this study, the scale index of depression or the prevalence of information are recorded in each signal file. However, if transcribed information is required within the signal file, detailed time information is recorded approximately when each guest-specific word starts and ends within the signal file. Evaluation through an external validator is very important because such annotation operations may have human errors in the operator. Because internal validation does not tend to be very effective, external validation is recommended whenever possible. It is important to perform this activity as often as possible. The signal data is mainly recorded properly, and the effectiveness of the annotation and the completeness of the documentation are evaluated.

3.2.2 Conventional Depressed Speech Corpora

The previous part describes the main elements from pre-experimental design to actual speech data collection to build a depressed speech corpus. Based on these collection protocols, various depressed speech corpora have been developed through previous studies and have been built to create an automatic depression diagnosis system. The overview of depression speech corpora is given in Table 3.1.

Among these, DAIC-WoZ [17] is the most popular audio-visual depression language corpus used in several system proposals for a long time. Based on clinical interviews designed to support the diagnosis of psychological distress conditions, the database was constructed using a speech corpus through clinical interviews with a computer agent. Through teleconference, face-to-face, automated agent, and WoZ-based speech data construction, vision information containing the face data of the subject being interviewed, speech recordings, and transcribed text data are provided. Eventually, they employed 621 subjects. In particular, WoZ-based speech data, widely used in speech-based system design, was collected through 142 participants. For each participant, DAIC-WoZ provides a patient health questionnaire (PHQ-8) [74] score, indicating depression severity. Moreover, labels of PHQ-8 scores are also provided to indicate the presence of depression. A score greater than or equal to 10 indicates that the participant is suffering from depression.

The second most popular speech corpus, AViD-Corpus, was collected by 292 participants in human-computer interactions by answering a series of queries (freeform parts). They were asked to recite excerpts of fables. Participants were labeled with a Beck Depression Index-II (BDI-II) score [26]. BDI-II is a list of 21 multiple-choice depression scores from 0-63. If the participant's BDI-II score exceeds 29, it indicates the presence (absence) of depression. The average age of the participants was collected at 31.5 years and was used as a target voice corpus for AVEC 2013 [18].

The Emotional Audio-Textual Depression (EATD)-Corpus [75] and Turkish databases [76] stand out among non-English depression speech corpus. EATD is the widest collection of depressed speech corpus collected in Chinese, and 162 participants (30 depressed participants) were collected for college students

(20s). The participant's depression scale measurement is made by SDS [77], and the total length of speech chunks is 2.26 hours. The Turkish database is a representative database collected under the initiative of the hospital and can be said to be a clinic-based depression corpus database built through interviews with clinics. BDI-II was used as a measure of determining depression. Finally, we constructed a depression speech corpus based on 70 participants.

DEPression and Anxiety Crowdsourced corpus is the most recently studied case of depression corpus and has been employed to collect large training datasets to identify candidate speech and language features specific to a given mental illness. Participants could collect several subjects by completing the tasks using Amazon Mechanical Turk, a platform where individuals are paid to complete short tasks online. PHQ-9 and GAD-7 were used as the standard depression scale, and 2,674 speech chunks were finally collected.

Table 3.1 Overview of depression speech corpora

Corpora	Language	Collecting Methods	Subjects	Age	Depression Scale	Validated by clinician	Additional Notes
Audio-Visual Depressive Language Corpus (AViD-Corpus) [18]	German	Interview, Reading	292	Mean age: 31.5	BDI-II	X	Audio-visual / AVEC 2013
Distress Analysis Interview Corpus (DAIC-WOZ) [17]	English	Face to face, Teleconference, Wizard-of-Oz, Automated agent	621	Age range: 18 - 60	PHQ-8	X	Audio-visual / Transcribed
EATD-Corpus [75]	Chinese	Interview	162 (30 depressed)	20s (student)	SDS	X	File length: 2.26 hours
Turkish database [76]	Turkish	Interview	70	Mean age: 34	BDI-II	X	Mean BDI-II : 23.45
DEPAC [78]	English	Interview	571	Age range: 18 - 76	PHQ-9, GAD-7	X	2,674 samples

3.2.3 Factors Affecting Depressed Speech Characteristics

As this study must detect the emotional aspects of the subject included in the speech, it should generalize various factors that may affect speech characteristics, which should be clearly designed early in the development of the depressive speech corpus. Several factors that must be controlled as much as possible to build a reliable depressed speech corpus are as follows: [79].

(1) Demographic/biological factors: race, gender, age, ethnicity

(2) Cultural background: mother tongue, language accent, dialect

(3) Social background and personal characteristics: social position, friendship, relationship with family, or personal taste and personality

(4) Current emotional state: anger, happiness, joy, or sadness

(5) voice pathology: speech disorders, intoxication, and respiratory tract infection

Given that the factors can directly affect the characteristics of human speech, they must be considered when selecting subjects for the corpus. The imbalance of the factors induces the tendency of speech characteristics, with a risk of minimizing the speech characteristics of depression. Another challenge in constructing a depression speech corpus is that it is challenging to recruit subjects diagnosed with depression. Most depression diagnoses employ a questionnaire method, and the scale measurement of this questionnaire method can be dynamically changed per social affects and current mood states. However, speech characteristics do not change as dramatically from day to day as sudden mood swings. Thus, to address the limitations, qualitative opinions must be considered through the measurement and interview by professional clinicians using the depression scale; much money and time are necessary to create a speech corpora that satisfies all prerequisites.

3.3 Motivations

Through the construction process of the depressed speech corpus, it was possible to understand what processes must be performed to build a reliable speech corpus, and what types of conventional depressed speech corpus have been proposed. This section introduces what motivation induced the development of a set of speech sentences for the depressed speech corpus.

3.3.1 Limitations of Conventional Depressed Speech Corpora

Not Validated by Clinician For all conventional depressed speech corpus, depression scales such as BDI-II, PHQ-8, PHQ-9, SDS, GAD-7, and HAM-D have been annotated as diagnostic evaluation tools. Before using this depression scale as an absolute judgment indicator for modeling and analysis, the following questions should be asked.

"Are these depression scales absolutely representative to judge the subject's depression? "

The depression scales are interpreted as showing symptoms of depression if more than a certain number of items are checked based on a questionnaire that checks for several symptoms. However, it is only a one-dimensional measure of depressive symptoms. For example, according to DSM-IV, five out of nine depressive symptoms should appear in major depression. Given that a fixed number of five items can refer to different items for different patients, completely different patients can meet these symptom requirements. Thus, this heterogeneity has serious limitations in the predictive validity of diagnosis related to treatment selection [80]. In addition, an improvement in the total depression scale during the administration of antipsychotic drugs does not in itself qualify the drug as an antidepressant [81]. The limitations of the depression scale are clearly revealed in the literature, and the final judgment of depression should be from a macroscopic perspective through interviews with the clinic, long-term observation, and the depression scale.

Does that mean the depression scale should not be used for the annotation of the depressed speech corpus? Although the depression scale has several limitations, it can be sufficiently used as one of the observation indices of this task. However, there are many limitations to using binary decisions to diagnose final depression. The final criterion for judgment should be whether there is a prevalence diagnosis by the clinic. This limitation is the reason the automatic depression diagnosis system should not be a model for predicting depression scale.

What is the optimal speech collection method? As noted, one of the most important factors in the early stages of speech corpus construction is to define the subject's speaking style, method, and contents. It has a profound influence on speech recording and the order and method of experiments:

(1) Reading Sentences

(2) Interview

(3) Wizard of Oz

(4) Descriptive Speech

The speaking method used by most conventional depressed speech corpora is the interview and WoZ method. The interview method is suitable for acquiring the subject's free speech and has been used in emotional speech recognition and various mood disorder diagnosis systems. Acquisition of speech by the interview method was considered to have the greatest advantage in that it was the closest method to daily life, as it naturally elicited the emotions of the subject. How the interviewer obtains only answers based on the questionnaire prepared in advance and the free interview data in a natural atmosphere have different characteristics. Prepared questionnaire-based interviews can be analyzed at the semantic level of answering questions; regarding free interviews, specific intentions can be carried out, as per the interviewer's competence. However, given that it is based on mutual communication with the interviewer, the social aspect can influence the interviewee to act unnaturally. WoZ-style interviews are human-computer interaction methodologies that can compensate for the noted shortcomings. It is possible to obtain more natural speech data than interview-based data. Moreover, text analysis through transcribed data is also possible for the above two methods. However, the free interview-based speech-collecting method has a fatal disadvantage in that the interviewer's intention can affect the subject's emotions. For example, if emotional induction occurs given the personal experience of a specific subject, it will cause emotional changes only in some of the depression groups. It acts as a very large error in setting up a speech corpus with a generalized representation. Further, prosodic factors can act as variables because the content pronounced for each topic is different, which is a disadvantage.

The speech recording method of reading sentences was mainly used in speech recognition tasks to construct phonemic pronunciation data. Given that it is a method of reading pre-written sentences, it is challenging to induce natural emotions, and because the subject has to intentionally pronounce certain sentences, it is considered challenging to construct natural speech data. Thus, it is rarely used as a method of constructing a depressive speech corpus. However, given that all subjects pronounce the same sentence, variables caused by prosodic differences are removed and are more suitable for development as a diagnostic tool. Interview-based data has real-world application limitations because interviewers must always exist, and the method of reading structured sentences is simpler in the form of diagnostic tools. This thesis builds a suppressed speech corpus of reading sentences by highlighting the advantages.

3.3.2 Attitude of Subjects to Depression: Masked Depression

As noted, the speech corpus of the sentence reading method has several limitations but with a suitable form as a depression diagnosis tool. Moreover, it can express natural emotions. However, depression shows very different behaviors in emotional speech. Emotions of depression deepen in the direction of reducing productivity in all humans, which may induce the inability to see various physical symptoms of depression. It is called masked depression. [82, 83, 84] Masked depression can occur more easily among celebrities and emotional workers who must show only good looks. A father who must endure the weight of social or family responsibilities can also suffer from masked depression. In particular, adolescents who have not yet completed their cognitive development often do not properly recognize depression and cannot express it. In this case, depression can be expressed in behavioral problems such as irritation, violence, deviant behavior, and school rejection. If the elderly suddenly lose their memory, they may not think of it as dementia at first, which may stem from depression. Masked depression is no longer used as a diagnosis, likely because the term is ambiguous and the list of related symptoms is so wide that it often induces misdiagnosis. Without a clear understanding of the effects of depression on the body, symptoms can be misunderstood as a physical disease. However, treating physical symptoms without treating underlying depression is not effective. It is said that the groups in which masked depression may mainly occur are

(1) The elderly

(2) Children and adolescents

(3) The socially marginalized (in the United States, Africans)

(4) Asian (patriarchal society)

(5) The chronically ill person

These factors of masked depression are factors that can occur in various age and social groups and are also the foundation for depression. The depression scale is a good example for people who, though not on the low scale, have been diagnosed with depression. When building an interview-based speech corpus, the appearance of masked depression is an obstacle because the subject feels burdened by the existence of the interviewer and hides the true feelings. Further, there is a limit to using the depression scale as an absolute judgment indicator in conventional depressed speech corpus. The purpose is to maximize masked emotions while minimizing the intervention of speech collectors to create a depressive speech corpus that overcomes the factors of masked depression. For this, a sentence reading type collection method is more suitable.

3.3.3 Emotions in Reading

It is necessary to eliminate masked emotion through the composition of sentences that the subject will utter. That is, a process that evokes specific emotions through reading is necessary. Accordingly, we extend the logic based on one psychological hypothesis.

Hypothesis: The subject's emotions be driven to a specific state by reading sentences containing emotional content

In many studies, emotional changes in the reading process affect the understanding of reading, though with a small but partial correlation between the emotions contained in the text and the emotions of the subjects. Lai's research [85] shows that brain activation in emotion-related areas is activated when negative and neutral sentences are read to understand emotions implied in sentences; thus, implicit emotions contribute to the activation of language-related fields. Hence, we intend to construct a sentence set with emotional intent, which is one of the most important motivations of this chapter.

3.3.4 Objectives of this Chapter

Based on the motivations in the previous section, we would like to propose a new method of depressed speech corpus. As one of the preliminary tasks, we would like to construct a sentence set to be read by the subject. We build a new type of depressed speech corpus to achieve the following objectives.

(1) We present a sentence reading-based speech corpus methodology and develop emotional sentences (negative and positive) to overcome the masked depression effect prior speech corpora did not consider.

(2) We develop sentences based on neutral emotions to observe changes in speech characteristics of EPS that may occur in the treatment process of antipsychotic drugs.

In particular, to achieve the second purpose, neutral sentences are constructed by considering the hypothesis of the utterance of emotional sentences established in the process of establishing motivation. That is, to observe the change in speech characteristics related to the decrease in speech motor function, sentences with neutral meaning were constructed to control the change in characteristics that may occur given the emotional sentences.

3.4 Proposed Methods

Two steps are needed for the development of sentences. First, words are selected to be placed in a sentence. Second, a complete sentence structure is designed based on the selected words. All sentences were constructed based on Korean. The following sentence development process was co-developed with psychiatrists at Seoul National University Bundang Hospital (SNUBH). The suggestion of emotional words in the word selection process was proposed by the SNUBH's psychiatrists. Finally, the arrangement of words, the composition of complete sentences, and the evaluation of final sentences were conducted jointly.

3.4.1 Selection of Words

Multiple Affect Adjective Checklist-Revised (MAACL-R) [86], the Korean version of the Positive Affect and Negative Affect Schedule (PANAS) [87], and the Korean Affective Word List [88] were used to select the words that would make up the sentences. Among them, the most widely known MAACL-R is a 70-item subset of the original 132 items, which includes five measures: anxiety, depression, hostility, positive effects, and sensory pursuit. PANAS is a self-report questionnaire comprising two 10-item scales to measure both positive and negative impacts. A revised Korean version was used in this study. Further, the Korean Affirmative Word List, which scaled the emotional impact of Korean words, verified the emotional values of all words in the sentence.

3.4.2 Structure of Sentence

After the process of forming the types and structures of sentences based on the words selected in the previous step, the type of sentence was defined in three ways—positive, negative, and neutral—because emotions largely comprise two factors: positive and negative emotions. Positive emotions are enthusiastic, active, high energy, and marked by full concentration, while negative emotions refer to the general dimensions of various hateful emotional states, including anger, contempt, nausea, guilt, fear, and unpleasantness. Such negative emotions are important factors that distinguish depression from anxiety [89, 90]. Neutral sentences simply comprised elements corresponding to the main components of the subject, object, predicate, and complement sentence. Regarding emotional sentences, adverbs reflecting each emotion were used. Moreover, sentences were designed to have a greater influence on emotions, as they used actual events that had a great influence on the Korean people.

Table 3.2 Mood-inducing sentences

Induced Mood	Number	Description
Neutral	T_1	The night and day on the spring and autumnal equinox are of equal length.
	T_2	Seoul is the capital city of Korea.
	T_3	The day when the moon appears the largest is known as a full-moon day.
	T_4	Fourteen plus eight is twenty-two.
	T_5	Dogs can smell better than people do.
Negative	N_1	I feel sad* whenever I think of the young high school victims of the Sewol ferry disaster‡.
	N_2	I always break into tears‡ when I recall those moments with my spouse who passed away‡.
	N_3	I feel distressed† because my child failed‡ the College Scholastic Ability Test again.
	N_4	I fall into despair‡ because the cancer relapsed again.
	N_5	I feel very hopeless‡ at the bankruptcy‡ of my company.
Positive	P_1	I feel joyful* that Korean national soccer team makes it to the semi-finals of the World Cup.
	P_2	I am happy* that my partner always stands by me.
	P_3	I'm so glad* because my child passed‡ the College Scholastic Ability Test.
	P_4	I am pleased* that the cancer is in full remission‡.
	P_5	I am very happy* that I finally get a job‡.

*: selected from the Multiple Affect Adjective Check List-Revised (MAACL-R) [86].

†: selected from Korean version of the Positive Affect and Negative Affect Schedule (PANAS) [87].

‡: selected from the Korean Affective Word List [88].

3.5 Results

3.5.1 Mood-Inducing Sentences (MIS)

MIS were developed to create a depressive speech corpus that overcomes the first goal of masked depression. We employed a mood induction protocol originally proposed by Velten [91] using a set of emotionally charged sentences written in Korean to manipulate mood experimentally. We developed 15 sets of MIS: five sentences for negative mood induction, another five for positive mood induction, and the remaining five for a matched neutral condition (Table 3.2).

3.5.2 Neutral Sentences for Extrapyramidal Symptom Analysis

The sentences to be pronounced by the speaker were constructed such that words that could affect the speaker's emotions were not included. The words to compose the read sentences can affect the speaker's emotions per the emotional meaning they contain; they can also have a great influence on the extraction of speech feature vectors [92]. Based on MAACL-R [93], PANAS [87], and Korean Affective Word List [88], we constructed neutral words that do not have an emotional impact and composed a set of 16 neutral reading sentences, listed in Table 3.3. In addition, by putting ”Ah” utterances corresponding to exclamations, we attempted to collect voice information of different patterns from voice characteristics generated by word pronunciation.

Table 3.3 Sentences of EPS speech corpus

Sentence Type	Description
SA	Ah
S1	The night and day on the spring and autumnal equinox are of equal length.
S2	Seoul is the capital city of Korea.
S3	The day when the moon appears the largest is known as a full-moon day.
S4	Fourteen plus eight is twenty-two.
S5	Dogs can smell better than people do.
S6	Sitting for long periods of time makes your back stiff.
S7	I go to church every Sunday.
S8	Seniors 65 and older can use the subway for free.
S9	Dinosaurs went extinct hundreds of thousands of years ago.
S10	The wine color of this bottle is red.
S11	If you go out on the main road, you can catch a taxi.
S12	The items are placed in a basket in a locker.
S13	what to eat for lunch?
S14	Do you know where the car keys are?
S15	Don’t forget to brush your teeth before going to bed
S16	Choose the one that best fits the text above

3.6 Summary

This chapter lists the limitations of conventional depressed speech corpora and explains what methods are available to design a new method of depressed speech corpus. We proposed a reading-sentence method rather than a free interview method to compensate for the trend caused by masked depression missed by all prior depressed speech corpus. Sentence sets were designed to construct it. MIS, an emotional word-based mood-inducing sensation, was developed based on the fact that it affects the reader's emotions when reading emotional content. Further, a neutral sentence set was also developed for EPS observation.

Chapter 4

Screening Depression in The Elderly

4.1 Introduction

The previous chapter identified the limitations of conventional depressive speech corpus, proposed a new method of constructing a depressive speech corpus, and proposed a sentence reading-based collection method beyond the widely-used interview-based method. Two types of sentence sets were developed for subjects to read, and MIS, constructed in chapter 3.5.1, is employed in this chapter. MIS is a sentence set developed to direct emotional changes in a speech by placing emotional words in sentences; that is, it is more practical and challenging to apply than conventional depression diagnostic system studies. The elderly were selected to identify the masked depression element, which is the limit of the conventional speech corpus noted in chapter 3.3.2. Research on the diagnosis model of depression in the elderly begins with the following motivations.

The difficulties in diagnosing depression in the elderly MDD is one of the most common and debilitating ailments in the elderly [94]. Current methods of evaluation of this disorder hinge on subjective complaints from patients and clinical judgment based on the symptoms and signs proposed by, for example, the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders-Text revision (DSM-IV-TR) [95]. Thus, it is challenging to be diagnosed when the patient is defensive or repressing the symptoms, especially in the elderly. Depression is high among low-income people, the elderly, and female seniors. Although it is a relatively common psychiatric disease among the elderly population, the symptoms often overlap with physical symptoms. Common depressive symptoms include mood loss, decreased motivation, decreased appetite, changes in the waterfront, anxiety, anxiety, fatigue, guilt, decreased concentration, and suicidal thoughts. The clinical pattern of elderly depression does not differ significantly from that of young-age depression, but there are several differences in the frequency of each symptom. Relative to depression among other age groups, melancholic depressive symptoms that are very similar to masked depression such as apathy, decreased mood responsiveness, decreased appetite, and excessive guilt are common. That is, the diagnosis of depression among the elderly is clinically challenging. Moreover, approximately 12% of the elderly population suffer from senile voice disorder; representative symptoms include limitation of sound intensity, weak voice, coughing, and hoarse voice, which are also factors that make it challenging to analyze speech in the elderly [96].

Limitations of related research Despite many attempts to study the diagnosis of depression in the elderly, most examinations have been socio-scientific, psychiatric, and neuropsychological studies. [97, 98, 99] These attempts define the depressive symptoms of the elderly and provide great insights. Several automatic classification systems for depression have been developed, their accuracy ranging between 60 and 80% [42]. However, they have several methodological drawbacks that this study addresses. First, some tests were developed using unstandardized speech samples, such as naturalistic interactions with family members [49] and autobiographical stories [50], that are subject to interference from environmental stimuli, interviewer bias, or willingness of the interviewee, relative to the standardized mood induction procedures [51]. Second, none considered the different emotional reactivity to the speech contents. Reportedly, people with depression show blunted reactivity to positive and negative emotional stimuli from the environment [52]. Third, acoustic features are inherently different between the elderly and young and middle-aged adults and between males and females. For example, a higher shimmer and jitter and a lower harmonics-to-noise ratio were reported among the elderly, relative to the young and middle-aged adults [53]. Moreover, elderly males exhibited lower fundamental frequency (F0), lower F0 standard deviation, higher absolute jitter, and higher soft phonation index relative to elderly females [54]. However, most prior studies analyzed the relationship between depression and acoustic features without considering age or sex differences but incorporating them into a single model [55, 50, 56]. Further, few addressed the diagnostic performance of voice-based tests for depression in the elderly for males and females, separately.

Hence, this chapter develops a voice-based screening test for MDD using the variation pattern of vocal acoustic features of elderly Koreans (males and females) while they read a series of MIS. We hypothesized that certain acoustic features can differentiate individuals with MDD from those without, and these features are different between males and females.

4.2 Korean Elderly Depressive Speech Corpus

This section introduces the actual process of building a Korean-based elderly depressive speech corpus based on MIS. It describes the selection of the recording population, the recording process, and other specifications and introduces the clinical assumptions established through the recording process. This course was conducted by geriatric psychiatrists at Seoul National University Bundang Hospital (SNUBH), and the design and planning of the speech corpus construction experiment were co-developed.

4.2.1 Participants

We recruited 61 individuals aged 60 or older with depression from the Korean Longitudinal Study on Cognitive Aging and Dementia (KLOSCAD) [100], an ongoing nationwide, community-based prospective cohort study of elderly Koreans. KLOSCAD sampled 6,818 participants randomly from 30 villages and towns across South Korea using residential roasters from November 2010 to October 2012 and followed them every two years. This addendum study enrolled KLOSCAD participants from the Jukjeon district of Yongin city, who were evaluated at SNUBH from March 2015 to May 2016. Geriatric psychiatrists made a diagnosis of MDD using the Korean version of the Mini International

Neuropsychiatric Interview (MINI-K) [101] following the criteria by the fourth edition of the DSM-IV-TR. All participants with MDD had their voices recorded properly, which was included in the analysis. A total of 160 individuals in the age-matched healthy control for each male and female group were recruited from KLOSCAD during the same period. From them, 17 individuals were excluded given a recording error, leaving 143 individuals for the final analyses.

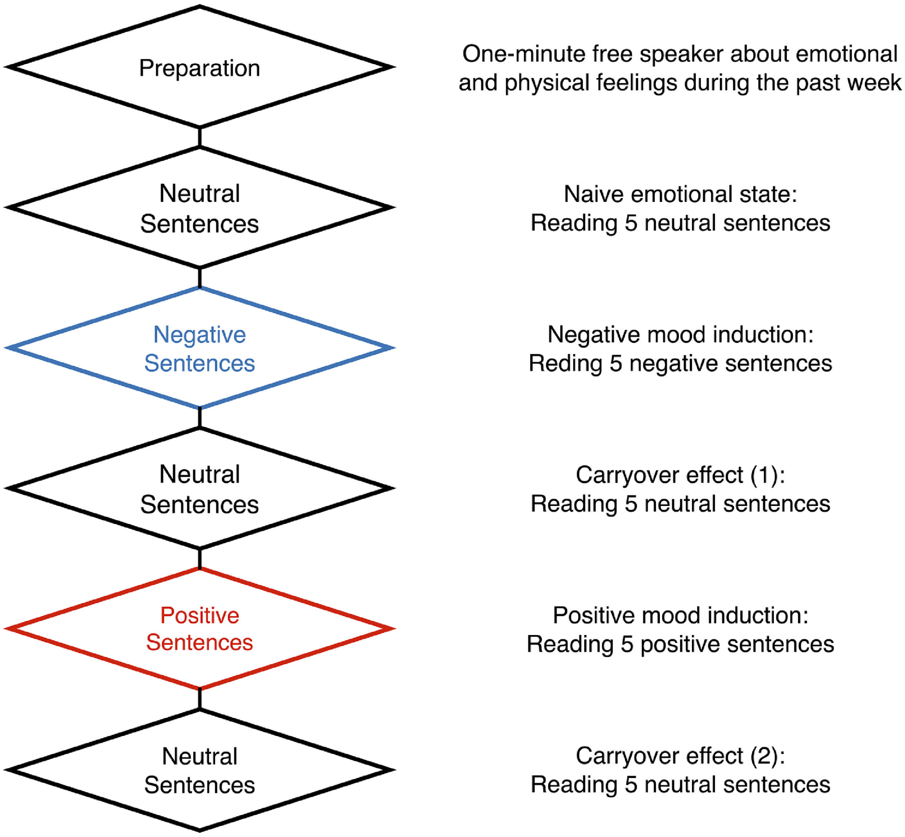
The exclusion criteria for all study participants were as follows: having evidence of impaired consciousness such as delirium; inability to read or understand sentences written in Korean; current diagnosis of any serious medical or neurologic disorders that negatively affect cognitive or language functions; and history of any substance dependence. These were assessed through a face-to-face standardized diagnostic interview for every participant, including physical and neurological examinations, using the Korean version of the Consortium to Establish a Registry for Alzheimer's Disease Assessment Packet Clinical Assessment Battery (CERAD-K-C) [102] and MINI-K by geriatric psychiatrists, CERAD-K Neuropsychological Assessment Battery [102, 103], Digit Span Test [104], and Frontal Assessment Battery [105] by trained neuropsychologists or nurses. We also conducted laboratory tests, including complete blood cell counts, chemistry panels, apolipoprotein E genotyping, and a serologic test for syphilis. A consensus diagnostic conference attended by geriatric psychiatrists confirmed the final cognitive status of the participants. Information on demographic variables; use of psychotropic drugs, including antidepressants, antipsychotics, and anxiolytics in the past month; the duration of the current episode, the number of lifetime depressive episodes, and the scores of the Korean version of the geriatric depression scale (GDS-KR) [106] were also obtained. This study was approved by the Institutional Ethics Review Board of SNUBH and written informed consent was sought from all participants.

4.2.2 Recording Procedure

One of the biggest keys to this study is to read the three emotional types (neutral, positive, and negative) sentences developed earlier in any order. The three emotional types were cross-arranged to confirm the effectiveness of the emotional type itself and study the relationship between emotional types. After reading one sentence, the induced emotional state was expected to be maintained when reading the next sentence. Thus, the interval between sentences was induced not to exceed two seconds to observe the transition of emotions between sentences. Before the induction, participants were asked to freely talk about their mood and physical states during the previous week for a minute. We then presented five neutral sentences (#1 – 5, in order of appearance), followed by five negative sentences (#6 – 10), five neutral sentences (#11 – 15), five positive sentences (#16 – 20), and five neutral sentences (#21 – 25) (Fig. 3.1). We placed an identical set of five neutral sentences immediately after the positive and negative MIS to explore the carryover effect of participants' previous mood status [107].

The carryover effect is a term used in clinical chemistry to describe the transfer of unwanted substances from one container or mixture to another, which, if applied to this study, affects the emotional state of the sentence to be read. That is, emotional metastasis observation is a study to confirm whether it is possible to maximize the difference between the depression group and the normal group by generating a combination of the direct effects of emotional

Fig. 4.1 Overview of recording session procedures.


Preparation
One-minute free speaker about emotional and physical feelings during the past week
Neutral Sentences
Naive emotional state: Reading 5 neutral sentences
Negative Sentences
Negative mood induction: Reding 5 negative sentences
Neutral Sentences
Carryover effect (1): Reading 5 neutral sentences
Positive Sentences
Positive mood induction: Reading 5 positive sentences
Neutral Sentences
Carryover effect (2): Reading 5 neutral sentences

content and the effects from previous types of sentences. Further, the one-minute free speech information in the preparation stage allows us to observe the difference between free speech and reading speech.

4.2.3 Recording Specification

Research physicians recorded speech data using a Tascam iXZ Microphone attached to participants' chests with a sampling rate of 44.1 kHz in a predefined outpatient clinic of SNUBH. To minimize the effect of the noise, the recording was made in a quiet room and a wind mask was attached to the microphone. These data were stored at a sentence-level unit in an unrecoverable format and then transferred to a cloud server where they were subsequently analyzed. We created a PowerPoint presentation for the MIS, which provided visual instruction through a Samsung Syncmaster S23B350T monitor and audio instruction through a Samsung SHS-260W headset. Participants were instructed to navigate the PowerPoint slides by themselves.

4.3 Proposed Methods

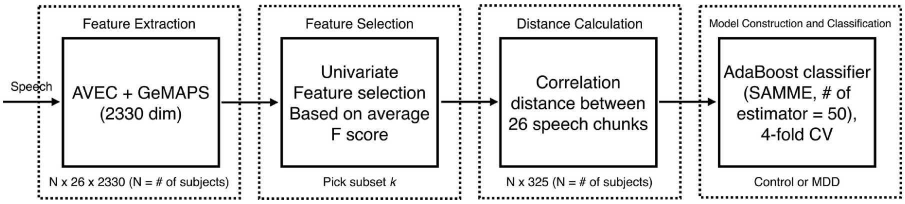
4.3.1 Voice-based Screening Algorithm for Depression

Fig. 4.2 presents the overall framework of the voice-based screening algorithm for depression (VoiSAD) from its preprocessing to the classification. The development of VoiSAD was designed sequentially with the extraction of acoustic features, a feature selection process that reduces the dimension of a feature vector, a distance calculation step that represents the distance between 26 sentences, and a final classification of the prevalence of depression using the distance vector formed in this way. Although the method of learning by listing speech feature vectors in frame units has been mainly used in previous studies, this study attempted to propose a semantic-level analysis method that merges them into sentence units and analyzes them. This chapter details each step.

4.3.2 Extraction of Acoustic Features

First, we removed the silences from the recorded speech samples using a voice activity detection algorithm from the VoiceBox toolkit for Matlab [108]. The

Fig. 4.2 Procedure for developing diagnostic tool and its overall system.


Feature Extraction
Speech
AVEC + GeMAPS (2330 dim)
N x 26 x 2330 (N = # of subjects)
Feature Selection
Univariate Feature selection Based on average F score
Pick subset k
Distance Calculation
Correlation distance between 26 speech chunks
N x 325 (N = # of subjects)
Model Construction and Classification
AdaBoost classifier (SAMME, # of estimator = 50), 4-fold CV
Control or MDD

AVEC = Audio/Visual Emotion Challenge, GeMAPS = Geneva Minimalistic Acoustic Parameter Set, dim = dimensions, SAMME = Stagewise Additive Modeling using a Multi-class Exponential loss function, CV = cross-validation, MDD = major depressive disorder.

speech data were then analyzed with OpenSMILE v2.1.0 [109] using the Audio-Visual Emotion Challenge 2013 (AVEC 2013) audio baseline feature set [18] and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [110]. Thus, the input audio signal was split into multiple overlapping frames obtained by calculating particular mid-level features. The frames compose a time-series sequence of feature vectors and were used for computing secondary feature statistics. The first one, AVEC 2013, comprises 2,268 features that include 32 energy- and spectral-related low-level descriptors (LLD) and six voicing-related fundamental LLD, delta coefficients of each of the LLD, and 10 voiced/unvoiced durational features (Table 4.1). The second set, eGeMAPS, contains 62 features, including a compressed set of 25 LLD (frequency-related, energy/amplitude-related, and spectral parameters) and percentile-related functionals (Table **??**). These feature sets show high robustness for emotional speech recognition tasks [111] and have been standardized through the years of the workshop [18, 70, 111]. Altogether, 2,330 features using the two sets were extracted, and each speech unit (a sentence) was quantized as one size of a vector with 2,330 feature dimensions with a frame length of 1.25 milliseconds. One speech unit was about 1 to 2 seconds long for each MIS and 4 to 5 seconds for the free-talking session.

Table 4.1 Low-level descriptors from AVEC 2013 and eGeMAPS

Feature set	**Parameters**	**LLD**
AVEC 2013	Energy- and Spectral- related (n = 32)	Loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz, 25%, 50%, 75%, and 90% spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity, flatness, and MFCC 1 – 16
	Voicing-related (n = 6)	F0 (sub-harmonic summation, followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), and logarithmic HNR
eGeMAPS	Frequency-related (n = 8)	Pitch, jitter, formant 1,2, and 3 frequency, formant 1, and formant 2-3 bandwidth
	Energy/ Amplitude-related (n = 3)	Shimmer, loudness, and HNR
	Spectral-related (n = 14)	Alpha ratio, Hammarberg index, spectral slope 0-500 Hz and 500-1500 Hz, formant 1, 2, and 3 relative energy, harmonic difference H1-H2, harmonic difference H1-A3, MFCC 1-4, and spectral flux

AVEC 2013, Audio-Visual Emotion Challenge 2013; eGeMAPS, extended Geneva Minimalistic Acoustic Parameter Set; LLD, low-level descriptor; MFCC, Mel-frequency cepstral coefficients; HNR, harmonics-to-noise ratio.

4.3.3 Feature Selection System and Distance Computation

After extracting acoustic features, we designed a feature selection system to choose a subset of significant features that could reflect the feature variation pattern between two speech units when an individual reads a series of MIS. One of the distinguishing attributes of this algorithm is that we only used the variation pattern of an individual as an input feature (not individual features, per se) when modeling the decision classifier, thereby ignoring the unique speech characteristics of the individual. Thus, the classifier can avoid the speaker verification problem by not trying to find a similar tone of voice and be free from the influences of voice aging. After that, we employed the concept of distance between two speech units to represent the acoustic features as follows.

In the feature selection system, given a set of acoustic features $Y = [y_1, y_2, \ldots, y_s]$ from an individual, each $y_s = [y_{s1}, y_{s2}, \ldots, y_{sf}]$ is standardized as $z_{sf} = (y_{sf} - \bar{y_f})/\sigma_f$ and a set of normalized features $Z_f = [z_{1f}, z_{2f}, \ldots, z_{sf}]$ is computed to reduce the dependency and redundancy of the feature and to improve its integrity, where s stands for the number of 26 speech units recorded consecutively (y_1: one minute-free talking, y_2: the first MIS, . . . ,y_{26}: the last sentence of MIS) and f stands for the feature dimensions. The differential between the two speech samples of z_{if} and $z_{jf} \in Z_f$, where $i<j$, can be described as $d_f = z_{if} - z_{jf}$. Moreover, a set of differentials, D $= [d_1, d_2, \ldots, d_f]$, from study participants is used in the feature selection system to select the most significant features that can reveal the difference in the feature variation patterns.

From the scientific perspective, the distance (i.e., dissimilarity) between two vectors is defined as a quantitative degree of how far apart two objects are [112]. The dissimilarity, d, between the speech vectors u and v was calculated from

$d = 1 - \frac{(u-\bar{u})\cdot(v-\bar{v}))}{\|(u-\bar{u})\|_2\|(v-\bar{v})\|_2}$, where $\bar{u}$ is the mean of the elements of u, $\|\cdot\|_2$ is the L-2 norm, and u·v is the dot product of u and v. Using the value of dissimilarity, we calculated the internal dissimilarity vector of a specific individual with the size of 325 (26 distinct speech samples taken two at a time). We also produced self-similarity matrices [113] for each of the four groups, control female, depressed female, control male, and depressed male, to visualize the degree of dissimilarity between two sentences in a 26 × 26 square matrix.

We employed a univariate filter method of feature selection based on the F-score. This method has several advantages in that it readily scales to high-dimensional datasets, and it is computationally fast and simple and independent of the classification algorithm [114, 115]. It could be particularly useful when each feature is considered separately. A feature relevance score is calculated with the averaged F-score of 325 observations. Ultimately, feature vectors with 1 × 325 dimension size computed from preprocessing are used as the input features for the classifier. It remains unknown as to how many and what kind of acoustic features will be used for calculating the dissimilarity between two feature vectors until the validation test is conducted.

4.3.4 Classification and Statistical Analyses

Regarding the demographic information, we used the Mann-Whitney U test to compare group differences for continuous variables and the $\tilde{\chi}^2$ test for categorical variables. The normality of the distribution of data was evaluated by the Shapiro-Wilk test.

We employed the AdaBoost classifier as a decision tool [116, 117, 118] to classify a certain variation pattern of an acoustic feature set between the depression and control groups. This boosting algorithm uses a set of weak learners to form a highly accurate prediction rule by calling the weak learner repeatedly. Additionally, it selects only useful dimensions of a feature set to improve the predictive power of the model, thereby avoiding the curse of dimensionality, minimizing the computational cost [119], and reducing the possibility of overfitting. We used Stagewise Additive Modeling using a Multi-class Exponential loss function algorithm [120], a multi-class generalization of AdaBoost, to measure the feature relevance score by calculating the feature importance based on the information gain [121]. The speech samples were split into three partitions for males and females separately—a training (60%), validation (20%), and test (20%) set—and four-fold cross-validation was used. The training set was used to build the classifier with 50 rounds of random sampling. We also generated a receiver operating characteristics curve and computed the area under the curve (AUC), sensitivity, and specificity for male and female participants, separately. The validation set was used to determine the optimal subset k from the feature selection system, which maximizes the AUC.

We conducted additional analyses demonstrating the emotional reactivity or its carryover effect after a participant read certain MISs to examine the validity of the MIS. We analyzed the effect of sex and mood status on the dissimilarity between two sentences by performing the analysis of covariance (ANCOVA) with sex and mood status as between-subject factors and age, years of education, and current use of psychotropics as covariates. These covariates are known to be associated with speech patterns or depression [53, 20, 54]. The Statistical Package for the Social Sciences (SPSS) for Windows, Version 20.0 (SPSS Inc., Chicago, IL, USA) was used to conduct statistical analyses, and we

Table 4.2 Demographic characteristics of participants.

Variables	Male		Female		*Statistics*[a]	
	Control (N = 50)	MDD (N = 18)	Control (N = 50)	MDD (N = 18)	*p*[b]	*p*[c]
Age, years	72 (5)	75 (6)	17 (5)	71 (6)	0.009	0.061
GDS, points	6 (4)	18 (6)	7 (4)	20 (6)	0.060	<0.001
Psychotropics use, N (%)	0 (0)	11 (61)	1 (1)	19 (44)	0.783	<0.001
Duration of the current episode, years	N/A	2.86 (2.34)	N/A	3.81 (4.67)	0.448	
Number of lifetime depressive episodes	N/A	1.48 (0.51)	N/A	1.46 (0.61)	0.896	
Antidepressants, *mg*[d]	N/A	87.50 (74.65)	N/A	97.19 (66.76)	0.773	
Antipsychotics, *mg*[e]	N/A	28.41 (13.39)	N/A	18.94 (0)	0.667	
Anxiolytics, *mg*[f]	N/A	13.29 (18.41)	5 (0)	7.44 (6.17)	0.228	

Values are mean (SD), unless specified otherwise.
a, Mann-Whitney U test for continuous variables, and χ2 test for categorical variables; b, p value between sex groups; c, p value between mood groups; d, chlorpromazine equivalent dose; e, imipramine equivalent dose; f, diazepam equivalent dose.
MDD = major depressive disorder, GDS = Geriatric depression scale, N/A = not applicable.

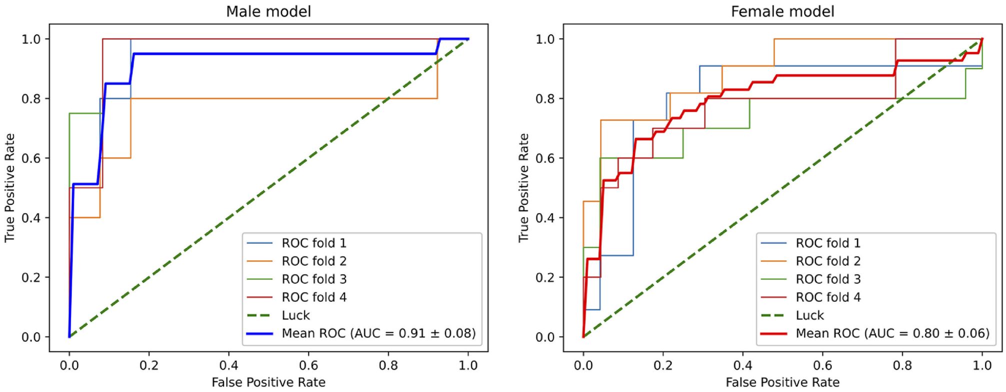
employed 2-sided significance at the 0.05 level.

4.4 Results

The demographic characteristics of the final sample are given in Table 4.2. The depression and control groups did not differ significantly in the mean age. The depression group had significantly higher GDS scores and current rates of psychotropic use. Males and females had comparable GDS scores and current rates of psychotropics use, while males were older than females. Additionally, depressed males and females did not differ in terms of the duration of the current episode, the number of lifetime depressive episodes, and the equivalent doses [122] of psychotropics. Excluded individuals because of recording errors (N = 17) were not different from those included in the final analyses in age (mean [SD], 71.7 [4.2] vs 71.6 [5.5]; p = 0.878), female proportion (52.9% vs 66.7%; 0.252), years of education (mean [SD], 13.4 [4.2] vs 12.7 [4.2]; p = 0.498), and depression severity (GDS score [SD], 7.7 [4.9] vs 10.6 [7.7]; p = 0.124).

Based on VoiSAD, the depression classification achieved an AUC of 0.911

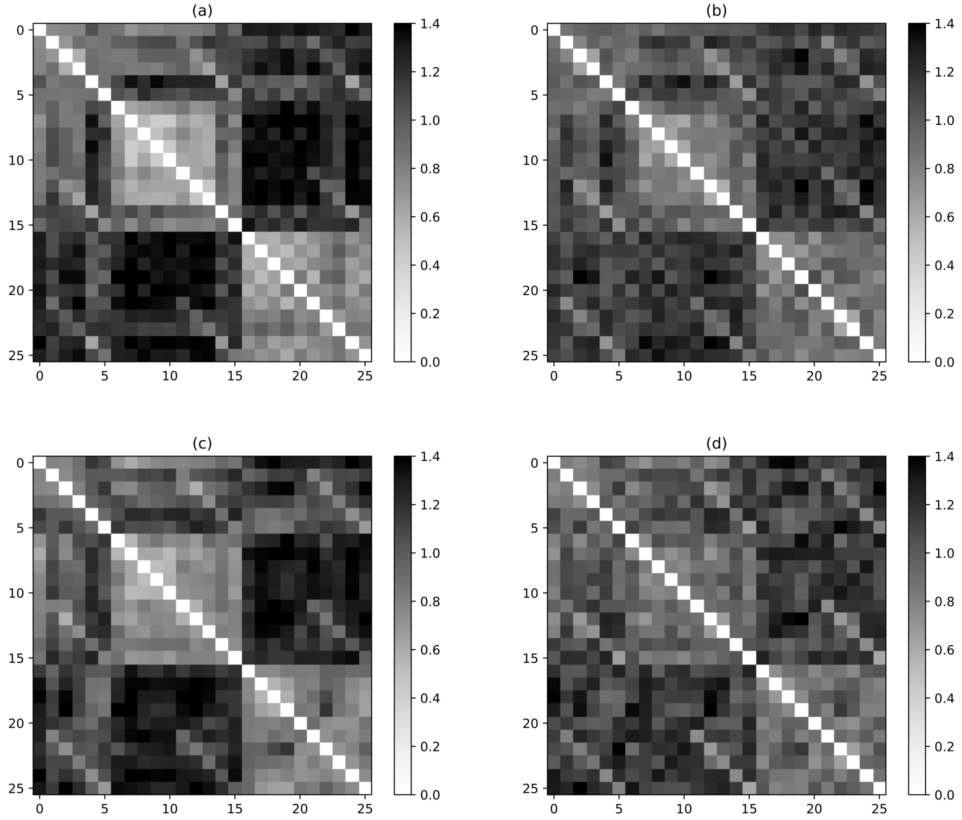
Fig. 4.3 Receiver operating characteristics (ROC) curve of the classification model.



AUC = Area under the curve.

(sensitivity 0.950, specificity 0.881, and accuracy 0.856 ± 0.584) for male participants and 0.799 (sensitivity 0.734, specificity 0.862, and accuracy 0.773 ± 0.347) for female (Fig. 4.3). The number of acoustic features of optimal subset k from the feature selection system ranges between 17 and 43 (mean [SD]; 28.53 [11.26] for males; 31.27 [12.28] for females), amounting to approximately 1% of all features extracted. Only these features were used to calculate the dissimilarity and classify those with MDD and without. The rest of the features did not contribute to the variation pattern, while an individual performs one-minute free-talking and reads 25 MIS. Acoustic features with the top five feature relevance score were as follows: for male participants, spectral and energy-related features including the linear regression slope of the first MFCC, arithmetic mean of loudness, arithmetic mean of the audio spectrum, percentile 20% of loudness, and root quadratic mean of the audio spectrum were found to be discriminative. For female participants, prosody-related features such as the root quadratic mean of fundamental frequency (F0), third inter-quartile of F0, arith-

Fig. 4.4 Self-similarity matrices of distance matrix in depressed and healthy control female and male groups.



(a) Control Female, (b) Depressed Female, (c) Control Male, and (d) Depressed Male.
* The x- and y- axes of the matrix represent the designated number of each mood-inducing sentence, and the greyscale intensity indicates the correlation distance between the two sentences.

metic mean of F0, percentile 80% of F0, semitone from 27.5 Hz, and percentile F0 showed significant discriminatory performance.

Fig. 4.4 depicts the self-similarity matrices showing the emotional reactivity and its carryover effect. The depressed group shows a shorter average correlation distance between negative ($\#6 - 10$) and positive ($\#16 - 20$) sentences than the control group ($F = 18.574$, $p < 0.001$) regardless of sex ($F = 1.368$, $p = 0.244$),

Table 4.3 Dissimilarity distance between free speech and MIS samples

	Male		Female	
S	Control	MDD	Control	MDD
1	1.117 ± 0.220	1.121 ± 0.231	0.946 ± 0.221	1.004 ± 0.273
2	1.125 ± 0.180	1.157 ± 0.174	0.913 ± 0.251	1.046 ± 0.250
3	1.166 ± 0.140	1.150 ± 0.186	0.866 ± 0.221	1.021 ± 0.218
4	1.051 ± 0.209	0.904 ± 0.249	1.154 ± 0.257	0.982 ± 0.246
5	1.161 ± 0.180	1.075 ± 0.234	1.092 ± 0.279	1.129 ± 0.228
6	0.817 ± 0.220	0.818 ± 0.235	0.788 ± 0.243	0.996 ± 0.288
7	0.953 ± 0.224	0.926 ± 0.229	0.753 ± 0.259	1.035 ± 0.284
8	0.893 ± 0.174	0.855 ± 0.156	0.819 ± 0.232	0.856 ± 0.284
9	0.968 ± 0.185	0.995 ± 0.181	0.852 ± 0.236	1.036 ± 0.229
10	1.033 ± 0.231	1.039 ± 0.186	0.876 ± 0.233	1.028 ± 0.229
11	1.156 ± 0.169	1.130 ± 0.213	1.021 ± 0.210	1.092 ± 0.271
12	1.193 ± 0.158	1.197 ± 0.219	0.919 ± 0.261	1.157 ± 0.232
13	1.125 ± 0.207	1.159 ± 0.168	0.914 ± 0.244	1.066 ± 0.229
14	1.031 ± 0.195	1.049 ± 0.210	1.207 ± 0.216	1.034 ± 0.242
15	1.095 ± 0.218	1.064 ± 0.189	1.119 ± 0.246	1.126 ± 0.233
16	1.012 ± 0.188	1.060 ± 0.270	1.029 ± 0.251	0.983 ± 0.215
17	0.992 ± 0.173	0.917 ± 0.211	1.093 ± 0.237	1.110 ± 0.272
18	1.024 ± 0.159	0.983 ± 0.251	1.187 ± 0.244	1.080 ± 0.264
19	1.035 ± 0.175	1.077 ± 0.199	1.235 ± 0.205	1.089 ± 0.248
20	1.068 ± 0.163	1.150 ± 0.145	1.193 ± 0.205	0.990 ± 0.205
21	1.189 ± 0.215	1.250 ± 0.125	1.201 ± 0.213	1.090 ± 0.260
22	1.267 ± 0.161	1.256 ± 0.201	1.156 ± 0.201	1.146 ± 0.204
23	1.194 ± 0.194	1.216 ± 0.186	1.103 ± 0.222	1.096 ± 0.197
24	1.133 ± 0.187	1.068 ± 0.261	1.332 ± 0.198	1.038 ± 0.211
25	1.184 ± 0.194	1.175 ± 0.191	1.299 ± 0.187	1.175 ± 0.171

indicating a dampened reactivity to MIS in the depressed group. The depressed group also shows a shorter correlation distance between the neutral sentences read immediately after the negative sentences (#11 – 15) and the positive sentences (#16 – 20) than the controls (F = 13.647, $p < 0.001$) regardless of sex (F = 0.168, p = 0.682). This situation was the case in the neutral sentences read after the positive (#21 – 25) and negative (#6 – 10) sentences regardless of sex (F = 5.392, p = 0.021 for mood status; F = 0.192, p = 0.661 for sex). The results indicated that depressed patients may have a reduced carryover effect of negative and positive MIS.

4.5 Discussion

Performance of proposed system The VoiSAD classified participants with MDD from the healthy controls with an AUC of about 0.9 in males and 0.8 in females, which represented an excellent to outstanding discriminatory performance [123]. Prior research shows that screening instruments using vocal acoustic features could successfully detect depression in young adults [55, 50, 49] or adults of all ages [56, 124]. Regarding the elderly population, one study examined 16 severely depressed (Center for Epidemiologic Studies Depression Scale score > 25) [125] and 16 non-depressed male participants aged 65 to 82 years old [126]. Feeding spectral features, such as spectral tilt and formants, and prosodic features, including pitch and energy, into a support vector machine (SVM) discriminative classifier, they reported a prediction accuracy of 81.3%, though their limited sample size and uniform sex composition of participants hinder generalizability. However, Fraser et al. investigated 65 depressed (Hamilton Depression Rating Scale score > 7) [127] and 65 non-depressed Alzheimer's disease (AD) patients (mean age [SD], 72 [9]; female 69%) [128]. Feeding textual and acoustic features into an SVM, they reported a classification accuracy of 65.0% for males and 58.8% for females. Although they employed textual features including parse constituents, vocabulary richness, and psycholinguistic measures, their relatively poor performance might stem from the inclusion of individuals with AD of unknown severity. Additionally, both studies determined participants of their depressed group based on screening questionnaires, which are unlikely to accurately detect and classify masked depression. This study made a diagnosis of MDD in a standardized way using MINI-K following the criteria by the DSM-IV-TR. Therefore, this study demonstrates that MDD in the elderly can be successfully detected using vocal acoustic features.

Uniqueness of this study Moreover, to increase the robustness and the reliability of the acoustic features and avoid the confounding effect of unique features of each participant (i.e., speaker identification), we did not use the raw acoustic features but computed the sentence-wise distances among 26 spoken sentences and used the variation pattern in distances as a final input to the classifier. The usefulness of the variation pattern is validated through the experiments, resulting in high AUC scores for unseen speech samples. Consequently, we escaped the problem of artifacts related to speaker identification that might be learned by the classification model, precluding its generalization.

Previous research shows that older adults have different features of depression than younger adults. Older adults face challenges in expressing their depressive mood and often hide their depressive mood behind somatic symptoms [9].

Furthermore, as elderly individuals are likely to live alone or experience social isolation [129], it is challenging to discover their depressive symptoms promptly. Therefore, if voice-based screening tests for depressive disorders using acoustic features such as VoiSAD are disseminated widely through the internet or information and communications technology, they could contribute to the early detection of masked depression, especially among the elderly population.

Which speech characteristic was a good discriminator? The results show that the vocal acoustic features that could discriminate individuals with MDD from healthy controls were quite different between males and females: spectral- and energy-related acoustic features for males and prosody-related features for females. The analyses show that, for males, different energy-related features manifested as a lower loudness (i.e., low intensity in individuals with MDD relative to healthy counterparts). Reportedly, the spectral-related features might correspond to a wide range of phonetic events of the speaker, including nasals, vowels, or fricatives reflecting a vocal tract configuration [130]. Moreover, the cepstral features extracted from a spoken utterance were known to be closely related to its linguistic content [131]. Therefore, when a uniform linguistic content such as MIS is analyzed, MFCC related to vocal tract features might be a sensitive biomarker of depression in males. However, for females, prosody-related features such as F0 were useful to discriminate MDD patients from healthy controls. F0 is known to be influenced by hormonal changes, including shifts in the testosterone-estrogen ratio, leading to thickening or edematous vocal folds in females [132]. Given that estrogen has been linked to a higher incidence of depressive disorder in females than in males [133], it is tempting to suppose that F0 may reflect the physiologic nature of MDD in females.

What was the tendency of the carryover effect? In this study, we collected voice data while participants read MIS developed from a series of emotionally charged Korean words using the Velten method [91]. It enabled us to analyze the voice reactivity to the negative or positive MIS and its carryover effect by examining the correlation distances of a set of sentences. From the self-similarity map and the ensuing ANCOVA, individuals with MDD showed a dampened reactivity to positive and negative stimuli and a reduced carryover effect. The findings accord with a recent study stating that MDD is better characterized by a significantly attenuated emotional reactivity to positively- and negatively-valenced stimuli, rather than a combination of increased negative reactivity (negative potentiation) and decreased positive reactivity (positive attenuation) [134].

Free Speech vs MIS Reading From 4.3, we can see the distance between 1-minute free speech and MIS reading. For men and women, health control has a closer distance from free speech when reading negative and neutral sentences (#11 - 15) and a higher distance when reading positive and neutral sentences (#21 - 25). However, the depression group did not have a large difference in the distance between special positive and negative sentences. There is a greater difference in the induction of positive than negative sentences in the discrimination of depression. Moreover, the similarity of free speech in everyday life to reading negative sentences confirmed that depression among the elderly is a common reality. The free speech of the depression group was generally very different from the characteristics of the sentence reading method.

Limitations There were several limitations worth mentioning. First, a relatively small sample size of individuals with MDD, especially male participants, might have affected the reliability of the analyses. Additionally, given the absence of an independent sample other than the recruited participants, we could not replicate the findings on the emotional reactivity and carryover effect. Second, though we adjusted the effect of psychotropics in the analysis model of the correlation distance, it remains necessary to examine the performance of the classification model for drug-naïve participants with depressive disorders to assure its validity. Third, interference from "demand characteristics" [135] is an issue for this type of mood-inducing procedure (MIP). It is defined as the summation of cues that conveys a predetermined experimental hypothesis to the subject, thus becoming a significant determinant of the subject's behavior. Reportedly, cognitive MIPs, including the Velten method, are vulnerable to this phenomenon [136]. In particular, this limitation should be supplemented when applied as a real-world application. The VoiSAD, however, may be unlikely to be influenced by demand characteristics because the carryover effect is under a non-subjective mechanism. Further, relative to other MIPs using music [137], film clips [138], autobiographical recall [139], or interviews by a trained interviewer, the MIS may be easier to administer and more widely applicable via the internet. Fourth, although we could classify MDD using machine learning, we could not give a definitive answer to which mechanism in depression was discriminative. Fifth, because the study targeted only those with MDD, it may show poor performance for other depressive disorders in DSM-IV-TR, such as dysthymic disorder, adjustment disorder with depressed mood, or other subthreshold MDD.

4.6 Summary

We constructed a Korean-based elderly depressive speech corpus based on MIS to overcome various limitations, including masked depression. Through this speech corpus, VoiSAD, an automatic depression diagnosis system, was proposed based on the distance of acoustic features between sentences, and men and women recorded high performance. Based on the dissimilarity matrix between MIS, the pattern of emotional transfer between the depression and normal groups was significantly different, proving the carryover effect. Vocal acoustic features while reading MIS may be a promising noninvasive biomarker of late-life MDD in both sexes. By confirming its validity through a large sample size and drug-naïve participants in future studies, the classification model should be widely deployed and easily administered for elderly individuals.

Chapter 5

Correlation Analysis of Antipsychotic Dose and Speech Characteristics

5.1 Introduction

In the previous chapter, MIS was developed to build a Korean-based elderly depressed speech corpus, and a sentence-based automatic depression diagnostic system (VoiSAD) based on feature distance between MIS was proposed to be free from bias from inherent speech characteristics. In this chapter, we will study the changes in speech characteristics induced by antipsychotic drugs, the most important factor that can affect speech in the diagnosis and treatment of depression. Accordingly, sentences were developed to observe symptoms from chapter 3.5.2, which is a set of sentences using only neutral sentences, unlike the method of chapter 4. The relationship between antipsychotic drugs and EPSs, the pattern of changes in speech characteristics, and the resulting research motivation are as follows.

Antipsychotic drug and extrapyramidal symptoms Antipsychotic drugs have been used as the main treatment for psychiatric disorders like schizophrenia because they are effective in various psychotic symptoms, such as hallucinations, delusions, and thinking disorders. Antipsychotic drugs non-selectively block dopamine D2 receptors of several dopamine pathways in the brain, inducing a decrease in dopamine transmission in the long term, which is known to show therapeutic effects. The degree of dopamine-2 receptor binding depends on the type of antipsychotic drug, and the degree of anticholinergic action that reduces extrapyramidal side effects also varies. Representative EPSs include Parkinson's syndrome, acute dystonia, akathisia, and tardive dyskinesia.

Changes in speech given extrapyramidal symptoms EPSs are often accompanied by various voice changes, such as hoarseness, tremor, and pronunciation problems caused by abnormal movements of muscles such as the larynx and vocal cords when speaking. These subtle changes in voice may appear earlier than the clinically observed EPSs. Assumedly, the developmental mechanism is similar to the changes in Parkinson's disease [140]. Previous studies report that subtle changes in the voice can be a precursor to Parkinson's disease, and various studies are underway on the possibility of early detection of Parkinson's disease via voice feature analysis [141, 142, 143, 144]. Antipsychotic-induced EPSs are diagnosed based on the patient's subjective report and the physician's physical examination. Abnormal Involuntary Movement Scale (AIMS) [145], Simpson-Angus Scale (SAS) [146], and drug-induced Anxiety Rating Scale

(Barnes Akathisia Rating Scale, BARS) [147] can be used to evaluate its pattern or severity using standardized scales. Voice changes often accompany EPSs, but are rarely considered when evaluating EPSs, and their diagnostic use is limited. Changes in voice characteristics per the level of antipsychotic drug administration have not been studied in-depth. In particular, the investigation of the correlation per dose may contribute to the early detection of EPSs, thereby contributing to dose adjustment of antipsychotic drugs and prevention of side effects.

Limitations of related studies Despite the medical seriousness of EPSs and their strong influence on speech, few studies directly analyze the correlation between speech and EPSs. The study [140] observing the change in the voice of the subjects with EPSs observed the change from the perspective of the clinic, with no quantitative analyses using a database. Most prior studies confirm EPSs from the perspective of Parkinson's disease. However, extrapyramidal and Parkinson's disease show similar symptoms, though the cause of the outbreak is different; if the EPSs persist severely, it has a causal relationship that can cause Parkinson's disease. Examples of quantitative analyses by recording the speech of subjects suffering from EPSs are limited, and Sinha's studies [148, 149] are representative. They compare the characteristics of various elements of antipsychotic drugs (dose, movement scale) and speech. Although it seems similar to the purpose of this study, limited experiments have been conducted on Risperidone drugs, and speech samples are collected in the form of free speech and sentence reading.

This chapter primarily records EPSs per the antipsychotic drug dose and constructs the Korean version of EPS speech corpus using neutral sentences for EPSs following the method devised in this study. It investigates the correlation of speech characteristics per antipsychotic doses based on the configured speech corpus. However, given that this study is an early-stage experiment to study the relationship between antipsychotic dose and negative characteristics, we mainly address the observation of extracted speech characteristics and dose prediction per the antipsychotic dose. Hence, the research problems are as follows.

(1) Can a direct correlation between antipsychotic dose and speech characteristics be observed?

(2) Is there a difference between subjects who experienced EPSs and those who did not?

(3) Which speech characteristics are strongly correlated with antipsychotic doses?

(4) Is there a correlation with indicators (SAS, AIMS, BARS) that evaluate EPSs?

5.2 Korean Extrapyramidal Symptoms Speech Corpus

5.2.1 Participants

The subjects were first selected to build a language corpus in this study, and antipsychotic drugs were administered among patients treated at the psychiatric outpatient clinic of Seoul National University Bundang Hospital from March 2018 to March 2019. Adults between the ages of 18 and 65 were selected for the collection. However, to exclude subjects who can affect speech regardless of whether antipsychotic drugs are administered, uncontrolled physical diseases, people who cannot read Korean, people with personality disorders who can affect participation in research, and recruitment and recruitment are excluded. Finally, 111 recording sessions were performed in 42 subjects, and 1,886 speech chunks were collected. The subjects were taking a total of eight antipsychotic drugs, including Aripiprazole and Clozapine, and EPS evaluation indicators such as SAS, BARS, and AIMS were also measured every recording session. The collection and recording of the subjects were conducted by a psychiatrist at Seoul National University Bundang Hospital. Table 5.1 shows detailed demographic characteristics.

This study was approved by the Institutional Ethics Review Board of SNUBH to collect baseline speech data from the selected subjects, before taking antipsychotic drugs, before increasing drug dosage, reaching maintenance dose, and when extrapyramidal side effects occur and written informed consent was sought from all participants.

5.2.2 Recording Process

Speech data were collected using a designated recording device in a closed medical room at Seoul National University Bundang Hospital, and the data were repeatedly collected whenever the dose of the antipsychotic drug the subject was taking was changed. When pronouncing the "Ah" sound (SA) corresponding to an exclamation, it was induced to be pronounced for two to three seconds without emotion while maintaining a constant pitch as much as possible.

Table 5.1 Description of EPS speech corpus

Participants	42
Recording Sessions	111
Speech Chunks	1,887
Gender	Male: 17, Female: 25
Diagnoses	Schizophrenia, Psychotic disorder, Bipolar disorder, Paranoid shizophrenia
Antipsychotic drugs	Aripiprazole, Clozapine, Amisulpride, Paliperidone, Risperidone, Olanzapine, Haloperidol, Quetiapine
EPS	Positive: 50, Negative: 61
	M (SD)
Age	32.43 (11.38)
Equivalent dose	12.83 (8.88)
SAS	0.45 (0.87)
BARS	0.44 (1.37)
AIMS	0.21 (0.59)

5.2.3 Extrapyramidal Symptoms Annotation and Equivalent Dose Calculations

The evaluation of EPSs, which is the basis of analysis in this chapter, was performed by one psychiatrist before the start of the recording session and was assessed on the AIMS, SAS, and drug-induced Anxiety Rating Scale (BARS). Through the above indicators, the psychiatrist finally diagnosed the onset of EPSs through interviews. Moreover, regarding the antipsychotic drug dose, which is the main comparative measure, the total antipsychotic drug being administered for each recording session after converting all the doses of several antipsychotic drugs into an olanzapine equivalent dose [122, 150, 151, 151] was calculated. Equivalent dose calculation is a necessary process because each subject has a different type of antipsychotic drug and most subjects take multiple antipsychotic drugs simultaneously.

5.3 Proposed Methods

5.3.1 Acoustic Feature Extraction

Table 5.2 describes the method of extracting speech features in this study. Speech analysis based on audio feature extraction has long been used for the analysis of emotions [152, 153], mood disorders such as depression [72], gender and age [154], and various medical conditions. Regarding the analysis of Parkinson's disease, which is the most representative disease caused by a motor disorder and most related to this study, various speech feature extraction methods have been proposed. The study employs Surfboard [155], a Python-based audio feature extraction package, which showed robust classification performance for a Parkinson's disease classification task relative to the existing widely used Opensmile [156] and Praat [157]-based methods.

Thirteen types of speech characteristics, including F0, Formant, MFCCs, Jitter, Shimmer, Harmonics to noise ratio, pitch period entropy, detrended fluctuation analysis, and energy (RMS, log-energy, sliding window of log-energy, loudness) related feature sets, and detailed derivation methods can be found in Lenain, Raphael, et al. [155]. A 405-dimensional feature vector was finally obtained by extracting feature vectors for each frame based on a specific window and hop size for each audio feature and combining them with 24 kinds of statistical methods. A detailed list of features can be found at 5.2.

Table 5.2 List of extracted speech features and statistical methods

Dynamics	Parameters
F0 contour	hop length seconds=0.01, method=’swipe’
F0 statistics	Mean, standard deviation of F0 contour
Log-energy	-
Sliding window log-energy	frame length seconds=0.04, hop length seconds=0.01
Formants	F1, F2, F3, F4
Loudness	-
Energy (RMS)	frame length seconds=0.04, hop length seconds=0.01
MFCCs	nmfcc=13, nfft seconds=0.04, hop length seconds=0.01
Jitters	p floor=0.0001, p ceil=0.02, max p factor=1.3
Shimmers	max a factor=1.6, p floor=0.0001, p ceil=0.02, max p factor=1.3
Harmonics-to-noise ratio	-
Pitch period entropy	-
Detrended fluctuation analysis	window lengths=[64, 128, 256, 512, 1024, 2048, 4096]
Statsitcal Methods	
- Mean, Standard deviation, Skewness, Kurtosis - First derivative mean, First derivative standard deviation, First derivative skewness, First derivative kurtosis - Second derivative mean, Second derivative standard deviation, Second derivative skewness, Second derivative kurtosis -First quartile, Second quartile, Third quartile - Q2-Q1 range, Q3-Q2 range, Q3-Q1 range - 1st percentile, 99th percentile, 99th-1st percentile range - Linear regression offset, Linear regression slope, Linear regression MSE	

Table 5.3 Statistical results of pearson correlation coefficient r (top 20 speech features)

	A	S1 S16	A + S1 S16
	M (SD)	M (SD)	M (SD)
Total	0.322 (0.051)	0.353 (0.048)	0.325 (0.049)
EPS = 0	0.291 (0.024)	0.280 (0.019)	0.244 (0.014)
EPS = 1	0.473 (0.061)	0.531 (0.053)	0.502 (0.049)
SAS > 0	0.501 (0.042)	0.520 (0.055)	0.498 (0.051)
SAS = 0	0.283 (0.023)	0.301 (0.033)	0.263 (0.025)

5.3.2 Speech Characteristics Analysis recording to Eq.dose

Further, to observe the change in the speech feature per the increase in the dose of the substituted antipsychotic drug (Eq.dose), Pearson's correlation analysis between each feature vector and Eq.dose was used as a basic analysis method. It was based on EPSs (EPS=1 or 0), types of utterance sentences (A, S1 to S16), and SAS, a test useful for measuring stiffness and convulsions associated with drug treatment. The correlation between speech characteristics and the equal dose was analyzed following these criteria. Moreover, after analyzing the correlation of each feature vector, the significance of the entire speech feature vector was analyzed using multivariate linear regression analysis.

5.4 Results

Table 5.3 shows the mean and standard deviation of the top 20 speech features with an absolute maximum value among the Pearson correlation coefficients measured per sentence type (A, S1 to 16, A+S1 to S16), EPS, and SAS scale.

In the group with EPSs (EPS = 1), regardless of the type of sentence, a high Pearson correlation coefficient of over 0.5 showed a clear linear relationship. Moreover, the reading sentences (S1 - S16) showed a higher correlation coefficient. Regarding the extrapyramidal group (EPS = 0), the correlation coefficient was between 0.2 and 0.3, indicating weak linearity, and there was a higher linear coefficient in exclamation utterances (SA) than in reading sentences. In the state before the onset of EPSs, exclamation utterances rather than reading sentence utterances are more suitable sentence forms for distinguishing speech characteristics. Regarding the SAS, we could examine the results of a similar tendency that could be found, per the presence of EPSs. At any rate, the group with SAS=0 showed very similar correlation coefficient results to the group with EPSs, and the group with SAS=0 showed very similar results to the group without EPSs. Although the SAS is being used as one of many indicators to determine EPSs, it shows the possibility that it can act as a very important indicator individually.

Table 5.4 shows 20 representative speech features selected through Pearson correlation analysis of all sentence groups (A+S1-S16). In the normal group (EPS=0), MFCCs were selected as a very important index for judgment. However, in the EPS group (EPS=1), energy-related features showed a high correlation coefficient. EPSs are expressed as abnormalities in muscles such as the larynx and vocal cords that move, inducing changes in energy-related voice characteristics closely related to the motor function of speech. Thus, frequency-related speech features such as MFCCs respond sensitively to dose changes before extrapyramidal manifestations and begin to affect speech organs after EPSs, with a high correlation with energy-related features.

Table 5.4 Selected speech features

	Selected features
EPS = 0	- Second quartile of Energy - Q3-Q2 range of Sliding window log-energy - (Second derivative standard deviation 2, First quartile 2/3/6/11, Second quartile 2/3/6, Third quartile 2/3/6, Mean 2/3/6, 99th percentile 2/3/6, Linear regression offset 6 of MFCCs
EPS = 1	- Log-energy - (Standard deviation, 99th percentile, Second quartile, Third quartile) of Sliding window log-energy - DFA (Detrended fluctuation analysis) - Loudness (Mean, Standard deviation, First derivative standard deviation, Second derivative standard deviation, Second quartile, Third quartile, Q2-Q1 range, Q3-Q2 range, Q3-Q1 range, 99th percentile, 99th-1st percentile range, Linear regression offset, Linear regression MSE) of Energy

Further, through Figure 5.1, the speech features with the highest correlation coefficient in each EPS group can be identified. The speech feature indicating the difference between Q3 and Q1 of energy yielded a Pearson correlation coefficient of r=0.586.

Fig. 5.1 Strongest correlated speech features of each extrapyramidal symptom group.

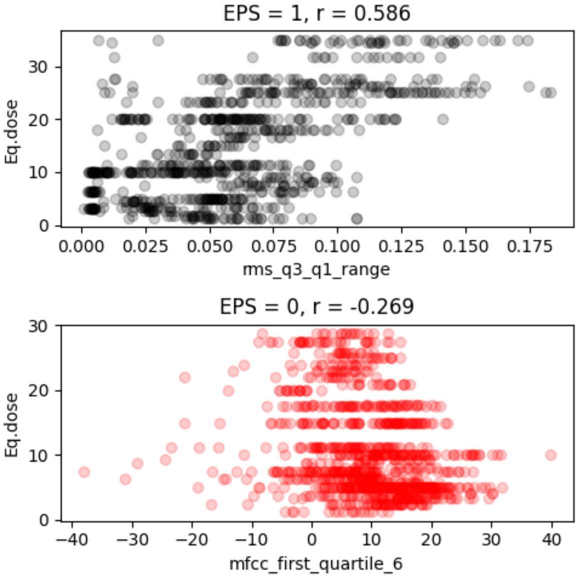
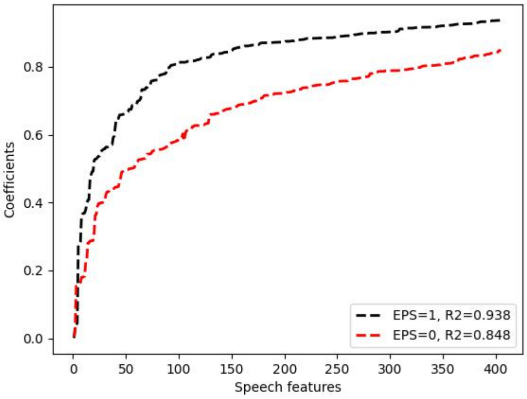


Figure 5.2 shows the correlation coefficient of the linear regression model through multiple linear regression analysis for all sentence groups. Through a 20% test group, correlation coefficients for each speech feature were obtained, and the correlation coefficient trend of the final multiple linear regression model was observed by adding them one by one in the order of speech features with high correlation coefficients. The correlation coefficient of the group with EPSs was very high at 0.938, and the group without EPSs also showed a high correlation coefficient of 0.848.

Fig. 5.2 Multivariate linear regression coefficients of each EPS group.



5.5 Discussion

Uniqueness of this study This chapter analyzes the correlation between EPSs and speech characteristics per the dose of antipsychotic drugs to directly confirm the effect of EPSs on antipsychotic drug users. It attempts to identify a rare correlation to address the change in speech characteristics that can be confirmed only by the observation of a psychiatrist and ascertain what kind of speech characteristics can be a good discriminator. The only relation to Sinha's work is that the speech method for securing speech samples is limited to a single syllabus such as "Ah" and "Uh," which is the most observable form of speech-motor control. However, relative to speech methods such as sentence reading or free speech, it is not a form that is mainly pronounced in everyday life. From Table 5.3, the correlation coefficient result of sentence "A" ("Ah" single syllabus pronunciation) was 0.473 in the group with EPSs, which was lower than 0.531 in the group that pronounced neutral sentences. Therefore, the speech collection method of sentence utterance allows for detecting EPSs more clearly than the single syllabus-type speech method. Moreover, Sinha's study was conducted only in patients who took Risperidone, one of the antipsychotic drugs. Thus, there is no evidence of the effectiveness of other drugs. This study calculated doses for a total of eight drugs, including Risperidone, Aripiprazole, and Clozapine, which are the majority of drugs used to treat antipsychotic drugs. The use of equivalent dose, a method of calculating drug dosage, is an unprecedented attempt and shows the possibility that the equivalent dose calculation method can be sufficiently applied to statistical modeling studies like this study.

Difference between EPS=0 and EPS=1 Table 5.3 shows the stark difference between the correlation coefficient of the group with EPS=1 and the group with EPS=0. The group with EPS=1 has the highest correlation of 0.531 when reading neutral sentences, and the combination analysis shows the lowest correlation of 0.244 when EPS=0 pronounced neutral sentences and "A" sounds. On average, the correlation value was nearly twice as large, showing evidence of the change in speech characteristics of EPSs. When neutral sentences are combined with "A" sounds, the correlation is somewhat reduced, which shows that the change in speech characteristics shown by the "A" sound and the neutral sentence is somewhat different.

Which speech characteristic was a good discriminator? From Table 5.4, 20 representative speech characteristics with high correlation could be examined, revealing differences between the EPS=0 group and the EPS=1 group. In the group with EPS=1, log-energy, loudness, and energy-related characteristics were mainly selected, and prosody or spectral characteristics were not observed at all. However, in the group with EPS=0, spectral-related characteristics such as MFCCs were mainly selected. It can be said to be the result of equally reflecting the causal relationship that EPSs induce a decrease in motor cortex function, inducing impairment in vocal cord speech function.

Why does the speech characteristic of EPS=0 correlate with the dose of the drug? This question is one of the most interesting parts of the results of this study. In general, EPSs are expected to appear gradually, but the basis for actually confirming them was insufficient. Moreover, patients who did not show symptoms outside of the vertebral body could sufficiently observe the progress of the speech-motor function. From the list of selected speech features, the characteristics related to spectral and pitch were the targets, and they will be crucial indicators when creating a predictive model for EPSs.

The relationship between Simpson-Angus Scale and extrapyramidal symptoms SAS is an index that measures movement involuntary ability and is one of the main indicators for determining EPSs. From Table 5.3, the group with SAS 0 or higher was highly correlated with neutral sentences (0.520) and ”A” sounds (0.501), while the group with SAS 0 showed a slightly lower correlation of 0.301 and 0.283, respectively. This result showed a very similar tendency to the presence of EPSs; it was confirmed that SAS was a very reliable method to determine EPSs. Moreover, interestingly, SAS brings the same change to the method and speech characteristics of the examination based on the physical function, such as the mobility and walking of the arms and legs.

Limitations Although this study is the first study to investigate the correlation between EPSs, antipsychotic drugs, and speech characteristics, it also has limitations. In the case of the Korean EPS (KEPS) speech corpus, it was challenging to observe the pattern of change within one patient given the difficulty of securing the population. If the observation unit was fixed as one person from the start of treatment of antipsychotic drugs in a patient to the onset of EPSs, more meaningful results could have been produced. Further, regarding the EPS prediction model through the dose of antipsychotic drugs, which is the ultimate goal to reach, it has a limitation that is challenging to model with the current speech corpus.

5.6 Summary

In this chapter, the KEPS speech corpus was constructed using the Korean-based antipsychotic dose, and the correlation with speech characteristics according to the antipsychotic dose was analyzed. The correlation coefficient of speech characteristics per antipsychotic drug administration in the EPS group was higher than that of the normal group, with a significant difference in the correlation of negative characteristics. The study showed the possibility that speech could be used as an indicator for early detection of EPSs in the future because it is related to the speech characteristics not only after the onset but also before the onset of EPSs.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis presents a new aspect to the establishment of an automatic diagnosis system for an MDD commonly called depression. Speech-based automatic depression diagnosis systems have been gradually developed by constructing a corpus of various methods, developing acoustic characteristics, and suggesting algorithms. However, they had limitations in their application as a diagnostic tool for depression given the lack of clinical assumptions and simple modeling with the depression diagnosis scale. This study tried to establish a new clinical hypothesis and suggest a new methodology for depressed speech analysis through analysis at the sentence level.

We developed a new method of depressive speech corpus and proposed sentence sets to overcome the masked depression effect (chapter 3). Accordingly, we developed a sentence-level diagnostic analysis system for elderly depression and proposed the effectiveness of the newly constructed speech corpus and a diagnostic analysis system method specialized for the elderly (chapter 4). Moreover, we investigated changes in speech characteristics that may occur during the period of taking antipsychotic drugs and changes in speech characteristics that could occur in the entire process from the prevalence of mental disorders to treatment (chapter 5). The major contributions of this thesis can be summarized as follows.

Development of mood-Inducing Sentence (MIS) that drive the speaker's emotions Depression patterns of the elderly differ significantly in characteristics from those of adults, as they are very passive in expressing emotions and have very monotonous characteristics given the degeneration of vocal organs such as the vocal cords. Thus, a strategy to maximize their emotional state was warranted. In particular, we developed MIS based on emotional words that affect the emotions of the elderly speaker, constructed a speech corpus, and proposed a new methodology for speech tasks related to emotions beyond depression.

Effects of reading emotional sentences and carryover effect Chapter 4 highlighted the psychological assumption that only reading emotional sentences could induce changes in the speaker's emotions. Based on this assumption, the change patterns of the normal and depressed groups were different. However, reading one or two emotional sentences was not sufficient to maximize the change in speech characteristics. The change was maximized when reading positive, negative, and neutral sentences alternately. It can be confirmed that a kind of "carryover effect" exists. That is, when a neutral sentence was read after reading a sentence with emotional content, especially negative content, the characteristic of maintaining the speech characteristic was found more clearly in the depression group. It could be interpreted as having a great correlation with the ability to recover emotions, which can be said to be a new attempt to overcome the masked depression effect.

Reflections of extrapyramidal symptoms to speech characteristics In Chapter 5, we observed not only depression but also changes in speech characteristics that may occur during the administration of antipsychotic drugs. An EPS speech corpus comprising neutral sentences was not developed to agitate changes in emotion. Accordingly, a correlation between speech characteristics and antipsychotic drug dose could be found in subjects who had already developed EPSs. Moreover, a high correlation was found in people who were taking antipsychotic drugs but did not develop symptoms. The corpus can be utilized in a diagnostic system for early detection of the EPS stage, which is said to be a very early stage of Parkinson's disease.

Validation of medical hypotheses through an engineering approach The original purpose of this thesis was to use an engineering approach to verify the medical hypothesis. Studies conducted up to this chapter have dealt with how symptoms experienced by depressed patients, from the prevalence of depression to treatment, are expressed in the patient's speech characteristics. However, through the various speech change characteristics of the patient based on the analyzed speech characteristics, opinions that viewed depression socially and medically were confirmed again.

Through the comparison of free speech and MIS speech characteristics, it was possible to derive an analysis supporting the socio-scientific analysis that depression in the elderly is widespread regardless of the prevalence of depression, and depression in the elderly corresponds to basic emotions. In addition, the conservative attitude toward depressive emotional expression, which is mainly found in East Asian men, was also confirmed through the analysis of MIS speech characteristics, which can be seen as evidence that social awareness of depression is needed. The correlation analysis between antipsychotic drugs and speech characteristics enabled various analyses related to the neurological approach and facts of depression.

Limitations The research of each chapter has different limitations, but the limitations from an integrated perspective based on this thesis can also be presented. Sentence sets (MIS, neutral) for driving human depression, which are the starting points of this thesis, should be considered a better collection method to compensate for the shortcomings while maintaining the advantages of the existing depressed speech corpus collection method. The problem of overcoming demand characteristics, an experimental phenomenon that creates unconscious behavioral changes that the subject interprets and aligns with the purpose of the experiment, is a problem that the methodologies of this dissertation must solve in order to move toward actual application. In addition, the fact that young people with depression can think exclusively about driving emotions and that the methodology is more suitable for patients with masked depression tendencies is also something to be solved for the generalization of this study. For modeling between antipsychotic use and speech characteristics, a more granular dose sample is needed, and the justification of sentence speech methods should be further supplemented by comparisons between different collection methods as well as sentence speech methods.

6.2 Future work

Advances in medical diagnostic tools The goal of this study is to develop a novel medical diagnostic tool. Therefore, it is essential to expand the speech corpus by considering more diverse demographic factors. Moreover, it is expected that diverse methodological attempts are needed for feature representation in sentence units. Further, regarding the sentences developed in Chapter 3, to develop for daily monitoring purposes, consideration should be given to the degree of human adaptation that can occur when reading the same sentence every day.

Extension to multimodality Speech, the main analysis biomarker in this study, is one of several biomarkers for analyzing depression. It is necessary to improve the performance of the automatic depression diagnosis tool through multimodality-integrated analysis and modeling with image, text, and electrogastrogram data. The methods proposed in this thesis can also be extended to other domains.

Additional methods for diagnosing extrapyramidal symptoms Chapter 5 furnishes an observational study to examine the correlation between EPS and the dose of antipsychotic drugs. Hence, to achieve the development of a diagnostic tool, more data collection (especially, speech data collection per one person's dose change), a model predicting the dose level, and a pre-trained network representing speech impairment construction studies are needed.

Bibliography

[1] I. Schwabe, Y. Milaneschi, Z. Gerring, P. Sullivan, E. Schulte, N. Suppli, J. Thorp, E. Derks, and C. Middeldorp, “Unraveling the genetic architecture of major depressive disorder: merits and pitfalls of the approaches used in genome-wide association studies,” *Psychological medicine*, vol. 49, no. 16, pp. 2646–2656, 2019.

[2] A. Qaseem, M. J. Barry, D. Kansagara, and C. G. C. of the American College of Physicians, “Nonpharmacologic versus pharmacologic treatment of adult patients with major depressive disorder: a clinical practice guideline from the american college of physicians,” *Annals of internal medicine*, vol. 164, no. 5, pp. 350–359, 2016.

[3] B. Geiselman and M. Bauer, “Subthreshold depression in the elderly: qualitative or quantitative distinction?” *Comprehensive Psychiatry*, vol. 41, no. 2, pp. 32–38, 2000.

[4] H. Lavretsky and A. Kumar, “Clinically significant non-major depression: old concepts, new insights,” *The American journal of geriatric psychiatry*, vol. 10, no. 3, pp. 239–255, 2002.

[5] "Symptoms - clinical depression - nhs," https://www.nhs.uk/mental-health/conditions/clinical-depression/symptoms/, (Accessed on 10/05/2022).

[6] "Depression assessment instruments," https://www.apa.org/depression-guideline/assessment, (Accessed on 10/05/2022).

[7] C. Mattiuzzi and G. Lippi, "Worldwide asthma epidemiology: insights from the global health data exchange database," in *International forum of allergy & rhinology*, vol. 10, no. 1. Wiley Online Library, 2020, pp. 75–80.

[8] D. J. Brody, L. A. Pratt, and J. P. Hughes, "Prevalence of depression among adults aged 20 and over: United states, 2013-2016," 2018.

[9] G. Sözeri-Varma, "Depression in the elderly: clinical features and risk factors," *Aging and disease*, vol. 3, no. 6, p. 465, 2012.

[10] B. D. Lebowitz, J. L. Pearson, L. S. Schneider, C. F. Reynolds, G. S. Alexopoulos, M. L. Bruce, Y. Conwell, I. R. Katz, B. S. Meyers, M. F. Morrison *et al.*, "Diagnosis and treatment of depression in late life: consensus statement update," *Jama*, vol. 278, no. 14, pp. 1186–1190, 1997.

[11] S. P. Namboodiri and D. Venkataraman, "A computer vision based image processing system for depression detection among students for counseling," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 503–512, 2019.

[12] G. Rees, E. K. Fenwick, J. E. Keeffe, D. Mellor, and E. L. Lamoureux, “Detection of depression in patients with low vision,” *Optometry and Vision Science*, vol. 86, no. 12, pp. 1328–1336, 2009.

[13] D. William and D. Suhartono, “Text-based depression detection on social media posts: A systematic literature review,” *Procedia Computer Science*, vol. 179, pp. 582–589, 2021.

[14] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, “A textual-based featuring approach for depression detection using machine learning classifiers and social media texts,” *Computers in Biology and Medicine*, vol. 135, p. 104499, 2021.

[15] U. R. Acharya, V. K. Sudarshan, H. Adeli, J. Santhosh, J. E. Koh, and A. Adeli, “Computer-aided diagnosis of depression using eeg signals,” *European neurology*, vol. 73, no. 5-6, pp. 329–336, 2015.

[16] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech communication*, vol. 71, pp. 10–49, 2015.

[17] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, “The distress analysis interview corpus of human and computer interviews,” UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, Tech. Rep., 2014.

[18] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings*

of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013, pp. 3–10.

[19] S. Grover, N. Malhotra *et al.*, “Depression in elderly: A review of indian research,” *Journal of Geriatric Mental Health*, vol. 2, no. 1, p. 4, 2015.

[20] D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[21] M. Fava, “The role of the serotonergic and noradrenergic neurotransmitter systems in the treatment of psychological and physical symptoms of depression,” *J Clin Psychiatry*, vol. 64, no. Suppl 13, pp. 26–29, 2003.

[22] Y. Liu, J. Zhao, and W. Guo, “Emotional roles of mono-aminergic neurotransmitters in major depressive disorder and anxiety disorders,” *Frontiers in psychology*, vol. 9, p. 2201, 2018.

[23] B. J. Sadock and V. A. Sadock, *Kaplan & Sadock’s concise textbook of clinical psychiatry.* Lippincott Williams & Wilkins, 2008.

[24] P. A. Noone, “The holmes–rahe stress inventory,” *Occupational Medicine*, vol. 67, no. 7, pp. 581–582, 2017.

[25] “Depression (major depressive disorder) - symptoms and causes - mayo clinic,” https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007, (Accessed on 10/05/2022).

[26] J. S. Beck, “Cognitive therapy: basics and beyond. new york,” *Guilford Press. Befera, MS & Barkley, RA (1985). Hyperactive and normal girls*

and boys: mother child interaction, parent psychiatric status and child psychopathology. Journal of Child Psychology and Psychiatry, vol. 26, no. 3, pp. 439–452, 1995.

[27] "Mental disorders," https://www.who.int/news-room/fact-sheets/detail/mental-disorders, (Accessed on 11/10/2022).

[28] Y. Li, Y. Lin, H. Ding, and C. Li, "Speech databases for mental disorders: A systematic review," *General psychiatry*, vol. 32, no. 3, 2019.

[29] "Mental disorders," https://www.who.int/news-room/fact-sheets/detail/mental-disorders, (Accessed on 11/10/2022).

[30] D. Iter, J. Yoon, and D. Jurafsky, "Automatic detection of incoherent speech for diagnosing schizophrenia," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 136–146.

[31] B. Wang, Y. Wu, N. Taylor, T. Lyons, M. Liakata, A. J. Nevado-Holgado, and K. E. Saunders, "Learning to detect bipolar disorder and borderline personality disorder with language and speech in non-clinical interviews," *arXiv preprint arXiv:2008.03408*, 2020.

[32] Z. N. Karam, E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. G. Mcinnis, "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4858–4862.

[33] Z. Pan, C. Gui, J. Zhang, J. Zhu, and D. Cui, "Detecting manic state of bipolar disorder based on support vector machine and gaussian mixture model using spontaneous speech," *Psychiatry investigation*, vol. 15, no. 7, p. 695, 2018.

[34] N. M. Etter, F. A. Cadely, M. G. Peters, C. R. Dahm, and K. A. Neely, "Speech motor control and orofacial point pressure sensation in adults with adhd," *Neuroscience letters*, vol. 744, p. 135592, 2021.

[35] L. Barona-Lleo and S. Fernandez, "Hyperfunctional voice disorder in children with attention deficit hyperactivity disorder (adhd). a phenotypic characteristic?" *Journal of Voice*, vol. 30, no. 1, pp. 114–119, 2016.

[36] "Speech emotion analysis - scholarpedia," http://www.scholarpedia.org/article/Speech_emotion_analysis, (Accessed on 10/06/2022).

[37] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[38] S. J. Garlow, J. Rosenberg, J. D. Moore, A. P. Haas, B. Koestner, H. Hendin, and C. B. Nemeroff, "Depression, desperation, and suicidal ideation in college students: results from the american foundation for suicide prevention college screening project at emory university," *Depression and anxiety*, vol. 25, no. 6, pp. 482–488, 2008.

[39] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[40] T. F. Quatieri and N. Malyska, “Vocal-source biomarkers for depression: A link to psychomotor activity,” in *Thirteenth annual conference of the international speech communication association*, 2012.

[41] K. Heser, F. Tebarth, B. Wiese, M. Eisele, H. Bickel, M. Köhler, E. Mösch, S. Weyerer, J. Werle, H.-H. König *et al.*, “Age of major depression onset, depressive symptoms, and risk for subsequent dementia: results of the german study on ageing, cognition, and dementia in primary care patients (agecode),” *Psychological medicine*, vol. 43, no. 8, pp. 1597–1610, 2013.

[42] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, “Multimodal assessment of depression from behavioral signals,” *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pp. 375–417, 2018.

[43] J. F. Greden and B. J. Carroll, “Psychomotor function in affective disorders: an overview of new monitoring techniques.” *The American journal of psychiatry*, 1981.

[44] Å. Nilsonne, J. Sundberg, S. Ternström, and A. Askenfelt, “Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression,” *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 716–728, 1988.

[45] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, “On the relative importance of vocal source, system, and

prosody in human depression," in *2013 IEEE International Conference on Body Sensor Networks.* IEEE, 2013, pp. 1–6.

[46] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker *et al.*, "From joyous to clinically depressed: Mood detection using spontaneous speech." in *FLAIRS Conference*, vol. 19. Citeseer, 2012.

[47] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of psychiatric research*, vol. 27, no. 3, pp. 309–319, 1993.

[48] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2015.

[49] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[50] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[51] K. R. Scherer, “Vocal markers of emotion: Comparing induction and acting elicitation,” *Computer Speech & Language*, vol. 27, no. 1, pp. 40–58, 2013.

[52] J. Rottenberg and A. C. Hindash, “Emerging evidence for emotion context insensitivity in depression,” *Current Opinion in Psychology*, vol. 4, pp. 1–5, 2015.

[53] A. Dehqan, R. C. Scherer, G. Dashti, A. Ansari-Moghaddam, and S. Fanaie, “The effects of aging on acoustic parameters of voice,” *Folia Phoniatrica et Logopaedica*, vol. 64, no. 6, pp. 265–270, 2012.

[54] S. A. Xue and D. Fucci, “Effects of race and sex on acoustic features of voice analysis,” *Perceptual and motor skills*, vol. 91, no. 3, pp. 951–958, 2000.

[55] N. W. Hashim, M. Wilkes, R. Salomon, J. Meggs, and D. J. France, “Evaluation of voice acoustics as predictors of clinical depression scores,” *Journal of Voice*, vol. 31, no. 2, pp. 256–e1, 2017.

[56] V. Mitra and E. Shriberg, “Effects of feature type, learning algorithm and speaking style for depression detection from speech,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2015, pp. 4774–4778.

[57] J. K. Darby and H. Hollien, “Vocal and speech patterns of depressive patients,” *Folia Phoniatrica et Logopaedica*, vol. 29, no. 4, pp. 279–291, 1977.

[58] H. Hollien, “Vocal indicators of psychological stress,” *Forensic psychology and psychiatry*, vol. 347, no. 1, pp. 47–71, 1980.

[59] H. Stassen, S. Kuny, and D. Hell, “The speech analysis approach to determining onset of improvement under antidepressants,” *European Neuropsychopharmacology*, vol. 8, no. 4, pp. 303–310, 1998.

[60] F. Tolkmitt, H. Helfrich, R. Standke, and K. R. Scherer, “Vocal indicators of psychiatric treatment effects in depressives and schizophrenics,” *Journal of communication disorders*, vol. 15, no. 3, pp. 209–222, 1982.

[61] J. D. Teasdale, S. J. Fogarty, and J. M. G. Williams, “Speech rate as a measure of short-term variation in depression,” *British Journal of Social and Clinical Psychology*, vol. 19, no. 3, pp. 271–278, 1980.

[62] A. C. Trevino, T. F. Quatieri, and N. Malyska, “Phonologically-based biomarkers for major depressive disorder,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–18, 2011.

[63] E. Keller, “The analysis of voice quality in speech processing,” *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pp. 54–73, 2004.

[64] S. Scherer, G. Stratou, and L.-P. Morency, “Audiovisual behavior descriptors for depression assessment,” in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 135–140.

[65] I. V. Bele, “The speaker’s formant,” *Journal of Voice*, vol. 20, no. 4, pp. 555–578, 2006.

[66] N. Cummins, J. Epps, and E. Ambikairajah, “Spectro-temporal analysis of speech affected by depression and psychomotor retardation,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7542–7546.

[67] M. Zimmerman, J. H. Martinez, D. Young, I. Chelminski, and K. Dalrymple, “Severity classification on the hamilton depression rating scale,” *Journal of affective disorders*, vol. 150, no. 2, pp. 384–388, 2013.

[68] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser, “Critical analysis of the impact of glottal features in the classification of clinical depression in speech,” *IEEE transactions on biomedical engineering*, vol. 55, no. 1, pp. 96–107, 2007.

[69] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, “Classification of depression state based on articulatory precision.” in *Interspeech*, 2013, pp. 2172–2176.

[70] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, “Avec 2014: 3d dimensional affect and depression recognition challenge,” in *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 2014, pp. 3–10.

[71] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang *et al.*, “Deep learning for depression recognition with audiovisual cues: A review,” *Information Fusion*, vol. 80, pp. 56–86, 2022.

[72] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.

[73] Y. Dong and X. Yang, "A hierarchical depression detection model based on vocal and emotional cues," *Neurocomputing*, vol. 441, pp. 279–290, 2021.

[74] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[75] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 6247–6251.

[76] Y. S. Özkanca, C. Demiroğlu, A. Besirli, and S. Celik, "Multi-lingual depression-level assessment from conversational speech using acoustic and text features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.* International Speech Communication Association, 2018.

[77] W. W. Zung, C. B. Richards, and M. J. Short, "Self-rating depression scale in an outpatient clinic: further validation of the sds," *Archives of general psychiatry*, vol. 13, no. 6, pp. 508–515, 1965.

[78] M. Tasnim, M. Ehghaghi, B. Diep, and J. Novikova, "Depac: a corpus for depression and anxiety detection from speech," in *Proceedings of the*

Eighth Workshop on Computational Linguistics and Clinical Psychology, 2022, pp. 1–16.

[79] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing.* John Wiley & Sons, 2013.

[80] P. Bech, "Rating scales in depression: limitations and pitfalls," *Dialogues in clinical neuroscience*, 2022.

[81] S. M. Eack, C. G. Greeno, and B.-J. Lee, "Limitations of the patient health questionnaire in identifying anxiety and depression in community mental health: many cases are undetected," *Research on social work practice*, vol. 16, no. 6, pp. 625–631, 2006.

[82] P. Shetty, A. Mane, S. Fulmali, and G. Uchit, "Understanding masked depression: A clinical scenario," *Indian journal of psychiatry*, vol. 60, no. 1, p. 97, 2018.

[83] K. Glaser, "Masked depression in children and adolescents," *American Journal of Psychotherapy*, vol. 21, no. 3, pp. 565–574, 1967.

[84] G. A. Carlson and D. P. Cantwell, "Unmasking masked depression in children and adolescents." *The American journal of psychiatry*, 1980.

[85] V. T. Lai, R. M. Willems, and P. Hagoort, "Feel between the lines: implied emotion in sentence comprehension," *Journal of Cognitive Neuroscience*, vol. 27, no. 8, pp. 1528–1541, 2015.

[86] B. Lubin and M. Zuckerman, *Manual for the MAACL-R: Multiple Affect Adjective Check List-Revised.* EdITS/Educational and Industrial Testing Service, 1999.

[87] H.-h. Lee, E.-J. Kim, and M.-k. Lee, "A validation study of korea positive and negative affect schedule: The panas scales," *Korean Journal of Clinical Psychology*, vol. 22, no. 4, pp. 935–946, 2003.

[88] B. Kim, "Compilation of the korean affective word list," *Unpublished master's thesis). University of Yonsei, Seoul, Korea*, 2010.

[89] D. Watson and A. Tellegen, "Toward a consensual structure of mood." *Psychological bulletin*, vol. 98, no. 2, p. 219, 1985.

[90] C. A. Hall, "Differential relationships of pleasure and distress with depression and anxiety over a past, present, and future time framework." Ph.D. dissertation, ProQuest Information & Learning, 1978.

[91] E. Velten Jr, "A laboratory task for induction of mood states," *Behaviour research and therapy*, vol. 6, no. 4, pp. 473–482, 1968.

[92] S. Lee, S. W. Suh, T. Kim, K. Kim, K. H. Lee, J. R. Lee, G. Han, J. W. Hong, J. W. Han, K. Lee *et al.*, "Screening major depressive disorder using vocal acoustic features in the elderly by sex," *Journal of Affective Disorders*, vol. 291, pp. 15–23, 2021.

[93] B. Lubin, R. Van Whitlock, D. Reddy, and S. Petren, "A comparison of the short and long forms of the multiple affect adjective check list—revised (maacl-r)," *Journal of clinical psychology*, vol. 57, no. 3, pp. 411–416, 2001.

[94] U. Nations, "Department of economic and social affairs," *Population Division*, 2015.

[95] DSM-IV-TR., *Diagnostic and statistical manual of mental disorders.* American Psychiatric Association, 2000.

[96] P. Pontes, A. Brasolotto, and M. Behlau, "Glottic characteristics and voice complaint in the elderly," *Journal of Voice*, vol. 19, no. 1, pp. 84–94, 2005.

[97] G. J. Lamberty and L. A. Bieliauskas, "Distinguishing between depression and dementia in the elderly: A review of neuropsychological findings," *Archives of Clinical Neuropsychology*, vol. 8, no. 2, pp. 149–170, 1993.

[98] S. F. Poissant, F. Beaudoin, J. Huang, J. Brodsky, and D. J. Lee, "Impact of cochlear implantation on speech understanding, depression, and loneliness in the elderly." *Journal of Otolaryngology–Head & Neck Surgery*, vol. 37, no. 4, 2008.

[99] C. Salzman and R. I. Shader, "Depression in the elderly. i. relationship between depression, psychologic defense mechanisms and physical illness," *Journal of the American Geriatrics Society*, vol. 26, no. 6, pp. 253–260, 1978.

[100] J. W. Han, T. H. Kim, K. P. Kwak, K. Kim, B. J. Kim, S. G. Kim, J. L. Kim, T. H. Kim, S. W. Moon, J. Y. Park *et al.*, "Overview of the korean longitudinal study on cognitive aging and dementia," *Psychiatry investigation*, vol. 15, no. 8, p. 767, 2018.

[101] S.-W. Yoo, Y.-S. Kim, J.-S. Noh, K.-S. Oh, C.-H. Kim, K. NamKoong, J.-H. Chae, G.-C. Lee, S.-I. Jeon, K.-J. Min *et al.*, "Validity of korean version of the mini-international neuropsychiatric interview," *Anxiety and mood*, vol. 2, no. 1, pp. 50–55, 2006.

[102] J. H. Lee, K. U. Lee, D. Y. Lee, K. W. Kim, J. H. Jhoo, J. H. Kim, K. H. Lee, S. Y. Kim, S. H. Han, and J. I. Woo, "Development of the korean version of the consortium to establish a registry for alzheimer's disease assessment packet (cerad-k) clinical and neuropsychological assessment batteries," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 57, no. 1, pp. P47–P53, 2002.

[103] D. Y. Lee, K. U. Lee, J. H. Lee, K. W. Kim, J. H. Jhoo, S. Y. Kim, J. C. Yoon, S. I. Woo, J. Ha, and J. I. Woo, "A normative study of the cerad neuropsychological assessment battery in the korean elderly," *Journal of the International Neuropsychological Society*, vol. 10, no. 1, pp. 72–81, 2004.

[104] G. J. Chelune, R. A. Bornstein, and A. Prifitera, "The wechsler memory scale—revised," in *Advances in psychological assessment.* Springer, 1990, pp. 65–99.

[105] T. H. Kim, Y. Huh, J. Y. Choe, J. W. Jeong, J. H. Park, S. B. Lee, J. J. Lee, J. H. Jhoo, D. Y. Lee, J. I. Woo *et al.*, "Korean version of frontal assessment battery: psychometric properties and normative data," *Dementia and geriatric cognitive disorders*, vol. 29, no. 4, pp. 363–370, 2010.

[106] J. Y. Kim, J. H. Park, J. J. Lee, Y. Huh, S. B. Lee, S. K. Han, S. W. Choi, D. Y. Lee, K. W. Kim, and J. I. Woo, "Standardization of the korean version of the geriatric depression scale: reliability, validity, and factor structure," *Psychiatry investigation*, vol. 5, no. 4, p. 232, 2008.

[107] J. Zhao, "The effects of induced positive and negative emotions on risky decision making," in *Talk presented at the 28th Annual Psychological Society of Ireland Student Congress, Maynooth, Ireland*, 2006, pp. 2018–2019.

[108] M. Brookes, "Voicebox: A speech processing toolbox for matlab. 2006," *URL http://www. ee. ic. ac. uk/... hp/staff/dmb/voicebox/voicebox. html. Available online*, 2003.

[109] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[110] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[111] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in

Proceedings of the 6th international workshop on audio/visual emotion challenge, 2016, pp. 3–10.

[112] S.-H. Cha, “Comprehensive survey on distance/similarity measures between probability density functions,” *City*, vol. 1, no. 2, p. 1, 2007.

[113] M. Cooper and J. Foote, “Automatic music summarization via similarity analysis.” in *ISMIR*. Citeseer, 2002.

[114] J. Hua, W. D. Tembe, and E. R. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[115] Y. Saeys, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[116] L. Breiman, “Arcing the edge,” Technical Report 486, Statistics Department, University of California at . . . , Tech. Rep., 1997.

[117] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[118] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[119] R. Duda, P. Hart, and D. Stork, “Pattern classification (pp. 526-528),” 1973.

[120] T. Hastie, S. Rosset, J. Zhu, and H. Zou, “Multi-class adaboost,” *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[121] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” in *Icml*, vol. 97, no. 412-420. Citeseer, 1997, p. 35.

[122] T. Inada and A. Inagaki, “Psychotropic dose equivalence in j apan,” *Psychiatry and clinical neurosciences*, vol. 69, no. 8, pp. 440–447, 2015.

[123] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[124] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, and T. Arai, “Major depressive disorder discrimination using vocal acoustic features,” *Journal of affective disorders*, vol. 225, pp. 214–220, 2018.

[125] L. S. Radloff, “The ces-d scale: A self-report depression scale for research in the general population,” *Applied psychological measurement*, vol. 1, no. 3, pp. 385–401, 1977.

[126] M. H. Sanchez, D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, and W. Jarrold, “Using prosodic and spectral features in detecting depression in elderly males,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[127] M. Hamilton, “A rating scale for depression,” *Journal of neurology, neurosurgery, and psychiatry*, vol. 23, no. 1, p. 56, 1960.

[128] K. C. Fraser, F. Rudzicz, and G. Hirst, "Detecting late-life depression in alzheimer's disease through analysis of speech and language," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 1–11.

[129] N. Cotterell, T. Buffel, and C. Phillipson, "Preventing social isolation in older people," *Maturitas*, vol. 113, pp. 80–84, 2018.

[130] K. U. Rani and M. S. Holi, "Gmm classifier for identification of neurological disordered voices using mfcc features," *IOSR Journal of VLSI and Signal Processing*, vol. 4, pp. 44–51, 2015.

[131] W. N. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentation," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 6, pp. 1884–1892, 2007.

[132] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age," 2011.

[133] T. J. Shors and B. Leuner, "Estrogen-mediated effects on depression and memory formation in females," *Journal of affective disorders*, vol. 74, no. 1, pp. 85–96, 2003.

[134] K. E. Hill, S. C. South, R. P. Egan, and D. Foti, "Abnormal emotional reactivity in depression: Contrasting theoretical models using neurophysiological data," *Biological psychology*, vol. 141, pp. 35–43, 2019.

[135] M. T. Orne, “On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications,” in *Sociological Methods.* Routledge, 2017, pp. 279–299.

[136] P. M. Kenealy, “The velten mood induction procedure: A methodological review,” *Motivation and emotion*, vol. 10, no. 4, pp. 315–335, 1986.

[137] P. M. Niedenthal and J. B. H. . M. B. Setterlund, “Being happy and seeing”happy”: Emotional state mediates visual word recognition,” *Cognition & Emotion*, vol. 11, no. 4, pp. 403–432, 1997.

[138] J. J. Gross and R. W. Levenson, “Emotion elicitation using films,” *Cognition & emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[139] R. C. Baker and D. O. Guttfreund, “The effects of written autobiographical recollection induction procedures on mood,” *Journal of Clinical Psychology*, vol. 49, no. 4, pp. 563–568, 1993.

[140] J. M. Pierre, “Extrapyramidal symptoms with atypical antipsychotics,” *Drug safety*, vol. 28, no. 3, pp. 191–208, 2005.

[141] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, “Advances in parkinson’s disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects,” *Biomedical Signal Processing and Control*, vol. 66, p. 102418, 2021.

[142] L. Jeancolas, D. Petrovska-Delacrétaz, G. Mangone, B.-E. Benkelfat, J.-C. Corvol, M. Vidailhet, S. Lehéricy, and H. Benali, “X-vectors: New quan-

titative biomarkers for early parkinson's disease detection from speech," *Frontiers in Neuroinformatics*, vol. 15, p. 578369, 2021.

[143] B. K. Varghese, D. Amali, and K. Devi, "Prediction of parkinson's disease using machine learning techniques on speech dataset," *Research Journal of Pharmacy and Technology*, vol. 12, no. 2, pp. 644–648, 2019.

[144] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth, "Deep learning approach to parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 717–720.

[145] G. M. Gharabawi, C. A. Bossie, R. A. Lasser, I. Turkoz, S. Rodriguez, and G. Chouinard, "Abnormal involuntary movement scale (aims) and extrapyramidal symptom rating scale (esrs): cross-scale comparison in assessing tardive dyskinesia," *Schizophrenia research*, vol. 77, no. 2-3, pp. 119–128, 2005.

[146] S. Janno, M. M. Holi, K. Tuisku, and K. Wahlbeck, "Validity of simpson-angus scale (sas) in a naturalistic schizophrenia population," *BMC neurology*, vol. 5, no. 1, pp. 1–6, 2005.

[147] ——, "Actometry and barnes akathisia rating scale in neuroleptic-induced akathisia," *European neuropsychopharmacology*, vol. 15, no. 1, pp. 39–41, 2005.

[148] P. Sinha, V. P. Vandana, N. V. Lewis, M. Jayaram, and P. Enderby, "Predictors of effect of atypical antipsychotics on speech," *Indian Journal of Psychological Medicine*, vol. 37, no. 4, pp. 429–433, 2015.

[149] P. Sinha, V. Vandana, N. V. Lewis, M. Jayaram, and P. Enderby, "Evaluating the effect of risperidone on speech: A cross-sectional study," *Asian Journal of Psychiatry*, vol. 15, pp. 51–55, 2015.

[150] S. Leucht, M. Samara, S. Heres, and J. M. Davis, "Dose equivalents for antipsychotic drugs: the ddd method," *Schizophrenia bulletin*, vol. 42, no. suppl_1, pp. S90–S94, 2016.

[151] P. H. Rothe, S. Heres, and S. Leucht, "Dose equivalents for second generation long-acting injectable antipsychotics: The minimum effective dose method," *Schizophrenia research*, vol. 193, pp. 23–28, 2018.

[152] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[153] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125 868–125 881, 2019.

[154] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Convolutional-recurrent neural network for age and gender prediction from speech," in *2019 Signal Processing Symposium (SPSympo)*. IEEE, 2019, pp. 242–245.

[155] R. Lenain, J. Weston, A. Shivkumar, and E. Fristed, "Surfboard: Audio feature extraction for modern machine learning," *arXiv preprint arXiv:2005.08848*, 2020.

[156] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[157] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

초 록

주요 우울 장애 즉 흔히 우울증이라고 일컬어지는 기분 장애는 전 세계인 중 3.8%에 달하는 사람들이 겪은바 있는 매우 흔한 질병이다. 유전, 노화, 사회적 요인, 신경전달물질 체계의 이상등 다양한 원인으로 발생하는 우울증은 조기 발견 및 일상생활에서의 관리가 매우 중요하다고 할 수 있다. 인간의 음성은 우울증을 관찰하기에 대표적인 바이오마커로 여겨져 왔으며, 음성 데이터를 기반으로한 자동 우울증 진단 시스템 개발을 위한 여러 연구들이 진행되어 왔다. 그러나 음성 말뭉치 구축의 어려움과 60세 이하의 성인들에게 초점이 맞추어진 연구, 정신과 의사들의 임상 소견을 바탕으로한 의학적 가설 설정의 미흡등의 한계점을 가지고 있으며, 이는 의료 진단 기구로 발전하는데 한계점이라고 할 수 있다. 또한, 항정신성 약물의 복용이 음성 특징에 미칠 수 있는 영향 또한 간과되고 있다.

본 논문에서는 위의 한계점들을 보완하기 위한 의미론적 수준 (문장 단위)에서의 음성 기반 자동 우울증 진단에 대한 연구를 시행하고자 한다. 우선적으로 감정의 변화가 음성 특징을 잘 반영되지 않는 노인층의 우울증 분석을 위해 감정 발화 문장을 개발하여 노인 우울증 음성 말뭉치를 구축하고, 문장 단위에서의 관찰을 통해 노인 우울증 군에서 감정 문장 발화가 미치는 영향과 감정 전이를 확인할 수 있었으며, 노인층의 자동 우울증 진단 시스템을 설계하였다. 최종적으로 항정신병 약물의 과복용으로 나타날 수 있는 대표적인 부작용인 추체외로 증상을 조사하기 위해 추체외로 증상 음성 말뭉치를 구축하였고, 항정신병 약물의 복용량과 음성 특징간의 상관관계를 분석하여 우울증의 치료 과정에서 항정신병 약물이 음성에 미칠 수 있는 영향에 대해서 조사하였다. 이를 통해 주요 우울 장애의 영역에 대한 포괄적인 연구를 진행하였다.

주요어: 주요 우울 장애, 자동 우울 진단 시스템 , 우울증 음성 말뭉치, 감정 발화 문장, 음성 특징, 추체외로 증상

학　번: 2013-22411