UNIVERSITY OF CALIFORNIA

Los Angeles

ACCURACY AND PRIVACY IN SPEECH-BASED MODELING OF MAJOR DEPRESSION: INNOVATIVE APPROACHES THROUGH DATA AUGMENTATION, AND SPEAKER IDENTITY DISENTANGLEMENT

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Vijay Ravi

2024

ABSTRACT OF THE DISSERTATION

ACCURACY AND PRIVACY IN SPEECH-BASED MODELING OF MAJOR DEPRESSION: INNOVATIVE APPROACHES THROUGH DATA AUGMENTATION, AND SPEAKER IDENTITY DISENTANGLEMENT

by

Vijay Ravi

Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2024
Professor Abeer Alwan, Chair

Major Depressive Disorder (MDD) is a prevalent mental illness that affects a significant portion of the global population. Despite its severity, traditional diagnostic methods often fail to identify and treat MDD effectively, highlighting the need for automated diagnostic tools. Recent research has identified speech signals as promising biomarkers for objectively detecting depression. However, the development of speech-based depression detection systems faces several challenges including data scarcity and privacy preservation. The sensitive nature of mental health data makes it difficult to collect large datasets required for training robust models. Moreover, many current approaches rely on features that can compromise patient confidentiality, hindering the adoption of these systems in clinical settings. This thesis presents novel methods to address these challenges and to enhance the performance and privacy of speech-based depression detection. The contributions include a frame rate-based data augmentation technique (FrAUG) to increase training data while preserving depression-related acoustic information. Additionally, five speaker identity disentanglement methods are proposed: adversarial loss maximization, loss equalization via Cross-Entropy, Variance, and KL Divergence, and unsupervised speaker disentanglement via cosine similarity minimization.

These methods aim to reduce the reliance on speaker identity during depression detection. The proposed techniques are evaluated on multiple datasets in two languages - English (DAIC-WoZ dataset) and Mandarin (EATD and CONVERGE datasets), demonstrating improved depression detection accuracy and reduced speaker separability compared to state-of-the-art approaches. Furthermore, the privacy preservation capabilities of these methods are quantified using gain of voice distinctiveness and de-identification scores, showcasing their potential for safeguarding patient privacy. By advancing speech-based depression detection in terms of accuracy and privacy, this thesis aims to facilitate the development of effective and secure diagnostic tools that can be readily adopted in clinical settings.

The dissertation of Vijay Ravi is approved.

Jonathan Flint

Gregory Pottie

Vwani Roychowdhury

Abeer Alwan, Committee Chair

University of California, Los Angeles

2024

Dedicated to my family

whose unwavering love and support made this possible.

Contents

List of Figures

List of Tables

Acknowledgements

I am deeply indebted to many people whose support made this dissertation possible. First and foremost, I extend my heartfelt gratitude to my doctoral advisor, Professor Abeer Alwan. Her exceptional mentorship, unwavering support, and constant inspiration have been invaluable. Under her guidance, I learned to tackle challenging problems, think critically about my work, and develop into a more capable researcher.

I am profoundly grateful to Professor Jonathan Flint for his advice and guidance. His insights were instrumental in shaping my research on speech-based depression detection. My sincere thanks also go to my Ph.D. committee members, Professor Vwani Roychowdhury and Professor Greg Pottie, whose wisdom and knowledge have greatly enriched my years at UCLA.

I have been fortunate to have outstanding mentors during my internships: Dr. Pierre Lanchantin, Dr. Sandesh Aryal, Dr. Hamid Mohammadi, Dr. Jing Liu, Dr. Athanasios Mouchtaris, Dr. Yile Gu, Ankur Gandhe, Dr. Kazuhito Koishida, Dr. Dung Tran, Professor Mirco Ravanelli, and Professor Samira Abbasgholizadeh Rahimi. Their guidance has significantly enhanced my research skills.

I owe a debt of gratitude to my undergraduate mentors, Professor Mohanasankar Sivaprakasam and Professor Ashish Sahani, for encouraging my initial foray into research and helping me find my bearings. I also thank Dr. Lakshmi Krishnan for her valuable guidance during my undergraduate internship.

My appreciation extends to my exceptional labmates at SPAPL. Our numerous engaging and thought-provoking conversations have kept me abreast of the latest research and provided inspiration when I faced challenges. Special thanks to Jom, Jinxi, Soo, Amber, Kaan and Gary for their guidance in my early days, and to AJ, Ruchao, Yunzheng, Huanhua, Ben, Balaji, Vishwas and Eray for their camaraderie. I am particularly grateful to Jinhan, whose collaborations were crucial to this work.

I sincerely thank the administrative staff of the ECE department: Deeona Columbia, Ryo Arreola, Julio Romero, Ylena Requena, Jose Cano, Vanessa Ramirez, Yadira Trejo, and Lorena Rodriguez for their generous assistance throughout my time at UCLA.

I am incredibly fortunate to have friends who made Los Angeles feel like a second home, celebrating my achievements and supporting me through challenges. Their constant encouragement and willingness to lend an ear during difficult times have been invaluable.

My deepest gratitude goes to my loving mother, Suma Ravi, and my father, N. V. Ravi, who have shaped me into the person I am today. Finally, I thank my sister, Ramyashree

Ravi, my greatest supporter. Her unwavering belief in my abilities has been a source of strength, especially during moments of frustration and self-doubt. I am eternally grateful to my family for their love, encouragement, and support through both the trials and joys of graduate school.

My Ph.D. journey would not have been the same without all these remarkable individuals. Thank you all for being there through every step of this incredible experience.

Vita

Education

Visvesvaraya Technological University, India Aug 2011 - May 2015

Bachelors in Engineering, Electronics & Communication Engineering

University of California Los Angeles Sep 2017 - June 2019

Masters in Science, Electrical & Computer Engineering

University of California Los Angeles Sep 2019 - Dec 2024

Doctor of Philosophy, Electrical & Computer Engineering

Relevant Work Experience

Applied Scientist Intern, Amazon	June - September 2019
Applied Scientist Intern, Amazon	June - September 2020
Research Intern, Microsoft Research	June - September 2022
Research Intern, MILA	January - March 2024

Relevant Publications

1. Ravi V., Wang J., Flint J., Alwan A. A Privacy-Preserving Unsupervised Speaker Disentanglement Method for Depression Detection from Speech. CEUR Workshop Proc. 2024 Feb;3649:57-63. PMID: 38650610; PMCID: PMC11034881.

2. Ravi, V., Wang, J., Flint, J., & Alwan, A. (2024). Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. Computer Speech & Language, 86, 101605.

3. Ravi, V., Wang, J., Flint, J., & Alwan, A. (2022, September). A step towards preserving

speakers' identity while detecting depression via speaker disentanglement. In Interspeech (Vol. 2022, p. 3338). NIH Public Access.

4. Ravi, V., Wang, J., Flint, J., & Alwan, A. (2022, May). Fraug: A frame rate based data augmentation method for depression detection from speech signals. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6267-6271). IEEE.

5. Ravi, V., Park, S. J., Afshan, A., & Alwan, A. (2019). Voice quality and between-frame entropy for sleepiness estimation. Interspeech 2019.

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

7. Wang, J., Ravi, V., & Alwan, A. (2023, August). Non-uniform speaker disentanglement for depression detection from raw speech signals. In Interspeech (Vol. 2023, p. 2343). NIH Public Access.

8. Wang, J., Ravi, V., Flint, J., & Alwan, A. (2024). Speechformer-CTC: Sequential modeling of depression detection with speech temporal classification. Speech communication, 163, 103106.

9. Di, Y., Mefford, J., Rahmani, E., Wang, J., Ravi, V., Gorla, A., Alwan, A., Zhu, T., & Flint, J. (2024). Genetic association analysis of human median voice pitch identifies a common locus for tonal and non-tonal languages. Communications Biology, 7(1), 540.

Chapter 1

Introduction

1.1 Speech-based Depression Detection

Major Depressive Disorder (MDD) is a severe mental health condition that negatively impacts an individual's emotions, thoughts, and behaviors. In extreme cases, MDD can lead to suicide. Globally, MDD affects more than 264 million people [47] and is projected to become the second leading cause of disability by 2030 [63]. The effects of mental health issues like MDD extend beyond significant economic and healthcare costs; they also have profound negative consequences for the affected individuals, their loved ones, and the wider community. The burden of MDD is further compounded by the fact that many individuals suffering from the disorder do not receive adequate treatment due to societal stigma, lack of access to mental health services, or insufficient awareness about the condition.

The diagnosis of MDD currently relies on subjective interviews conducted by psychologists and self-reported surveys [50]. However, this approach can be influenced by several factors, such as the availability of mental health professionals, patients' willingness to disclose their symptoms, and the societal stigma surrounding mental health treatment [39]. These barriers can lead to delayed diagnosis or under-reporting of symptoms, hindering timely intervention.

Additionally, MDD is currently diagnosed based on a set of symptoms and this approach leaves open the possibility that MDD may actually consist of several distinct conditions, each with its own unique causes and optimal treatment approaches. Accumulating evidence suggests that MDD is not a single, homogeneous disorder, but rather a heterogeneous collection of related conditions [49]. This heterogeneity in the presentation and underlying causes of MDD poses challenges in developing targeted treatments and predicting treatment outcomes. Early intervention through automatic evaluations for individuals showing initial signs and symptoms of MDD can therefore be crucial in preventing the condition from progressing to more severe stages. By identifying and treating MDD early, it may be possible to alleviate some of the worst consequences associated with the disorder, including suicidal thoughts and behaviors.

Such objective assessment techniques must be secure, efficient, accessible, and scalable so as to ensure widespread adoption and effective early detection of MDD across diverse populations. These techniques should prioritize patient privacy, utilize resources optimally, be capable of adapting to the varying needs of different communities and demographics, and be easily integrated into existing healthcare systems. By meeting these criteria, objective assessment methods can play a critical role in identifying individuals at risk for MDD and in facilitating timely intervention, ultimately improving patient outcomes and reducing the overall burden of the disorder on society.

Automatic objective screening mechanisms based on Electroencephalogram (EEG) and Magnetic Resonance Imaging (MRI) have been used to automatically predict mental health states in the past [1, 56, 62]. These methods have proven to be highly effective in clinical settings. However, these techniques are often impractical for widespread use due to several limitations. First, they require specialized equipment that is expensive to acquire and maintain, making them inaccessible to many healthcare providers and patients. Second, conducting these tests is logistically challenging, as they often require dedicated facilities and appointments, which can be inconvenient for patients and limit the number of individuals who can be assessed. Finally, interpreting the results of these tests requires highly trained personnel, such as radiologists or neurologists, who may not be readily available in all healthcare settings. These factors combined make it difficult to scale up the use of EEG and MRI for the early detection and diagnosis of MDD in large populations.

Among other methods, the human voice has emerged as a promising biomarker for mental health. Speech, as an information-rich data source, has been shown to effectively capture the mental [19, 82] and emotional states [72, 77] of the human mind. Some characteristics speech, such as tone, pitch, rhythm, and rate, can provide significant insights into a person's mental state. Moreover, speech data can be collected and analyzed non-invasively, without the need for expert supervision, making it a practical and efficient alternative. Unlike EEGs and MRIs, which require a controlled environment and specialized equipment, speech can be recorded using readily available devices like smartphones or voice-assistants (eg: Siri, Alexa [81]) and analyzed using advanced machine learning algorithms. This accessibility allows for continuous monitoring and early detection of mental health issues in a more naturalistic setting.

By extracting representations from speech data, a model can be trained to predict the prevalence of mental health disorders (see Figure 1.1). Advanced algorithms can analyze various features of speech, such as prosody, articulation, and fluency, to detect subtle changes that may be indicative of mental health conditions. These models can then be integrated into mobile applications or telehealth platforms, providing users with real-time feedback and facilitating timely interventions.

1.2 Current Challenges

Although speech-based automatic objective screening mechanisms for MDD have gained popularity in recent years [10, 71, 91], several challenges remain unresolved. Among others,

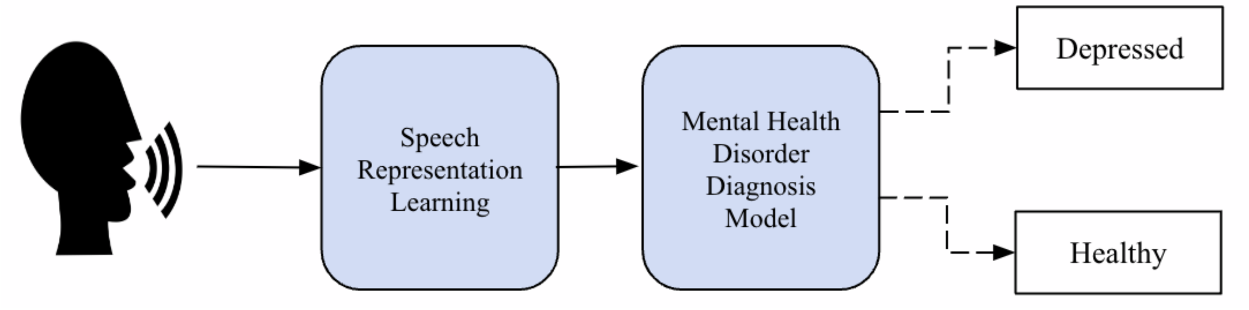

Speech Representation Learning
Mental Health Disorder Diagnosis Model
Depressed
Healthy

Figure 1.1: Schematic representation of the speech-based automatic assessment system for MDD diagnosis. The system utilizes speech representation learning techniques to extract meaningful features from speech data. These features are then fed into a mental health disorder diagnosis model, which classifies the input speech as belonging to either individuals with depression or healthy individuals.

data scarcity and privacy preservation are two critical issues that hinder the development and deployment of these systems.

Data scarcity limits the ability to train robust machine learning models, as high-quality, annotated datasets are often difficult to obtain. Without sufficient data, the models may fail to capture the variability in speech patterns across different populations, leading to reduced effectiveness and potential biases in mental health assessments.

Privacy preservation is a major concern because speech data contains sensitive personal information that must be protected to maintain user trust and comply with data protection regulations. Safeguarding the highly personal and confidential mental health information of patients is crucial to prevent harm such as discrimination, stigma, or social exclusion. Moreover, breaches of privacy could lead to misuse of data, further exacerbating these issues. Additionally, individuals may hesitate to seek mental health care if they feel their information is not secure, which can hinder the adoption of objective screening systems and result in untreated conditions and negative health outcomes.

1.2.1 Data Scarcity in Speech-based Depression Detection

In the realm of speech-based depression detection, data scarcity presents a significant obstacle, particularly highlighted by the limitations of publicly available datasets such as DAIC-WoZ [104] and EATD [96] and the proprietary dataset, CONVERGE [55]. The DAIC-WoZ (Distress Analysis Interview Corpus-Wizard of Oz) dataset, while valuable, contains a relatively small number of speakers (142 speakers), with an even smaller subset diagnosed with depression (42 speakers). This limited representation hinders the development of robust machine learning models capable of generalizing across diverse populations. Similarly, the EATD (Extended Audio-Visual Depression Corpus) faces the same challenge, offering a restricted number of recordings from depressed individuals (162 speakers in total, 30 speakers with depression). The CONVERGE dataset (from the China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology study), however, presents an interesting case, providing a large number of recordings from depressed speakers (4217 non-depressed speakers and 3742 depressed speakers). Despite this, a noteworthy limitation is that the dataset comprises solely female participants in order to focus on a more genetically homogeneous sample. While datasets like CONVERGE provide valuable insights, the absence of gender diversity and the overall scarcity of data underscore the urgent need to adopt techniques to maximize the utilization of available dataset resources. One such technique is data augmentation, which involves artificially expanding the training data by generating transformations of the original samples. Augmentation methods enhance dataset diversity and richness by applying operations such as pitch or formant shifting, speaking rate changes [3], or adding noise to existing data, thereby boosting the robustness and generalization ability of machine learning models. However, in the context of depression detection, data augmentation isn't as straightforward due to the counterproductive nature of manipulating acoustic cues such as pitch or speed, which are correlates of depression [19].

1.2.2 Privacy-preservation in Speech-based Depression Detection

Speech-based methods for depression detection often rely on speaker-related information, also known as speaker-identity features [25, 28, 30, 57, 84]. However, the use of these features raises privacy concerns, as they can be used to uniquely identify individuals through automatic speaker identification (SID) [99] and verification models [80]. One specific privacy threat is the membership inference attack [44, 97], in which malicious actors could compromise a patient's privacy by determining whether their data was used to train the model. The potential misuse of speaker-identity features presents a significant challenge in developing speech-based depression detection systems. While these features have proven to be effective in capturing depression-related cues, their ability to uniquely identify individuals poses a risk to patient privacy.

To mitigate these concerns, it's essential for models in speech-based depression detection to prioritize safeguarding individuals' privacy. Rather than depending on speaker-specific details, the emphasis should be on capturing broader patterns that differentiate between depressed and non-depressed groups. By prioritizing the extraction of non-identifying features, these models can foster a more privacy-oriented approach within speech-based mental health research.

While the above-mentioned privacy concerns are significant, an excessive reliance on individual speaker characteristics in depression detection models can introduce dataset biases, ultimately impairing their modeling capability. These biases may manifest in overfitting to speakers in the training set, resulting in inaccurate diagnoses for unseen speakers. This raises important questions about whether depression detection can be performed in a manner invariant to speaker identity and whether certain components of speech characterize a speaker without relevance to their mental health status. However, these critical questions remain largely unexplored within the speech research community.

1.3 Literature Review

The widespread adoption of digital voice assistants has simplified the process of collecting speech data, leading to a surge in research and development efforts aimed at creating objective speech-based screening systems for Major Depressive Disorder (MDD) [58, 89, 104]. This increased accessibility to speech data has opened up new possibilities for early detection and monitoring of mental health disorders.

Early research in this domain focused on understanding the impact of MDD on human speech patterns. Seminal studies, such as those conducted by Nilsonne et al. [67] and Andreasen [5], revealed that MDD is characterized by distinct verbal cues, including monotonic speech, specific vocabulary choices, and abnormal disfluencies. These findings laid the foundation for further exploration of speech-related features in the context of depression.

In recent years, researchers have built upon these initial findings and have identified clear differences in the acoustic features of speech between individuals with and without depression [19, 35]. These studies have leveraged advanced signal processing techniques and machine learning algorithms to extract and analyze a wide range of acoustic parameters, such as pitch, formants, and spectral characteristics. The results have consistently shown that depressed individuals exhibit unique patterns in their speech, such as reduced pitch variability, slower speaking rate, and altered voice quality.

1.3.1 Acoustic Features

Researchers have investigated a wide array of acoustic features for speech-based depression detection. One study [93] utilized statistics of spectral features, such as spectral tilt and formant frequencies, in combination with pitch and energy, to predict depression. Another study [116] focused on vocal prosody features, including switching pauses and pitch, for estimating depression severity. A third study [4] demonstrated that jitter, shimmer, energy, and loudness features were robust indicators of depression in both read and spontaneous speech. In [22, 23], pitch features were used to identify a common locus for tonal and non-tonal languages in the context of depression detection through a multisite genetic study. While frame-level features have been commonly used, several studies have proposed alternative representations. Some researchers [18, 21, 78] introduced the use of fixed-length i-vectors for depression detection, drawing inspiration from speaker identification research [37]. Another study [2] extended the i-vector representation to encompass voice quality features and performed a score-level fusion with Opensmile features [31] achieving an impressive F1-classification score of 94.9% using the pilot portion of the CONVERGE dataset. A comparative study [24] investigated the effectiveness of voice source-related features, such as linear prediction residual signals, homomorphically filtered voice source signals, and zero frequency filtered signals, against vocal tract-related high-frequency features. They found that voice source-related features outperformed vocal tract-related features in detecting depression. More recently, a novel approach [95] proposed the use of articulatory features obtained through acoustic inversion for depression detection.

1.3.2 Model Architectures

In addition to exploring various acoustic features, researchers have also focused on improving the backend model architectures for depression detection. Early studies employed traditional machine learning methods, such as Support Vector Machines (SVM) [92], Gaussian Mixture Models [100], and Random Forest classifiers [66]. However, in recent years, deep learning approaches have gained prominence due to their superior performance compared to conventional pattern recognition techniques [17, 41, 61, 88, 96, 110].

Among the deep learning architectures, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, have been extensively applied in depression detection tasks. For instance, a study proposed DepAudioNet, a CNN-LSTM framework that utilized mel-spectrogram features for depression detection [61]. The DepAudioNet model has been used as a baseline in many prior studies [8, 34], including this thesis, and achieved an F1-Score of 0.52 for depression classification using the DAIC-WoZ dataset. Another approach combined Mel Frequency Cepstral Coefficients (MFCCs) with a pre-trained RNN model, which was initially trained on a Speech Emotion Recognition (SER) task, to enhance depression prediction performance [88]. The effectiveness of an encoder-decoder structure, where the encoder was pre-trained on Automatic Speech Recognition (ASR) and subsequently fine-tuned for depression detection, was also investigated [41].

More recently, researchers have proposed innovative techniques to further improve the performance of deep learning models. One study introduced a method that aggregated mel-spectrograms using a NetVLAD network [6] to generate fixed-length segment-level embeddings, which were then used to train a GRU model for depression classification [96]. This setup achieved an F1-Score of 0.66 on the EATD dataset. Another study employed an Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Network (ECAPA-TDNN) model with MFCC features for depression detection [107]. Moreover, a novel self-supervised learning mechanism called instance-discrimination learning was specifically designed for depression detection tasks [109].

1.3.3 Data Augmentation

Data augmentation techniques have been widely explored in speech processing to improve the robustness and generalization of models, particularly when dealing with limited training data. One popular method is Vocal Tract Length Perturbation (VTLP) [46], which simulates variations in speaker characteristics by modifying the frequency spectrum of speech. This technique has been shown to enhance the performance of automatic speech recognition systems by generating diverse training samples. Another approach, proposed in [42], involves rotating spectrograms to create new samples, effectively increasing the amount of training data. Similarly, [87] employed a combination of noise addition, pitch-shifting, and speed perturbation to augment the training data, resulting in improved robustness of speech emotion recognition models. In [70], a multi-window data augmentation technique was introduced specifically for emotion recognition tasks. This method leverages multiple frame-widths to capture varying temporal contexts, enabling the model to learn more comprehensive representations of emotional cues in speech. These data augmentation techniques have demonstrated significant potential in enhancing the performance and generalization of speech processing models, particularly in scenarios where labeled data is scarce.

1.3.4 Speaker-Identity and Depression Detection

Numerous prior studies have investigated the utilization of speaker-related features in depression detection, particularly concerning speaker identity. Acoustic features like x-vectors [30,84] and other speaker embeddings [25,28] have demonstrated effectiveness in diagnosing a speaker's mental state. Nevertheless, these features inherently encode information about the speaker's identity [99], which poses challenges to privacy preservation, a pivotal aspect in the acceptance of digital mental health screening systems [59].

1.3.5 Privacy Preserving Speech Processing

In previous research, adversarial speaker normalization has been assessed within the domain of Speech Emotion Recognition (SER) [38,52,117]. In [117], speaker-invariant domain adaptation was conducted on multi-modal features, including speech, text, and video. Another study [52] proposed a technique employing gradient reversal with entropy loss to disentangle emotion and speaker information. Furthermore, [38] fine-tuned a pre-trained Hubert model using gradient-based adversarial learning. However, fine-tuning such models often necessitates large amounts of in-domain data and is computationally intensive. Moreover, these investigations utilized the IEMOCAP and MSP-Improv datasets, which are mono-lingual and comprise acted audio data [13, 14].

Given the critical importance of privacy preservation in speech-based depression detection, numerous studies have endeavored to tackle this challenge. Approaches such as federated learning [11] and sine wave speech [29] have been explored to protect patient identity; however, these methods frequently result in a performance decline in depression detection. More recently, adversarial learning (ADV) has shown an improvement in depression detection performance at the expense of reduced speaker classification accuracy [83,85]. These techniques are discussed in Chapters 5 and 6. In the study [108], non-uniform adversarial weights (NUSD) were identified as superior to vanilla adversarial methods when applied to raw audio signals. Additionally, [119] demonstrated the effectiveness of incorporating reconstruction loss alongside an autoencoder for achieving speaker disentanglement, leading to enhanced depression detection performance.

1.4 Technical Contribution

In this thesis, we address the problems of data scarcity and privacy preservation by introducing two frameworks: 1) Frame-rate based data augmentation and 2) speaker-disentanglement for depression detection.

The first framework focuses on data augmentation to alleviate the issue of limited training data. We propose a novel method that creates new feature samples by varying the frame-width and frame-shift during the feature extraction process. By adjusting these frame-rate parameters, the model is exposed to different sets of time-frequency resolutions during training. This approach ensures that the acoustic parameters thought to correlate with the speaker's mental state (e.g., pitch, formant frequencies, speaking rate) [2, 19] are preserved and not inadvertently modified. The proposed method is evaluated on two different datasets, input acoustic features, and models, and is shown to outperform two commonly used data augmentation techniques. This demonstrates the effectiveness and versatility of our frame-rate based data augmentation approach in improving the performance of depression detection models, particularly when faced with limited training data.

The second framework addresses the privacy concerns associated with using speaker-identity features in speech-based depression detection. We introduce five novel speaker-disentanglement method that aim to separate depression-related cues from speaker-specific characteristics. By minimizing the presence of speaker-identity information in the extracted features, we reduce the risk of individuals being uniquely identified through their speech data. This is achieved through the use of adversarial learning , loss equalization and unsupervised cosine similarity minimization techniques, where the model is trained to generate features that are discriminative for depression detection but uninformative for speaker identification. The effectiveness of our speaker-disentanglement approach is validated through experiments on multiple datasets, demonstrating its ability to maintain depression detection performance while preserving speaker privacy.

1.5 Thesis Organization

The thesis is organized into eight chapters, each focusing on a specific aspect of the research. Chapter 2 describes the databases used in the study and the feature extraction process. Chapter 3 discusses the models employed, the evaluation metrics used to assess their performance, and the training scheme adopted. Chapter 4 delves into the details of the frame-rate based data augmentation technique, presenting the methodology, experiments, and results. Chapter 5 introduces the adversarial speaker disentanglement approach, which aims to separate depression-related cues from speaker-specific characteristics to protect patient privacy.

Chapter 6 presents loss equalization based speaker disentanglement methods, comparing its performance with the adversarial approach. Chapter 7 explores an unsupervised speaker disentanglement technique, which aims to extract depression-related features without relying on speaker-labels for the training data. Finally, Chapter 8 concludes the thesis, summarizing the key findings, contributions, and potential future directions for research in speech-based depression detection with a focus on data augmentation and privacy preservation.

Chapter 2

Databases, and Features

In this chapter, we present the details of the databases and feature extraction methods used in our work. We begin with a description of the characteristics of each database, including the number of speakers, utterances, sampling rate, and any relevant metadata. Subsequently, we discuss the specific acoustic features and the corresponding extraction techniques used with these databases to evaluate the proposed methods.

2.1 Datasets

Experiments were conducted using two publicly available depression-related datasets - the DAIC-WOZ [104], and EATD [96] and one proprietary dataset - CONVERGE [55]. The datasets are described in the following subsections and the datasets' details are summarized in Table 2.1

2.1.1 DAIC-WOZ

The Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) [104] database comprises audio-visual interviews of 189 participants, male and female, who underwent evaluation of

Table 2.1: Summary of depression datasets used in this study. Cases refers to ‘depressed’ class and Controls is ‘non-depressed’ class.

	DAIC-WOZ	EATD	CONVERGE
Language	English	Mandarin	Mandarin
Number of Participants	142	162	7959
Gender	M&F	M&F	F
Cases/Controls	42/100	30/132	3742/4217
Train/Evaluation	105/37	83/79	
Sampling Rate (Hz)	16000	16000	16000
Total Duration (Hours)	22.56	2.26	391
Publicly Available	✓	✓	X

psychological distress. Each participant was assigned a self-assessed depression score through the patient health questionnaire (PHQ-8) method [50]. Audio data belonging only to the participants were extracted using the time-labels provided with the dataset. Recordings from session numbers 318, 321, 341 and 362 were excluded from the training set because of time-labelling errors. The test partition of the database was not used in this thesis because of the unavailability of the ground-truth labels resulting in a total of 142 speakers. The resulting dataset consists of a total of 22.56 hours of audio data. In line with prior research [9, 24, 34, 61, 111], methods using the DAIC-WOZ dataset are trained using data from 105 speakers and evaluated using the data from 37 speakers from the development split of the dataset.

2.1.2 EATD

The EATD Corpus (Emotional Audio-Textual Depression dataset) [96] consists of audio recordings and corresponding text transcripts obtained from interviews with 162 Mandarin-speaking participants, including both males and females. During the interviews, each participant responds to three randomly chosen questions and completes the Self-Rating Depression Scale (SDS) questionnaire [118], a widely used screening tool for assessing depression. In this dataset, participants with an SDS score exceeding 52 are classified as depressed. The dataset includes 30 depressed volunteers and 132 non-depressed volunteers. For the purpose of our study, we focus solely on the audio portion of the dataset, which has a total duration of 2.26 hours and a sampling rate of 16kHz. We adhere to the provided database description for data partitioning, utilizing 83 speakers for training and 79 speakers for evaluation.

2.1.3 CONVERGE

The second depression database used in this paper is in Mandarin and was developed as a part of the China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology (CONVERGE) study [55]. The CONVERGE study focused on subjects with increased genetic risk for MDD, and to obtain a more genetically homogeneous sample, only women participants were recruited. Each participant was interviewed by a trained interviewer. The diagnoses of depressive (dysthymia and MDD) disorders were made with the Composite International Diagnostic Interview (Chinese version) [102], which classifies diagnoses according to the Diagnostic and Statistical Manual of Mental Disorders fourth edition (DSM-IV) criteria. The database includes recordings of the interviews from 3742 individuals classified as suffering from MDD and 4217 healthy individuals. All audio recordings were collected with a sampling rate of 16kHz. The database was randomly split into 60%, 20%, and 20% for the train, development, and evaluation sets, respectively. This database contains a total of 391 hours of audio data and is characterized by a large degree of phonetic and content variability.

2.1.4 Speaker Recognition Datasets

The speaker identification (SID) models are pretrained using two publicily available datasets - VoxCeleb for English SID and CN-Celeb for Mandarin SID. The MUSAN library is used to augment noise during the SID training process.

VoxCeleb

The VoxCeleb dataset is a large-scale audio-visual dataset designed for speaker recognition research [65]. It consists of over 1 million utterances from more than 7,000 speakers, extracted from YouTube videos. The dataset is divided into two parts: VoxCeleb1 and VoxCeleb2. VoxCeleb1 contains over 100,000 utterances from 1,251 speakers, while VoxCeleb2 expands the dataset to include over 1 million utterances from 6,112 speakers. The development part of the VoxCeleb2 dataset with 5994 speakers is used as training data to train English Speaker ID model.

CN-CELEB

This is a large-scale Mandarin speaker recognition dataset collected 'in the wild'. The dataset consists of two subsets, CN-Celeb1 and CN-Celeb2. All the audio files are coded as single channels and sampled at 16kHz with 16-bit precision. For CN-Celeb1, it contains more than 130,000 utterances from 1,000 Chinese celebrities, and covers 11 different genres in real world. CN-Celeb2 contains more than 520,000 utterances from 2,000 Chinese celebrities, and covers 11 different genres [33]. A combination of data from CN-Celeb1 and CN-Celeb2 with 2800 speakers is used as training set to train the Mandarin Speaker ID model.

MUSAN

Data augmentation techniques, including additive noises and reverberation, are employed to increase the diversity and amount of training data during SID training. The process involves using the MUSAN dataset [98], which contains over 900 noises, 42 hours of music, and 60 hours of speech, as well as simulated room impulse responses (RIRs) for reverberation. A 3-fold augmentation strategy is used, combining the original "clean" training list with two augmented copies, where each recording is randomly modified using one of the following techniques: babble (adding 3-7 speakers), music (adding a single music file), noise (adding MUSAN noises at intervals), or reverb (convolving with simulated RIRs).

2.2 Feature extraction

2.2.1 Mel Frequency Cepstral and spectral features

Mel-frequency cepstral coefficients (MFCCs) and Mel frequency spectrograms represent the overall spectral envelope of the speech signal, and are closely related to the phonetic information in speech at the frame level. Inspired by speech processing literature, we use MFCCs and mel-spectrograms as input features [80,88]. For the data-augmentation study, the English experiments were performed using 20-dimensional mel spectrograms whereas the Mandarin experiments used 30-dimensional MFCCs. The window size and shift for feature extraction were varied as explained in Chapter 4. For the speaker identity disentanglement study, Mel-spectrograms are extracted using a Hanning window of length $w = 1024$ samples (64ms) and a hop size of $h = 512$ samples (32ms). The dimensionality of Mel-spectrogram features is either 40 or 80, depending on the model size.

2.2.2 ComParE 2016 Acoustic Feature Set

The ComParE (Computational Paralinguistics ChallengE) 2016 feature set has been used in paralinguistics analysis [48,66,90,94]. This set consists of 130 features which includes, among others, F0, energy, spectral, cepstral coefficients (MFCCs) and voicing related frame-level features which are referred to as low-level descriptors (LLDs). They also include the zero crossing rate, jitter, shimmer, the harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. We used the TUM's open-source openSMILE system to extract the ComParE16 features [32].

2.2.3 High-level Features

Recently, self-supervised learning (SSL) models have emerged as powerful tools in speech-processing applications [115]. These models can learn general speech representations from large volumes of unlabeled data, capturing patterns that are not specific to any particular task. One of the key advantages of SSL models is that they can be adapted to various downstream tasks, either through fine-tuning or by using them as feature extractors. In our research, we employ SSL models as feature extractors, maintaining the pre-trained weights without further adjustment. We extract frame-level features from a selection of SSL models and utilize these representations as input for our depression detection system.The specific SSL models used in this dissertation are as follows:

1. Wav2vec2.0 [7]: We employ the base pre-trained Wav2vec2.0 model, which is readily accessible through the fairseq toolkit [69]. The model's hidden dimension is set to 768. Our choice of Wav2vec2.0 is motivated by its remarkable performance across a wide range of speech-related tasks, as evidenced by its results on the SUPERB benchmark [115]. Notably, Wav2vec2.0 stands out as one of the earliest SSL models developed specifically for speech processing applications.

2. ContentVec [74]: As an extension of the HuBERT model [43], ContentVec incorporates speaker disentanglement to capture more content-related information while minimizing speaker-related information, hence its name. In our studies, we utilize the 100-cluster base model of ContentVec, which has a hidden dimension of 768. The feature extraction process for ContentVec closely resembles that of Wav2vec2.0, ensuring consistency in our approach.

3. WavLM [15]: WavLM is an SSL model known for its robustness in domain-mismatched

scenarios, such as noisy conditions, which is attributed to its signal reconstruction component. To extract features, we employ the base model configuration of WavLM, resulting in a feature dimension of 768.

4. Whisper [76]: Recently introduced, Whisper is a large-scale, weakly supervised model that has demonstrated state-of-the-art performance on speech-recognition tasks, surpassing other SSL models. Although it is weakly supervised, we include Whisper in this section for comparison with other large-scale pre-trained speech models. For our research, we select the base English-only model, which has a hidden dimension of 512. The feature extraction process is carried out using the OpenAI toolkit [12].

Since these pre-trained models are unavailable in Mandarin, high-level features are only investigated on the English DAIC-WOZ dataset.

Chapter 3

Models, Evaluation Metrics and Training Scheme

3.1 Models

In this study, multiple models were used to demonstrate that the proposed methods are generalizable and do not depend on architectural choices for improved performance. Table 3.1 provides a summary of the model architectures used in our experiments. It is important to note that all models were trained from scratch, ensuring a fair comparison and avoiding any potential bias from pre-training.

3.1.1 CNN-LSTM

The CNN-LSTM model, based on the DepAudioNet framework [61] and implemented following [8], was chosen as one of the baselines. The DepAudioNet model is a well-accepted baseline in the speech-based depression detection literature [9, 34]. The network parameters, including the number of hidden layers, learning rate, dropout probability, etc., were selected empirically. The architecture consists of 1D convolutional layers (*Conv1D*) with parameters

Table 3.1: Summary of Model architectures used in this study. *'Conv'* indicates convolutional layer. *'LSTM'* indicates Long Short-term Memory Layer. *'FC'* indicates fully connected layer. The number of layers and dimensions of each varies with the dataset size and/or input features.

Model Architecture	Initial Layers	Hidden Layers	Output Layer
CNN-LSTM	*Conv*	*LSTM*	*FC*
X-Vector+CNN	*Conv*	*Conv*	*FC*
ECAPA-TDNN	*Conv*	Time-Dilated *Conv*	*FC*
LSTM-only	*LSTM*	*LSTM*	*FC*

such as channels (C), kernel size (K), and stride (S), as well as recurrent LSTM layers with a hidden state dimension (H).

The *Conv1D* layers were followed by ReLU non-linearity, a max-pooling layer with a kernel size of 3, and a dropout layer. The final prediction layers (fully connected layers, whose inputs were the last hidden state of the preceding LSTM layer) generated the depression state predictions. In case of speaker disentanglement experiments, two branches of separate prediction layers were applied where one branch predicted the depression label and the other predicted the speaker label. The dimensions of the speaker branch prediction layer were dataset dependent: 107 for DAIC-WOZ, 83 for EATD and 1185 for CONVERGE. For MDD prediction, a sigmoid activation was applied, and the binary cross-entropy loss was used. For the SID branch in speaker disentanglement experiments, cross-entropy loss was used without any output activation.

3.1.2 X-vector Embeddings with CNN

This model consists of two parts - 1) the x-vector extractor and the downstream CNN model. This model setup was selected because the x-vector backend generates fixed length embeddings for variable length inputs which allows us to analyze the effects of the proposed data augmentation method on such model architectures. In this setting, bottleneck embeddings extracted from the pretrained X-vector model [99] are used with a fully convolutional downstream classification model. The x-vector model, made up of time-dilated neural network (TDNN) consists of 5 layers of TDNN convolutional layers followed by average pooling and two fully-connected layers. The implementation of the TDNN model is based on [51]. The downstream depression classification model is made up of two CNN layers followed by two fully connected layers.

3.1.3 ECAPA-TDNN

ECAPA-TDNN is a model architecture originally proposed for speaker recognition tasks [20] and currently represents the state-of-the-art in speaker identification (SID) and Speech Emotion Recognition (SER [79]). In this thesis, we propose a modified version of the original ECAPA model to adapt to the smaller training dataset of depression classification and address the inherent class imbalance problems. The modifications include empirically adjusting the kernel (K) and stride (S) of the input convolution layer, the number of channels (C) in the intermediate layers, the attention dimension, the embedding dimension, and the dimensions of the prediction layers. For speaker disentanglement experiments, the prediction layer dimension is 107 based on the number of speakers in the training set of the DAIC-WoZ dataset.

3.1.4 LSTM-only

The LSTM-only architecture was employed for high-level features of the DAIC-WOZ dataset. This architecture was utilized to process the latent representations obtained from the SSL models, which were used as encoders (feature extractors) in this study. By leveraging the LSTM-only architecture, we aim to capture the temporal dependencies and long-term contextual information present in the high-level features extracted from the SSL models, enabling effective depression classification.

The model architecture consisted of an input LSTM layer with a hidden state dimension of $H = 256$, followed by 5 hidden LSTM layers, each having the same hidden state dimension as the input layer. Similar to the CNN-LSTM model, the output of the last hidden state of the preceding LSTM layer served as input to the prediction layer. The dimensions of the final prediction layer were 107 for the DAIC-WoZ dataset and 83 for EATD.

3.2 Evaluation Metrics

The evaluation metrics used to compare models are presented in this section. Every model is evaluated on it's ability to classify depression status. In addition, for the speaker disentanglement experiments, the ability to disentangle speaker identity is also measured.

3.2.1 Depression Detection

Depression detection is evaluated using the macro average F1-score (F1-AVG) of depression (F1-D) and non-depression (F1-ND) classes computed at the speaker level. We opted to report F1-AVG because it provides a balanced representation of both D (Depression) and ND (Non-Depression) prediction capabilities. For a givenconfusion matrix [TN, FP, FN, TP], then the F1-AVG is calculated as:

$$\text{Precision}_D = \frac{TP}{TP + FP} \tag{3.1}$$

$$\text{Recall}_D = \frac{TP}{TP + FN} \tag{3.2}$$

$$\text{F1}_D = 2 \times \frac{\text{Precision}_D \times \text{Recall}_D}{\text{Precision}_D + \text{Recall}_D} \tag{3.3}$$

$$\text{Precision}_{ND} = \frac{TN}{TN + FN} \tag{3.4}$$

$$\text{Recall}_{ND} = \frac{TN}{TN + FP} \tag{3.5}$$

$$\text{F1}_{ND} = 2 \times \frac{\text{Precision}_{ND} \times \text{Recall}_{ND}}{\text{Precision}_{ND} + \text{Recall}_{ND}} \tag{3.6}$$

$$\text{F1-AVG} = \frac{\text{F1}_D + \text{F1}_{ND}}{2} \tag{3.7}$$

Where:

- TP is the number of True Positives
- TN is the number of True Negatives
- FP is the number of False Positives
- FN is the number of False Negatives
- D represents the Depressed class
- ND represents the Non-Depressed class

3.2.2 Speaker-Separability and Identification

Inspired by the Voice-privacy literature [68, 103], we employ two metrics to quantify speaker separability and identification: Gain of Voice Distinctiveness (G_{VD}) and De-Identification Score (DeID). G_{VD} is measured in decibels (dB), while DeID is expressed as a percentage. A G_{VD} value of 0 dB indicates that the voice distinctiveness remains the same before and after disentanglement. A negative G_{VD} value signifies a decrease in speaker distinctiveness, while a positive value suggests an increase. In the case of DeID, a score of 100% represents an optimal de-identification strategy, whereas a score of 0% indicates that the disentanglement approach has no effect on speaker identification. The following equations define G_{VD} and DeID as described in [68, 103].

$$G_{VD} = 10log_{10}\frac{D_{diag}(M_{dd})}{D_{diag}(M_{oo})} \tag{3.8}$$

$$\text{DeID} = 1 - \frac{D_{diag}(M_{od})}{D_{diag}(M_{oo})} \tag{3.9}$$

where M_{dd}, M_{oo} and M_{od} are voice similarity matrices and $D_{diag}(M)$ is called the diagonal dominance. In this paper, o stands for the baseline (original) model and d stands for the disentangled model. A voice similarity matrix $M_{AB} = (M(i,j))_{1\leq i\leq N, 1\leq j\leq N}$ is defined for a set of N speakers where each entry $M(i,j)$ defines the similarity between speaker i and j, calculated as:

$$M_{AB}(i,j) = sigmoid(\frac{1}{n_i n_j}\sum_{\substack{1\leq k\leq n_i \; and \; 1\leq l\leq n_j \\ k\neq l \; if \; i=j}} LLR(x_k^{(i)}, x_l^{(j)})) \tag{3.10}$$

where $LLR(x_k^{(i)}, x_l^{(j)})$ is the log-likelihood-ratio obtained from Probabilistic Linear Discriminant Analysis (PLDA) model between segment k from speaker i and segment l from speaker j. n_i and n_j are number of segments from speaker i and speaker j, respectively. A and B denoted the models from which speaker representations $x_k^{(i)}$ and $x_l^{(j)}$ are taken, respectively.

The diagonal dominance is defined as the absolute difference between average diagonal and off-diagonal elements as follows:

$$D_{diag}(M) = \left| \sum_{1\leq i\leq N} \frac{M(i,i)}{N} - \sum_{\substack{1\leq j\leq N \; and \; 1\leq k\leq N \\ j\neq k}} \frac{M(j,k)}{N(N-1)} \right| \tag{3.11}$$

3.3 Training and Evaluation Scheme

In general, to avoid overfitting the models to the training set, one or more of the following mechanisms were adopted - 1) random cropping and selection of segments to ensure class imbalance does not influence results along with aggregation of 5 models (empirically chosen, similar to 5-fold cross validation) trained with different random seeds to average the effects of random segmentations, 2) reduction of learning rate using a factor of 0.9 when the validation loss does not reduce for two successive epochs and 3) dropout with $p = 0.6$ for LSTM-only, 0.5 for ECAPA-TDNN, and 0.05 for CNN-LSTM models.

In the case of the DAIC-WOZ dataset, all three above-mentioned methods were used. Each utterance was randomly cropped into fragments of the length of the shortest utterance, and each fragment was further segmented into multiple segments. Segment lengths were set to 3.84 seconds, which corresponds to 120 frames for Mel-spectrogram, 61440 samples for raw-audio, 200 frames for Wav2vec2.0 features, and 193 frames for ContentVec, WavLM, and Whisper. A training subset was generated by randomly sampling, without replacement, an equal number of depression and non-depression segments. Five separate models were trained for each experiment using randomly generated training subsets.

In contrast, for the EATD and the CONVERGE datasets, only mechanisms 2 and 3 were used. Since EATD had utterances with equal duration and CONVERGE had a sufficient number of samples, segments were generated without random cropping and sampling but the segment length was kept the same as DAIC-WOZ (3.84s). Each experiment was performed by training only one model using all of the training data. Same as before, this was done to ensure that the improvements observed from the proposed approach were not attributed to training data sub-sampling.

At the evaluation stage, segment-level prediction scores are rounded to 0 or 1, representing “non-depressed’ or “depressed’ classes, respectively. Then, each model generates a speaker-level prediction score by averaging all segment-level scores. For experiments conducted on the DAIC-WOZ dataset where more than one model is trained, 5-model prediction aggregation is performed using two different methods - averaging (5M-AVG) or majority voting (5M-MV). For the averaging method (5M-AVG), speaker-level scores from all models are averaged and rounded for each individual. In contrast, for the majority voting method (5M-MV), speaker-level scores for all models are first rounded, and then a majority vote is taken. All rounding operations use a threshold of 0.5 to determine the final predicted class label for each individual.

For the speaker-separability experiments, Probabilistic Linear Discriminant Analysis (PLDA) models are trained using embeddings of 25 speakers (randomly selected). For G_{VD} computation, two PLDA models are trained separately - one using embeddings from the baseline and the other using embeddings from the disentangled model. On the other hand, for DeID computation, a single PLDA model is trained by combining embeddings from both baseline and disentangled models. Evaluation of G_{VD} and DeID is done on the remaining 10 speakers. For each speaker, to reduce computational complexity, 50 segments are randomly chosen and similarity matrices are generated from the equations described in Section 3.2.2. The Log-likelihood scores in the referenced equations are computed using the trained PLDA models. The experiments are repeated three times using different random seeds and the average G_{VD} and DeID are reported.

Lastly, all model hyperparameters, including learning rate, batch size, and learning rate decay, are kept the same for both the baseline and the corresponding disentanglement experiments. The only hyperparameter that varies is the λ parameter, which controls the degree of disentanglement. For baseline experiments, λ is set to 0, while for disentanglement experiments, λ is selected empirically to achieve the desired level of disentanglement in the latent representations.

3.4 Chapter Summary

In this chapter, we provide a comprehensive overview of the models, evaluation metrics and the model training scheme employed in our study. We describe the characteristics and properties of each dataset, including the DAIC-WOZ (English), EATD (Mandarin), and CONVERGE (Mandarin) datasets, highlighting their relevance to our research. Furthermore, we discuss the various input features utilized, such as Mel-spectrograms, raw audio signals, and high-level features extracted from pre-trained self-supervised learning models. We also present the architectures of the models used in our experiments, including the CNN-LSTM, X-vector with CNN, ECAPA-TDNN, and LSTM-only models, along with the training and evaluation mechanism. Finally, we introduce the depression classification evaluation metrics, such as the F1 score, as well as the privacy-related metrics, including the Gain of Voice Distinctiveness (GVD) and De-Identification Score (DeID), which are used to assess the performance and effectiveness of our proposed methods in terms of both depression detection accuracy and speaker privacy preservation.

Chapter 4

Data Augmentation for Depression Detection from Speech

In this chapter, we discuss our work, FrAUG, [84] on a frame rate-based data augmentation method for depression detection from speech.

4.1 Background

Recently, speech-based automatic diagnosis of depression has gained significant momentum [2, 4, 24] and advancements in deep learning have pushed their performance to newer heights [41, 42, 61, 70, 87, 114]. However, data scarcity still remains one of the major challenges in building reliable systems for MDD modeling purposes because collection of such data can be expensive and challenging. Therefore, there is a need to adopt data augmentation strategies to increase the amount of training data. However, conventional data augmentation techniques (e.g., Vocal-Tract Length Perturbation, VTLP [46], pitch-shifting and speed perturbation [87]) can be counter-productive when applied to para-linguistic applications such as depression detection because manipulation of acoustic features such as pitch or formants may result in loss of useful information related to the underlying health condition.

Previously, Generative Adversarial Network (GAN) [40] based data augmentation was proposed for depression detection [114]. However, GANs themselves require a significant amount of training data to be effective. In [70], a multi-window data augmentation was proposed for emotion recognition which used multiple frame-widths. However, the methods proposed in [42, 70, 87] were not compared with conventional data augmentation techniques and were only evaluated using one model that is trained from scratch.

In contrast, in our work, a frame rate based data augmentation technique is proposed for the task of depression detection from speech signals. New feature samples were created by varying the frame-width as well as the frame-shift during the feature extraction process. By changing the frame rate parameters, the model was provided with different sets of time-frequency resolutions during the training stage. This ensured that acoustic parameters which are thought to correlate with the mental state of the speaker (e.g., pitch, formant frequencies, speaking-rate etc. [2, 19]) were not inadvertently modified. Additionally, it is shown that the proposed method outperforms conventional data augmentation methods and can also be applied in the pretraining stage of a model.

4.2 Method

In this section, we describe the proposed data augmentation technique, FrAUG. Given an input speech signal $x[n]$, the windowing and feature extraction process for spectral features can be represented as:

$$X_r[k] = \sum_{m=0}^{L-1} x[m]w[rR - m]e^{-j(2\pi k/N)m}, \tag{4.1}$$

where, $w[rR - m]$ is the sliding window, $r \in \mathbb{Z}$, N is the DFT size, L is the frame-width and R is the frame-shift [75]. The windows overlap by $O = L - R$. R and O are usually specified as a fraction of L, which is specified in time or number of samples.

Changing the values of L and R and thereby the frame rate, changes the time-frequency resolution of the extracted features. A smaller window length leads to a wide-band spectrogram with better time-resolution whereas a larger window length results in a narrow-band spectrogram with better frequency resolution (See Figure 4.1). Conventionally, to balance resolutions between time and frequency, the parameters L, R and O are fixed. A Hamming window with $L = 25ms$ and $R = 40\%$ (i.e $R = 10ms$) is the most common configuration [73].

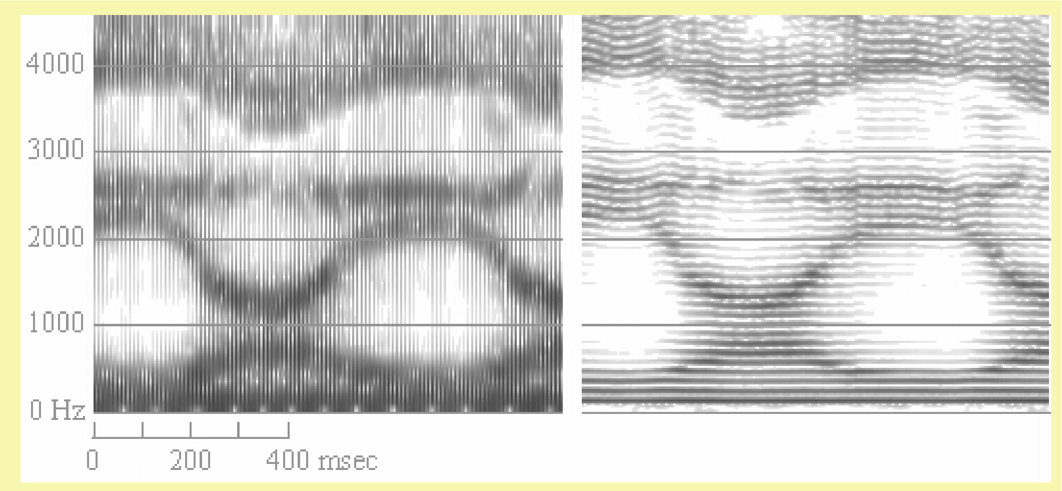


Figure 4.1: Wide-band and Narrow-band spectrograms for the same speech signal

In FrAUG, given a baseline frame-rate with parameters (L_1, R_1), we augment the training data with features extracted using multiple frame rates with parameters L_i and R_j where $i, j \in \mathbb{N}$. For example, to perform an 8-fold augmentation, frame widths of L_2, L_3 and frame-shifts R_2, R_3 are used along with baseline parameters, resulting in 9 different combinations such as (L_1, R_2),(L_2, R_3),(L_3, R_2), etc. Thus, the model is provided with 8 additional time-frequency resolutions in the training stage. The main advantage of the proposed method is that it does not alter vocal tract or voice source parameters and is independent of the dataset and model used. FrAUG can be extended to other acoustic features as well.

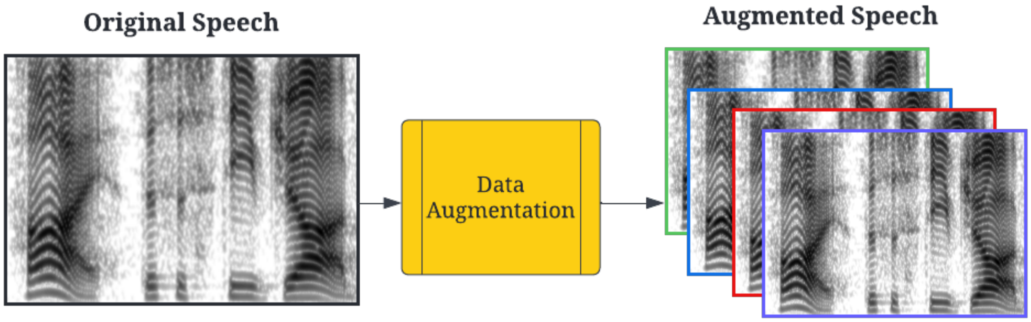


Figure 4.2: FrAUG provides multiple time-frequency resolutions at the training stage

4.3 The Experimental Setup

The proposed method was applied on two different models using three distinct datasets and two different input acoustic features. For the English DAIC-WOZ dataset and the Mandarin EATD dataset, a DepAudioNet [61] was trained using mel-spectrograms. To demonstrate that the proposed method generalizes to other model frameworks, for the CONVERGE dataset, a pretrained x-vector embedding generator trained on MFCCs followed by a convolutional neural network (CNN) backend was used. The models used in this work are described in the following subsections.

4.3.1 Models

DepAudioNet

For the English DAIC-WoZ dataset, the CNN-LSTM model based on the DepAudioNet framework (explained in Section 3.1) was used. The inputs to the model were 40-dimensional Mel-Spectrograms and the model parameters were - one *Conv1D* layer ($C = 128$, $K = 3$, $S = 1$) and two unidirectional LSTM layers ($H = 128$). Pre-processing of features, as explained in Section 3.3, using random sampling and cropping, was applied.

For the baseline experiments, mel-spectrograms were extracted with frame rate parameters of $L = 64ms$, and $R = 50\%$, as proposed in [61]. When FrAUG was applied, training data were augmented with up to 8 folds using additional frame rates with parameters $L = 32ms$ and $L = 128ms$ and an overlap R of 25% and 10%. The augmentation frame rates were chosen empirically. Even when augmentation was applied, mel-spectrograms for test and development sets were extracted at the baseline frame rate.

For the EATD dataset, the model parameters were the same as those used for the DAIC-WOZ dataset when Mel-spectrograms were used as input features. However, pre-processing of the training data was not adopted. Similar to the DAIC-WoZ dataset, baseline experiments used mel-spectrograms with frame rate parameters of $L = 64ms$, and $R = 50\%$ and up to 8 folds of data augmentation was evaluated.

4.3.2 X-vector Embedding with CNN Classifier

For the CONVERGE dataset, the x-vector embeddings with downstream CNN classifier framework was used (explained in Section 3.1). The x-vector model was pre-trained using CN-Celeb [33], a Mandarin speaker ID dataset. A Kaldi recipe was followed for training the x-vector model [53]. After pre-training the x-vector model, embeddings for the CONVERGE dataset were generated which were then used to train a downstream CNN network for classifying depression. Frame-rate-based data augmentation was only applied during the training of the downstream network i.e. embeddings were extracted for the augmented depression training data along with the unaugmented development and test data.

x-vector embeddings for the baseline experiments were generated using MFCCs extracted with frame rate parameters of $L = 25ms$ and $R = 40\%$, as proposed in [99]. When FrAUG was applied, x-vectors were generated using MFCCs extracted with additional frame rate parameters of $L = 10ms$, $L = 32ms$ and $R = 50\%$, $R = 25\%$. Similar to the previous experiment, augmentation frame rates were chosen empirically and up to 8-fold data augmentation was evaluated. Test and development set features were always extracted at the baseline frame rates.

4.4 Results and Discussion

The effectiveness of the proposed approach is demonstrated in three stages – first, using the DAIC-WoZ English data. Then, the performance of the proposed method is compared to conventional data augmentation techniques and lastly, the generalizability of the proposed method is evaluated by applying it to two different datasets in Mandarin. For the EATD dataset, the training data was used with no pre-processing. For the CONVERGE dataset a different backend system was used with different input acoustic features compared to the DAIC-WoZ dataset. Model performance is reported in terms of the F1-score [16] which is the harmonic mean of precision and recall. Statistical significance ($p < 0.05$) was evaluated using the McNemar's test [64]

4.4.1 Multi Frame Rate Training

In the first set of experiments, DepAudioNet models were trained on the DAIC-WOZ dataset using different combinations of single frame rate and multiple frame rates. In this work, DepAudioNet was chosen as a baseline mainly because of DepAudioNet's open-source code [9]. Performance comparison of single rate training versus multiple frame rate training on the development set of the DAIC-WOZ dataset is shown in Table 4.1. The baseline frame rate of $L = 64ms, R = 50\%$ has an F1-score of 0.619. This is comparable to the reported F1-scores of 0.610 in prior works [9, 61]. In contrast, the best performing configuration is the one with 5-fold data augmentation with multiple frame rate hyper-parameters of $L = 64ms, 128ms$ and $R = 50\%, 25\%, 10\%$. Higher folds of data augmentation were also evaluated but 5-fold produced the best results.

The best performing system has an F1-score of 0.656, a relative improvement of 5.97%

(p = 4.72e−6) when compared to the baseline. This best performing configuration is better than any of the single frame rate performances, including when only the frame-widths are manipulated as in [70]. A possible explanation for this result might be that a particular combination of time-frequency resolutions, provided to the model in FrAUG, contains depression-related information that is not available to the model when trained using single frame rate features.

Table 4.1: Results, in terms of F1-score, comparing single frame rate training versus multi frame rate training using DepAudioNet and the DAIC-WOZ development set. *L* and *R* represent frame-width and frame-shift, respectively. * denotes the baseline F1-score. The best F1-score is boldfaced.

↓ L\R →	50%	25%	10%	50%, 25%	50%, 10%	25%, 10%	50%, 25%, 10%
32ms	0.601	0.604	0.569	0.606	0.633	0.562	0.604
64ms	0.619*	0.638	0.587	0.612	0.620	0.599	0.613
128ms	0.648	0.627	0.588	0.618	0.628	0.638	0.616
32ms, 64ms	0.637	0.607	0.579	0.615	0.617	0.623	0.602
64ms, 128ms	0.635	0.612	0.576	0.620	0.625	0.617	**0.656**
32ms, 128ms	0.623	0.633	0.590	0.647	0.610	0.602	0.615
32ms, 64ms, 128ms	0.626	0.607	0.546	0.655	0.600	0.582	0.596

4.4.2 FrAUG versus Conventional Data Augmentation Methods

To compare FrAUG with conventional data augmentation techniques, DepAudioNet models were trained using the DAIC-WOZ dataset with FrAUG, noise augmentation [99] and VTLP-based augmentation [60, 113]. The noise augmentation method was similar to the one used in Kaldi. The MUSAN library [98] was used to augment every utterance with randomly chosen foreground noise samples at SNRs of 0,5,10, or 15 dB [99]. The VTLP augmentation was based on the nlpaug library [60] and the method proposed in [46]. For every augmentation method, up to 8-folds of data augmentation was applied and the best performing configuration was selected. The results comparing these augmentation methods are presented in Table 4.2. In case of noise, 7-fold augmentation performed the best and for VTLP, 3-fold augmentation was the best. In contrast, for FrAUG, 5-fold augmentation performed the best.

Table 4.2: Results, in terms of F1-score, for depression detection on the DAIC-WOZ dataset comparing proposed method with conventional data augmentation techniques. The best F1-score is boldfaced.

Augmentation Strategy	F1-AVG	Data Augmentation
Baseline	0.619	None
Noise [98]	0.579	7x
VTLP [46]	0.630	3x
FrAUG	**0.656**	5x

As seen in Table 4.2, FrAUG outperforms noise augmentation by 13.2% ($p = 3.43\text{e}{-6}$) and VTLP by 4.1% ($p = 4.92\text{e}{-6}$). One possible explanation for this result is that VTLP alters the spectral shape and therefore might be preserving less information about the depressive state of the speaker. In case of noise augmentation, a domain mis-match between training and development data (noisy vs clean) may be the reason for degraded performance. This shows that FrAUG can serve as an effective data augmentation strategy for depression detection without interfering with task-related acoustic information.

4.4.3 Extension to EATD and CONVERGE Dataset

To show that the proposed approach is independent of the dataset, the pre-processing, the model or the input acoustic feature, it was evaluated on the EATD and the CONVERGE datasets using embeddings extracted from a pre-trained x-vector system and the DepAudioNet without pre-processing, respectively. For the Converge dataset, the extracted embeddings were used to train a CNN model to classify utterances as cases (depressed) or controls (healthy). 3x, 5x and 8x data augmentation was applied.

The effectiveness of FrAUG when applied to the EATD and CONVERGE datasets is evident from the results presented in Tables 4.4 and 4.3. For the CONVERGE dataset, in comparison to the baseline F1-score of 0.674 (development) and 0.664 (test), the best

Table 4.3: Results, in terms of F1-score, for depression detection on the EATD dataset using DepAudioNet model, with and without FrAUG. The best F1-score is boldfaced.

L,R Configuration	F1-Avg	Data Augmentation
Baseline (L=64ms,R=50%)	0.517	None

L=64ms, 128ms
R=50%, 25%
	0.523	3x

L=64ms, 128ms
R=50%, 25%, 10%
	0.549	5x

L=32ms, 64ms, 128ms
R=50%, 25%, 10%
	0.551	8x

Table 4.4: Results, in terms of F1-score, for depression detection on the CONVERGE dataset using x-vector embeddings with a CNN classifier as the backend, with and without FrAUG. The best F1-score is boldfaced.

L,R Configuration	development	Test	Data Augmentation
Baseline (L=25ms,R=40%)	0.674	0.664	None

L=10ms, 25ms
R=40%, 25%
	0.708	0.699	3x

L=10ms, 25ms
R=40%, 50%, 25%
	0.721	0.710	5x

L=10ms, 25ms, 32ms
R=40%, 50%, 25%
	0.729	**0.723**	8x

performing configuration (8-fold augmentation) has a performance of 0.729 and 0.723, respectively. This is an improvement of 8.21% on the development set ($p = 6.03\text{e}{-6}$) and 8.77% on the test set ($p = 5.91\text{e}{-6}$). Even though the downstream model was trained on x-vector embeddings and not on the acoustic features themselves, FrAUG improves the classification performance. This is a rather significant outcome because this shows that FrAUG can be beneficial in improving system performance even when applied to downstream tasks after the pre-training step. An important implication of this result is that FrAUG can be applied irrespective of the model training style - supervised pre-training, training from scratch, etc.

Similarly, when FrAUG is applied to the EATD dataset, the baseline performance (0.517) improves by 6.57% when 8x data augmentation is applied showing that the performance gains from the proposed method cannot be attributed to the random cropping and segment selection. Additionally, the performance of both models improves consistently with increasing amounts of training data. These results also show that the proposed approach does not depend on specific frame rates. Instead, the FrAUG configuration can be considered as a hyperparameter and can be tuned in a way similar to other model parameters.

4.5 Chapter Summary

In this chapter, a data augmentation method, called FrAUG, was proposed for depression detection from speech. Training data were augmented with new feature samples created by varying the frame-width as well as the frame-shift parameters during feature extraction. Thus, the proposed approach did not modify vocal tract or voice source related parameters and hence preserved acoustic information that may be important for MDD modeling purposes. The proposed method of data augmentation performed better than a baseline system with no augmentation and two commonly used data augmentation methods. Lastly, the generalizability of the said method was demonstrated by improvements in classification performance on two different datasets with a different model and different input features.

FrAUG improved the classification performance of DepAudioNet [61] trained using mel-spectrograms on the DAIC-WOZ and the EATD datasets, and of a downstream network trained with x-vector embeddings generated from a pre-trained model [99] using MFCCs on the CONVERGE dataset. It can therefore be suggested that the proposed method is independent of the dataset, the language, the input acoustic features, data pre-processing, the model or the model training style. Frame rate based data augmentation, therefore, can be reliably used to increase the amount of training data and will prove to be useful in the development of large-scale MDD screening systems.

Chapter 5

Adversarial Speaker Disentanglement

5.1 Introduction

Depression detection through speech analysis is gaining traction in the research community. Various acoustic features, such as x-vectors [84], i-vectors [21], and other speaker embeddings [26], have demonstrated their effectiveness in diagnosing a speaker's mental state. However, these features also carry information about the speaker's identity [99], which can be detrimental to privacy preservation, a crucial factor in the adoption of digital mental health screening systems [59]. Consequently, a significant question that remains unanswered is whether depression detection can be performed in a manner that is invariant to speaker identity. Furthermore, it is unclear if there are components of the speech signal that characterize a speaker but may not be relevant to their mental health status. Recently, two studies introduced algorithms to preserve privacy during depression detection: [27] proposed sine-wave speech representation, and [101] employed federated learning. However, both studies reported a degradation in depression detection performance while attempting to preserve patient privacy. In this chapter, we introduce the paradigm of adversarial learning as a means to disentangle speaker and depression characteristics, aiming to address the challenges of privacy-preserving depression detection while maintaining high accuracy.

Adversarial speaker normalization has previously been explored in the context of emotion recognition [38, 52, 117]. Yin et al. [117] utilized multi-modal features (speech, text, and video) to perform speaker-invariant domain adaptation for emotion recognition. Li et al. [52] proposed a gradient reversal technique combined with an entropy loss to disentangle emotion and speaker information. Gat et al. [38] fine-tuned a pre-trained Hubert-base model [43], which has 300M parameters, using gradient-based adversarial learning. However, fine-tuning such large models can be data and computationally intensive, requiring substantial amounts of in-domain data. Furthermore, these studies employed the IEMOCAP and MSP-Improv datasets, which consist of monolingual and acted audio data [13, 14], limiting the generalizability of their findings to real-world scenarios. Additionally, previous work does not quantify the amount of speaker information that was disentangled, making it challenging to understand and analyze the effectiveness of the proposed methods.

In contrast, we propose adversarial disentanglement of speaker-identity and depression information, using speech-features only, on datasets with conversational and non-acted speech. In addition to Mel-spectrograms and raw-audio signals, we propose the use of CompareE16 features as well as latent representations from four large-scale pretrained models as the input features. Further, unlike prior studies in emotion recognition, we show that the benefits of the proposed method extend to another language (Mandarin). Lastly, we measure the voice-privacy attributes using metrics like De-Identification Scores (DeID) and Gain of Voice Distinctiveness (G_{VD}) to show that the proposed method improves depression detection performance while simultaneously reducing speaker-separability and identification. The two above-mentioned metrics quantify the amount of speaker information that is disentangled.

5.2 Preliminary Experiments

Preliminary experiments are conducted on the English dataset, DAIC-WOZ [104], to investigate the aspects of privacy preservation and speaker bias in the context of depression detection. The database is described in detail in Section 2.1.

5.2.1 Privacy Preservation in Depression Detection

As previously mentioned, the utilization of speaker-related features, such as speaker embeddings, can inadvertently lead to individual identification. For instance, in our preliminary work, embeddings from an ECAPA-TDNN model, recognized for its state-of-the-art performance in Speaker Identification (SID), were employed to train a straightforward support vector classifier (SVC) SID system. This configuration achieved an SID accuracy of 88% on the DAIC-WOZ dataset, a well-known dataset for depression detection in English [104]. Notably, the ECAPA-TDNN model was primarily optimized for depression detection rather than speaker prediction. This underscores the potential privacy risks associated with depression detection frameworks heavily reliant on speaker-related features.

5.2.2 Speaker-Bias in Depression Detection

In addition to the well-documented privacy concerns linked with an excessive reliance on speaker features [83], another potential issue is the model's susceptibility to overfitting on the speakers present in the training set. To explore this matter, a straightforward approach involves normalizing speaker information across all utterances in a dataset. This is achieved by utilizing a voice conversion (VC) system to transform all speakers' utterances into the voice of a single speaker, followed by training the depression classification system on the converted dataset. An improvement in depression classification performance following the single-speaker conversion process compared to the baseline suggests that speaker-identity-related features might introduce bias in depression detection. Therefore, this section presents a preliminary VC experiment conducted using the DAIC-WOZ dataset.

The VQMIVC (Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion, [106]) system, known for its state-of-the-art performance in voice conversion (VC), was employed to convert all speakers in the DAIC-WOZ dataset into a single speaker(p334_047). Several additional steps were taken to ensure the quality of the converted utterances. Firstly, each utterance was segmented into non-overlapping 50-second clips, and the conversion was applied to each clip separately, followed by concatenation. Secondly, to address the issue of audio loudness discrepancy, the loudness of each segment in DAIC-WOZ was scaled to match the maximum loudness of the reference waveform before conversion. Additionally, the quality of the converted audio files was manually verified. The target speaker p334_047 was chosen because it was provided with the demo of the VC model. Another target speaker, p225_038, was evaluated, but the conversion quality was found to be poor.

For major depressive disorder (MDD) classification, the DepAudioNet model [61] is selected. Both the baseline and voice conversion (VC) experiments use the same feature processing, model hyperparameters, configurations, and dataset splits, as described in Section 2.2 and 3.1. The results of the VC experiment are reported in Table 5.1 in terms of F1-AVG, which is the macro average of the F1-Scores for the two classes - depressed (D) and non-depressed (ND). The metric is explained in Section 3.2.

Table 5.1: Depression Detection performance in terms of F1-AVG on the DAIC-WOZ dataset, with and without voice conversion (VC), using the DepAudioNet model trained using Mel-Spectrograms.

Experiment	F1-AVG
DepAudioNet [61]	0.6081
DepAudioNet+VC	0.6237

Table 5.1 illustrates that converting all utterances into a single speaker improves depression

classification performance, with the F1-AVG increasing from 0.6081 for the DepAudioNet baseline to 0.6237 for the VC DepAudioNet. This finding supports the hypothesis that some speaker-related features may introduce bias in depression detection.

However, employing voice conversion (VC) to address speaker bias in depression detection may not be an ideal solution for several reasons. Firstly, even state-of-the-art (SOTA) VC systems can lead to content loss for certain speakers [74], potentially resulting in the loss of depression-related information during conversion. Secondly, there may be a dataset-domain discrepancy between the VC training data (e.g., VCTK [105]) and the target dataset (e.g., DAIC-WOZ), which could still preserve speaker information and introduce bias. Notably, the VCTK dataset comprises accented read speech by native English speakers from the UK, while DAIC-WOZ features spontaneous American English speech directed towards a robotic AI assistant. Furthermore, differences in channel attributes such as loudness between the two datasets could affect the success of VC systems trained on VCTK when evaluated on DAIC-WOZ [45]. Lastly, converting an entire dataset using VC can be computationally intensive and necessitates meticulous manual verification, rendering it impractical for real-world applications.

5.3 Adversarial Learning

We propose a loss-based adversarial learning mechanism for speaker-disentangled depression detection. Inspired by the domain-adversarial training proposed in [36], our approach involves a loss minimization-maximization technique.

Let the number of samples in a training batch be N. The loss used for the prediction of MDD binary labels is:

$$L_{MDD} = -\frac{1}{N}\sum_{n=1}^{N}[Y_n \cdot \log{(p_n)} + (1 - Y_n) \cdot \log{(1 - p_n)}] \tag{5.1}$$

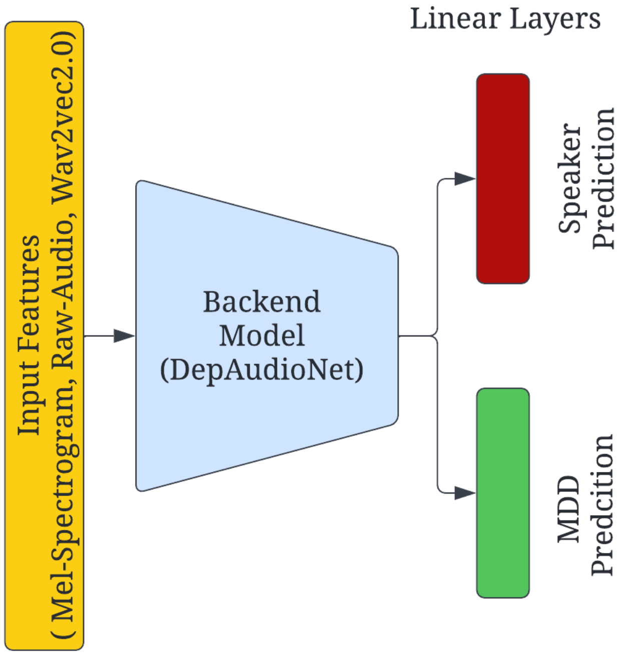


Figure 5.1: Block diagram representing adversarial disentanglement of speaker and depression characteristics.

$Y_n \in \{0, 1\}$ is the class label for the n^{th} sample and p_n is the probability that sample n's label is depressed. If we denote the total number of unique speakers as M, the adversarial loss for speaker ID prediction is defined as -

$$L_{adv} = -\frac{1}{N}\sum_{n=1}^{N}[\log \frac{\exp{(x_{n,\hat{n}})}}{\sum_{m=1}^{M}\exp{(x_{n,m})}}], \tag{5.2}$$

where $x_{n,m}$ is the score of the n^{th} sample's speaker ID being predicted as speaker m where $m \in 1, 2, ..., M$. And, $\hat{n}$ is the coordinate for the ground-truth speaker ID of sample n.

To train the model in a speaker-identity-invariant manner, during optimization, we minimize the depression loss and maximize the speaker prediction loss. This can be written as:

$$L_{total\_adv} = L_{MDD} - \lambda(L_{adv}) \tag{5.3}$$

where λ is an empirically determined hyperparameter that controls how much of the speaker loss contributes to the total loss. Initial λ values were selected to be similar to those reported in the adversarial learning literature for emotion recognition(1e-3) [117]. We experimented with higher and lower values and chose the best performing λ values. By minimizing depression loss and maximizing speaker prediction loss, we force the model to focus more on depression-discriminatory information and ignore some speaker-discriminatory information, thereby making the model invariant to changes in some speaker-specific characteristics.

5.4 Experimental Details

In this chapter, we conducted experiments using three datasets: DAIC-WOZ (English) [104], a subset of CONVERGE (Mandarin) [55], and EATD [96]. We employed three different backend models: 1) a modified version of DepAudioNet [61], 2) an ECAPA-TDNN model, and 3) an LSTM-only model. For the English dataset, we evaluated the performance using all seven input features, while for the Mandarin datasets, we focused on Mel-Spectrogram and Raw-Audio signals only. The choice of model parameters and features were empirically determined during baseline pilot experiments.

5.4.1 Models

CNN-LSTM

For the DAIC-WOZ dataset, when using 40-dimensional Mel-spectrograms as input, the model comprised one *Conv1D* layer ($C = 128$, $K = 3$, $S = 1$) and two unidirectional LSTM layers ($H = 128$). When using raw audio signals, two *Conv1D* layers ($C\_1 = 128$, $K\_1 = 1024$, $S\_1 = 512$, $C\_2 = 128$, $K\_2 = 3$, $S\_2 = 1$) and two LSTM layers ($H = 128$) were employed. For 130-dimensional ComParE16 features, the model consisted of one *Conv1D* layer ($C = 256$, $K = 3$, $S = 1$) and two unidirectional LSTM layers ($H = 256$).

For the EATD dataset, the model parameters were the same as those used for the DAIC-WOZ dataset when Mel-spectrograms and raw audio signals were used as input features. In contrast, for the CONVERGE dataset and 40-dimensional Mel-spectrograms, the model included two *Conv1D* layers ($K\_1 = 3$, $S\_1 = 1$, $K\_2 = 3$, $S\_2 = 1$) and four unidirectional LSTM layers ($H = 128$). When using raw audio signals, two *Conv1D* layers ($K\_1 = 1024$, $S\_1 = 512$, $K\_2 = 3$, $S\_2 = 1$) and four LSTM layers ($H = 512$) were utilized.

ECAPA-TDNN

For the DAIC-WOZ dataset with Mel-spectrograms as input, the model consists of one *Conv1D* layer ($C = 128$, $K = 5$, $S = 1$) followed by three SE-Res2Blocks, each with identical channel dimension, kernel size, and stride ($C = 128$, $K = 5$, $S = 1$). The three SE-Res2Blocks have increasing dilation steps of 2, 3, and 4. Our experiments revealed that using 80-dimensional

Mel-spectrograms yielded better performance compared to 40-dimensional ones. In addition to Mel-spectrograms, we also investigate the use of raw audio signals as input features. In this case, one input convolution layer ($C = 128$, $K = 1024$, $S = 512$) is followed by three SE-Res2Blocks with the same dimensions as those used with Mel-spectrograms.

For both Mel-spectrograms and raw audio signals, the attention dimension is set to 64, and the embedding dimension is set to 128. The final projection layer is similar to the CNN-LSTM architecture, but the input to the prediction layers comes from the embedding layer, as opposed to the last hidden state of the LSTM layer in the CNN-LSTM model.

LSTM-only

The LSTM-only model, explained in Section 3.1 is used with the SSL features (Section 2.2) and the DAIC-WOZ datasets.

5.5 Results and Discussion

The experimental results are organized and discussed as follows - First, we focus on the DAIC-WOZ dataset and compare the performance of the proposed adversarial speaker disentanglement methods with the baseline approaches that do not employ disentanglement. We evaluate all considered model-feature combinations and limit our discussion of results to segment-level probability averaging, following previous studies [8, 34, 61]. Next, we extend the proposed method to the EATD and CONVERGE datasets to assess its generalizability and effectiveness across different languages and domains. Unless otherwise specified, the reported relative improvements are statistically significant, as determined by the McNemar's test [64].

5.5.1 Speaker Disentanglement with DAIC-WOZ

Figures 5.2 and 5.3 shows the relative change in MDD classification F1-AVG, and the speaker separability metrics G_{VD} (in dB) and $De - ID$ (in%) for each model-feature combination when ADV is applied, respectively. Detailed results, in terms of F1-Score for 5M-AVG and 5M-MV are presented in Tables 5.2 and 5.3, respectively. Speaker-separability results are presented in Table 5.4.

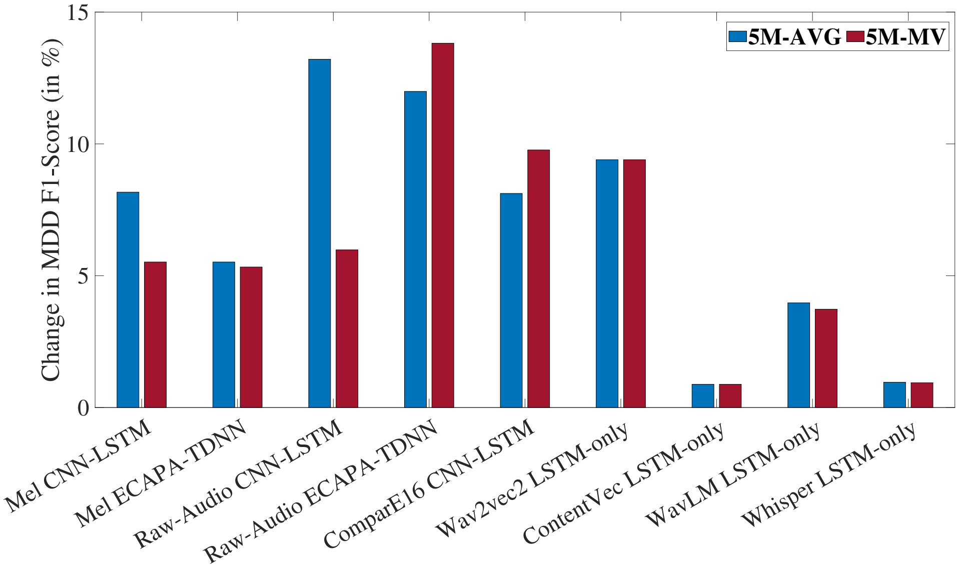
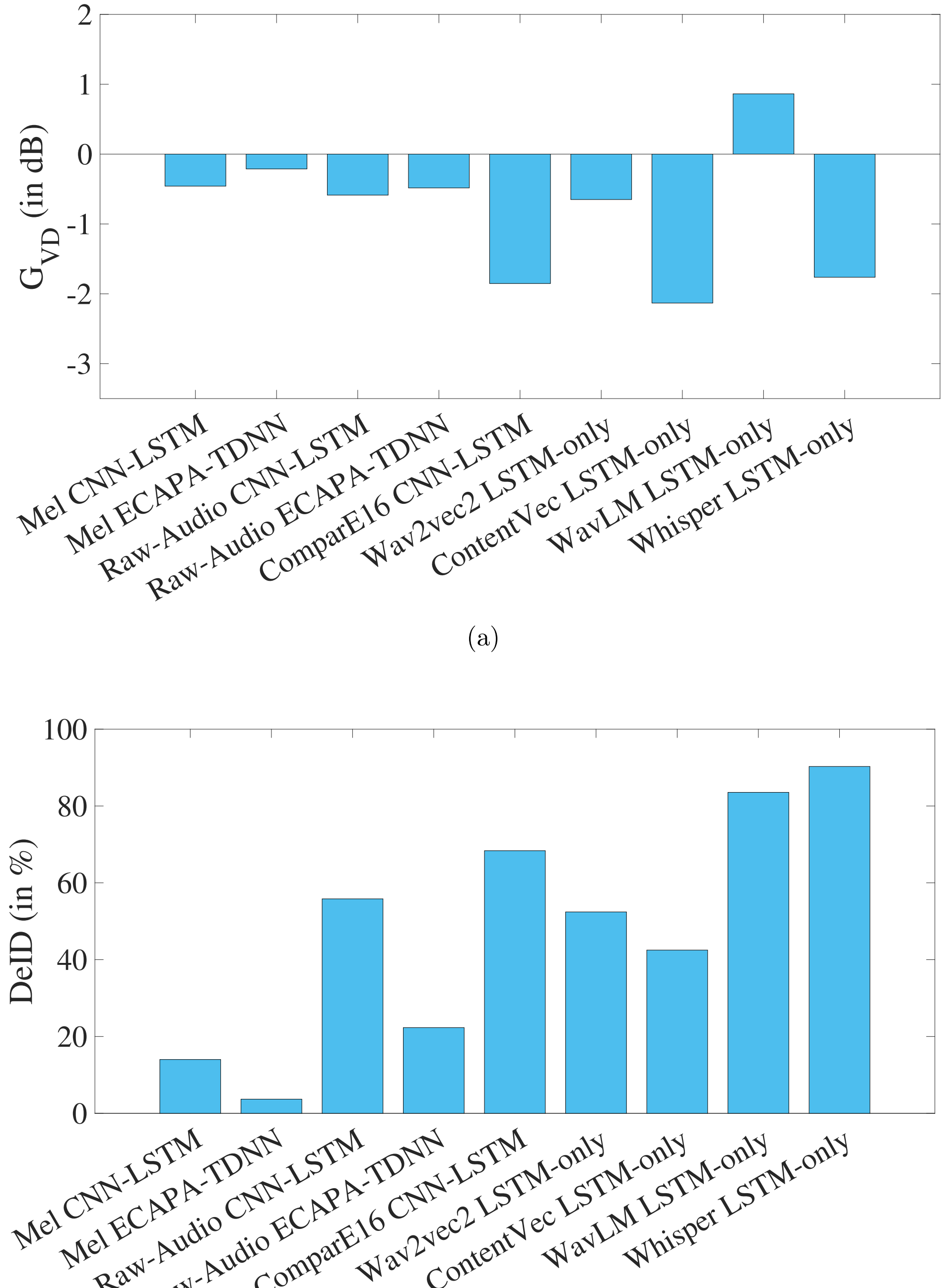


Figure 5.2: Relative improvements, in percentage, in MDD classification F1-Score when speaker disentanglement is applied in the form of ADV. The X-axis of each plot represents the 9 different feature-model combinations. 5M-AVG and 5M-MV refer to the averaging and majority voting aggregation of the 5 models, respectively. A higher value indicates a greater improvement in depression detection after speaker disentanglement is applied.

Across all experiments, it was observed that the MDD F1-AVG score increases when speaker disentanglement is applied, while the G_{VD} is negative in 8 out of 9 scenarios indicating a reduction in speaker separability. On average, over 9 experiments, there was an improvement of 6.53% in MDD F1-AVG. Improvements in MDD detection were statistically significant [64] in 6 out of the 9 experiments (relative change obtained with ComparE16, ContentVec, and Whisper were not statistically significant). Although positive trends were observed in all



(a)

(b)

Figure 5.3: (a) G_{VD} in dB and (b) $De-ID$ in %, respectively, for each experiment when speaker disentanglement is applied in the form of ADV. The X-axis of each plot represents the 9 different feature-model combinations.

experiments, results for Raw-Audio with ECAPA-TDNN, ContentVec with LSTM-only, and WavLM with LSTM-only are selectively discussed below.

The ECAPA-TDNN model is trained with raw-audio signals, and the baseline setup without disentanglement achieves an F1-AVG score of 0.6196 (5M-AVG) and 0.6941 (5M-MV). Recall that 5M-AVG and 5M-MV refer to the averaging and majority voting aggregation of the 5 models, respectively. The best-performing configuration is obtained when adversarial loss maximization is applied to the ECAPA-TDNN model with raw audio signals as input. The F1-AVG increases to 0.6939 (5M-AVG) and 0.7900 (5M-MV). This configuration has a G_{VD} of -0.48 dB which indicates a reduction in speaker separability when ADV is applied and a DeID of 22% that indicates a partially successful masking of speaker identities.

ContentVec with LSTM-only, on the other hand, resulted in smaller improvements when speaker disentanglement was applied. For example, the improvement in F1-AVG is only 0.88% for both 5M-AVG and 5M-MV. Although the improvements in F1-AVG were small, the G_{VD} was -2.13 dB, the lowest among all features with a DeID of 42.5%. It is possible that because ContentVec already includes 3 speaker disentanglement stages, features extracted from it have lost much speaker-identity-related information, and therefore, another disentanglement approach improves depression detection performance only marginally but can severely degrade speaker separability, which is desirable in the context of speaker disentanglement.

In contrast, it was observed that Speaker G_{VD} was negative for all scenarios except WavLM LSTM-only experiments where GvD was 0.863 dB. However, the DeID for WavLM was 83.55% indicating that although the speaker identities were successfully masked when ADV was applied (because of positive DeID), they were still separable (positive G_{VD}).

5.5.2 Extension to EATD and CONVERGE datasets

To evaluate the generalizability of the proposed speaker-disentanglement method to a different language, we applied it to the EATD and the CONVERGE datasets. Results are summarized

Table 5.2: Results, in terms of F1-Score (5 model logit average - 5M-AVG), for speaker disentanglement through ADV using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	5-Models Logit Average						
				F1-Score			Confusion Matrix			
				F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6081	0.6977	0.5185	15	8	5	7
		Yes (α = 4e-5)	293k	0.6578	0.7556	0.5600	17	6	5	7
	Δ (in %)	-	-	8.17	8.30	8.00	-	-	-	-
	ECAPA-TDNN	No	515k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 5e-6)	529k	0.6941	0.7727	0.6154	17	6	4	8
	Δ (in %)	-	-	5.52	2.26	9.89	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6259	0.7755	0.4762	19	4	7	5
		Yes (α = 3e-6)	459k	0.7086	0.8085	0.6087	19	4	5	7
	Δ (in %)	-	-	13.21	4.26	27.82	-	-	-	-
	ECAPA-TDNN	No	595k	0.6196	0.7391	0.5000	17	6	6	6
		Yes (α=2e-4)	609k	0.6939	0.8163	0.5714	20	3	6	6
	Δ (in %)	-	-	11.99	10.45	14.28	-	-	-	-
CompareE16 (384x130)	CNN-LSTM	No	1.15M	0.5791	0.7234	0.4348	17	6	7	5
		Adv (α = 5e-3)	1.18M	0.6261	0.8077	0.4444	21	2	8	4
	Δ (in %)	-	-	8.12	11.65	2.21	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=4e-6)	3.7M	**0.7472**	0.8627	0.6316	22	1	6	6
	Δ (in %)	-	-	9.40	10.24	8.28	-	-	-	-
Contentvec-100 (193x768)	LSTM-only	No	3.6M	0.7287	0.7907	0.6667	17	6	3	9
		Yes (α=1e-2)	3.7M	0.7351	0.7805	0.6897	16	7	2	10
	Δ (in %)	-	-	0.88	-1.29	3.45	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6429	0.7143	0.5714	15	8	4	8
		Yes (α=4e-7)	3.7M	0.6684	0.7442	0.5926	16	7	4	8
	Δ (in %)	-	-	3.97	4.19	3.71	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6438	0.7660	0.5217	18	5	6	6
		Yes (α = 3e-5)	3.4M	0.6500	0.8000	0.5000	20	3	7	5
	Δ (in %)	-	-	0.96	4.44	-4.16	-	-	-	-

Table 5.3: Results, in terms of F1-Score (5 model majority voting - 5M-MV), for speaker disentanglement through ADV using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	5-Models Majority Voting						
				F1-Score			Confusion Matrix			
				F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 4e-5)	293k	0.6941	0.7727	0.6154	17	6	4	8
	Δ (in %)	-	-	5.52	2.26	9.89	-	-	-	-
	ECAPA-TDNN	No	515k	0.7086	0.8085	0.6087	19	4	5	7
		Yes (α = 5e-6)	529k	0.7464	0.8261	0.6667	19	4	4	8
	Δ (in %)	-	-	5.33	2.18	9.53	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6686	0.7917	0.5455	19	4	6	6
		Yes (α = 3e-6)	459k	0.7086	0.8085	0.6087	19	4	5	7
	Δ (in %)	-	-	5.98	2.12	11.59	-	-	-	-
	ECAPA-TDNN	No	595k	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=2e-4)	609k	**0.7900**	0.8800	0.7000	22	1	5	7
	Δ (in %)	-	-	13.82	13.89	13.75	-	-	-	-
CompareE16 (384x130)	CNN-LSTM	No	1.15M	0.6941	0.7727	0.6154	17	6	4	8
		Adv (α = 5e-3)	1.18M	0.7619	0.8571	0.6667	21	2	5	7
	Δ (in %)	-	-	9.77	10.92	8.34	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=4e-6)	3.7M	0.7472	0.8627	0.6316	22	1	6	6
	Δ (in %)	-	-	9.40	10.24	8.28	-	-	-	-
Contentvec-100 (193x768)	LSTM-only	No	3.6M	0.7287	0.7907	0.6667	17	6	3	9
		Yes (α=1e-2)	3.7M	0.7351	0.7805	0.6897	16	7	2	10
	Δ (in %)	-	-	0.88	-1.29	3.45	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=4e-7)	3.7M	0.7200	0.8000	0.6400	18	5	4	8
	Δ (in %)	-	-	3.73	3.53	4.00	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6686	0.7917	0.5455	19	4	6	6
		Yes (α = 3e-5)	3.4M	0.6749	0.8235	0.5263	21	2	7	5
	Δ (in %)	-	-	0.94	4.02	-3.52	-	-	-	-

Table 5.4: Speaker separability results, in terms of G_{VD} (in dB) and $DeID$ (in%), for speaker disentanglement through ADV using the DAIC-WOZ dataset. The best G_{VD} and $DeID$ are bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	G_{VD} (in dB)	DeID in (%)
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	-	-
		Yes (α = 4e-5)	293k	-0.4584	14.01
	ECAPA-TDNN	No	515k	-	-
		Yes (α = 5e-6)	529k	-0.2118	3.69
Raw-Audio (61440x1)	CNN-LSTM	No	445k	-	-
		Yes (α = 3e-6)	459k	-0.5868	55.83
	ECAPA-TDNN	No	595k	-	-
		Yes (α=2e-4)	609k	-0.4843	22.32
ComparE16 (384x130)	CNN-LSTM	No	1.15M	-	-
		Adv (α = 5e-3)	1.18M	-1.8526	68.37
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	-	-
		Yes (α=4e-6)	3.7M	-0.6503	52.43
Contentvec-100 (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=1e-2)	3.7M	**-2.1326**	42.50
WavLM-base (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=4e-7)	3.7M	0.863	83.55
Whisper-base (193x512)	LSTM-only	No	3.4M	-	-
		Yes (α = 3e-5)	3.4M	-1.7630	**90.29**

in Tables 5.5 and 5.6.

Table 5.5: Results, in terms of F1-AVG, Confusion-Matrix, G_{VD} and DeID, for speaker disentanglement through ADV using the development set of EATD dataset. TN, FP, FN, and TP are True Negative, False Positive, False Negative, and True Positive, respectively. The best F1-Score is bold-faced.

				Confusion Matrix					
Feature-Model	Speaker Disentanglement	# Params	F1-AVG	TN	FP	FN	TP	G_{VD} (in dB)	DeID (in %)
Mel-Spectrogram	No	415k	0.5166	58	10	9	2	-	-
CNN-LSTM	ADV ($\alpha = 2e - 4$)	430k	0.5756	56	12	7	4	-1.3478	63.49
Raw-Audio	No	445k	0.643	62	6	7	4	-	-
CNN-LSTM	ADV ($\alpha = 4e - 5$)	456k	**0.720**	62	6	5	6	-0.8827	51.21

For the EATD dataset, the results demonstrate that applying ADV to the CNN-LSTM model trained on Mel-Spectrograms leads to a 5.9% increase in F1-AVG, from 0.5166 to 0.5766, with a G_{VD} of -1.3478dB and a DeID of 63.49%. When Raw-audio is used as the input feature, similar improvements in depression detection are observed. The F1-AVG for MDD prediction increases by 11.99%, from 0.6430 for the baseline model to 0.7201 for the proposed method ($\lambda = 3e - 5$), with a G_{VD} of -0.8827dB and a DeID of 51.21%. These findings suggest that speaker-identity-related information poses a significant challenge across multiple datasets, and our proposed method demonstrates the potential to effectively mitigate these issues. Recall that for the EATD dataset, the evaluations were performed using the development splits as provided with the dataset description (see Section 2.1 for details)

For the CONVERGE dataset, the results show improvements in depression detection performance, but the privacy attribute enhancements are limited. When using Mel-Spectrograms, the depression detection F1-AVG increases from 0.879 to 0.890, but the G_{VD} is only -0.12 dB, with a DeID score of 17.53%. Similarly, when RAW-Audio signals are used as input, the F1-AVG improves from 0.829 to 0.857%, but the G_{VD} is a mere -0.0147 dB, and the DeID is only 4.13%. The homogeneity of the dataset, which consists of only female speakers and severely depressed participants, may contribute to the limited effectiveness of the speaker disentanglement methods in this case. It is possible that the characteristics of this dataset hinder its ability to fully benefit from the proposed approach.

Table 5.6: Results, in terms of F1-AVG, Confusion-Matrix, G_{VD} and DeID, speaker disentanglement through ADV using the test set of the CONVERGE dataset. TN, FP, FN, and TP are True Negative, False Positive, False Negative, and True Positive, respectively. The best F1-Score is bold-faced.

				Confusion Matrix					
Feature-Model	Speaker Disentanglement	# Params	F1-AVG	TN	FP	FN	TP	G_{VD} (in dB)	DeID (in %)
Mel-Spectrogram	No	415k	0.879	475	139	125	446	-	-
CNN-LSTM	ADV ($\alpha = 2e - 4$)	430k	**0.890**	512	102	94	477	-0.1217	17.53
Raw-Audio	No	445k	0.829	443	131	147	424	-	-
CNN-LSTM	ADV ($\alpha = 3e - 5$)	456k	0.857	456	144	132	439	-0.0147	4.13%

5.6 Chapter Summary

Rewritten paragraph: In the field of depression detection, features such as x-vectors and i-vectors have proven to be valuable. However, despite their effectiveness, these features also carry information about the speaker's identity, which can raise concerns about privacy in the context of a major depressive disorder (MDD) diagnosis system. Excessive reliance on speaker-identity features may hinder the adoption of speech-based assessment methods due to privacy considerations.

To address this challenge, we propose an adversarial disentanglement approach in this chapter, aiming to separate speaker identity from depression status. Our method demonstrates that speaker-identity invariant models can enhance MDD classification performance across various features and multi-lingual datasets. When applied to the English DAIC-WoZ dataset, our approach yields a 13.82% improvement over the baseline when using an ECAPA-TDNN model trained on Raw-Audio signals. Similar improvements are observed across all seven model-feature combinations, along with a reduction in speaker separability, as measured by the Gain of Voice Distinctiveness (G_{VD}) and De-Identification ($DeID$) metrics.

Furthermore, we show that our method generalizes to datasets in another language, specifically Mandarin, through experiments on the EATD and CONVERGE datasets. On the EATD dataset, an 11.99% improvement is achieved when using Raw-Audio with a CNN-LSTM model, resulting in an F1-Avg of 0.72. For the CONVERGE dataset, using Mel-Spectrograms with a CNN-LSTM model leads to an F1-Avg of 0.890, compared to the baseline F1 score of 0.879, representing a 1.25% improvement. These results demonstrate that by mitigating the influence of speaker-specific characteristics, our approach emphasizes the essential features related to depression, ultimately enhancing diagnostic performance.

Chapter 6

Speaker Disentanglement via Loss Equalization

6.1 Introduction

Speaker information can hinder the privacy attribute of a depression detection system and, in some cases, act as a bias factor, leading to incorrect decision-making. To address this challenge, Chapter 5 proposed a speaker-identification (SID) loss-maximization approach using an adversarial training mechanism. While such loss maximization approaches have been widely used in speech-related tasks, the adversarial SID loss is unbounded due to the log-function in Eq.5.2, which can sometimes result in poor model convergence [112]. Furthermore, during cross-entropy loss optimization in the SID branch, as shown in Eq. 5.2, only the probability of the specific speaker $\hat{n}$ corresponding to that sample x_n is considered. In other words, the numerator contains only the probability of the target speaker, while the denominator uses probabilities for all speakers. Since the denominator serves as a normalizer, its value is shared across all samples. As a result, the numerator becomes the primary contributor to the loss term for the given target speaker, leaving the other probabilities unused, which can limit the potential for disentangling speaker information.

To address the limitations of adversarial loss maximization, this chapter introduces loss equalization-based approaches. Three variants of loss equalization are proposed: loss equalization with variance, loss equalization with cross-entropy, and loss equalization with KL (Kullback–Leibler) divergence. We start with an L2-regularization-based approach in loss equalization with variance followed by classification-loss-based approach in loss equalization with cross-entropy, and finally a probability-distribution matching approach in loss equalization with KLD. The main idea behind these methods is to equalize the loss contributions across all speaker coordinates, ensuring that the model does not rely on any specific speaker information for depression detection. By doing so, the proposed approaches aim to effectively disentangle speaker information from the depression-related features, leading to a more privacy-preserving and unbiased depression detection system. Experiments are conducted across various settings to evaluate the performance of the proposed loss equalization methods and compare them with their adversarial counterpart. The voice-privacy attributes are measured using $DeID$ and G_{VD} as was done in Chapter 5. The results demonstrate that the loss equalization approaches can achieve comparable or even better performance than the adversarial loss maximization method, highlighting their effectiveness in addressing the challenges associated with speaker information in depression detection systems.

6.1.1 SID-loss Equalization with Variance

To overcome the limitations of adversarial loss maximization, a loss equalization-based approach is proposed. Instead of forcing the model to make wrong predictions about speaker identity, equalization methods tend to confuse the model so that it is unable to distinguish speaker classes through a uniform regularization process similar to an L_2 norm. The equalization loss is formulated as follows:

$$L_{Evar} = \frac{1}{N}\sum_{n=1}^{N}[||\sigma(x_n) - e||^2] \tag{6.1}$$

where $e = [1/M, 1/M, ..., 1/M]$ is the vector that assigns equal probability to each speaker in a uniform manner, with length M and x_n is the M-dimensional output logit obtained from the model and σ is the softmax function to convert logits to probabilities. The number of samples in a training batch is denoted as N. Since the new loss term is meant to be minimized, the objective function is defined as follows:

$$L_{total\_Evar} = L_{MDD} + \lambda(L_{Evar}), \tag{6.2}$$

In the initial experiments using Eq. 6.1, it was observed that the model learned to predict the e-vector very easily within a few epochs without learning to disentangle speakers i.e., the speaker prediction branch was overfitting to directly predict the e-vector without tangible speaker disentanglement. We refer to this situation as the "trivial" solution. To avoid this scenario, additive noise ($U(0, 1)$) is injected into the vector e. This method is referred to as Loss equalization with Variance (LEV) in further sections.

6.1.2 SID-loss Equalization with Cross-Entropy

In LEV, loss-equalization is achieved via the L_2 loss. Alternatively, loss-equalization can also be achieved by minimizing the Cross-Entropy loss between the speaker prediction probabilities and a ones-vector of the same dimension. Mathematically, the equalization loss can be formulated as:

$$L_{Ece} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}[y_{n,m} \cdot \log\sigma(x_{n,m}) + (1 - y_{n,m}) \cdot \log(1 - \sigma(x_{n,m}))] \tag{6.3}$$

Where $y_n = [1, 1, ..., 1]$ is the M-dimensional target vector and x_n is the M-dimensional output logits of the models for the n^{th} sample, respectively. σ is a Softmax function to convert logits to probabilities. Since $y_{n,m} = 1$ for all n, m, the above equation can be simplified as -

$$L_{Ece} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}[\log\sigma(x_{n,m})] \tag{6.4}$$

Therefore, the total loss can be written as -

$$L_{total\_ECE} = L_{MDD} + \lambda(L_{Ece}), \tag{6.5}$$

This method is referred to as Loss equalization with Cross-Entropy (LECE) in further sections.

6.1.3 SID-loss Equalization with KL Divergence

Another approach to achieve speaker disentanglement is by manipulating the distribution of the SID-prediction logits. We hypothesize that a uniform distribution for SID logits can help in disentangling speaker identity and MDD characteristics. To achieve this, we propose to minimize the KL-divergence loss between the normalized predicted logits and a uniform vector e. We denote this method as LEKLD in the following sections. The KL-divergence based equalization loss is formulated as:

$$\begin{aligned} L_{EKL} &= L_{KL}(x, e) \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} e_m \cdot (\log(e_m) - \log(\sigma(x_{n,m}))) \end{aligned} \tag{6.6}$$

where $x_{n,m}$ and e_m stand for the m^{th} element in predicted logits x_n and uniform vector e, respectively and σ is the Softmax function. Thus, the final loss with KL-divergence term is computed as:

$$L_{total\_EKL} = L_{MDD} + \lambda(L_{EKL}), \tag{6.7}$$

6.2 Experimental Details

In this chapter, we conducted experiments using two datasets: DAIC-WOZ (English) [104], and the EATD [96]. Unlike the experiments in Chapter 5, the CONVERGE dataset was not included in the evaluations for this chapter. This decision was made based on the observations reported in Chapter 5, where minimal gains were achieved when applying speaker disentanglement techniques to the CONVERGE dataset. The limited improvements in performance and privacy metrics on this dataset suggest that the speaker information may not be a significant factor in the depression detection task for this particular dataset. This could be attributed to the homogeneous nature of the CONVERGE dataset, which consists of only female speakers and severely depressed participants. As a result, focusing on the DAIC-WOZ and EATD datasets, which showed more promising results with speaker disentanglement, allows for a more targeted evaluation of the proposed loss equalization methods and their effectiveness in improving depression detection while preserving speaker privacy.

In terms of the backend model, we employed three different backend models: 1) a modified version of DepAudioNet [61], 2) an ECAPA-TDNN model, and 3) an LSTM-only model. For the English dataset, we evaluated the performance using all seven input features, while for the Mandarin dataset, we focused on Mel-Spectrogram and Raw-Audio signals only. All model parameters such as number of layers, number of channels of convolutional filters, stride, kernel size, etc were the same as those described in Section 5.4

6.3 Results and Discussion

The results section is organized in a similar way to that used in Chapter 5 and is as follows - first, we present the results of the DAIC-WOZ dataset and compare the performance of the proposed loss-equalization based speaker disentanglement methods with the baseline approaches that do not employ disentanglement. Next, we compare the performance of loss-equalization methods with the adversarial method proposed in Chapter 5. Lastly, the proposed method is extended to the EATD dataset to assess its generalizability and effectiveness across different languages and domains. Unless otherwise specified, the reported relative improvements are statistically significant, as determined by the McNemar's test [64].

6.3.1 Loss Equalization with Variance (LEV)

Figures 6.1 and 6.2 show the relative improvement in MDD classification F1-AVG, and the speaker separability metrics G_{VD} (in dB) and $De - ID$ (in %) for each model-feature combination when LEV is applied, respectively. Detailed results, in terms of F1-Score for 5M-AVG and 5M-MV are presented in Tables 6.1 and 6.2, respectively. Speaker-separability results are presented in Table 6.3. Similar to ADV, this loss function results in improvements in MDD detection across all 9 experiments, with an average improvement in F1-AVG of 6.69%. This is accompanied by a negative G_{VD} in 8 out of the 9 experiments. Improvements in MDD detection were statistically significant [64] in 5 out of the 9 experiments (relative change obtained with Raw-Audio, ComparE16, ContentVec, and Whisper were not statistically significant). For LEV, we discuss results from Wav2vec2 LSTM-only, ComparE16 CNN-LSTM, Whisper LSTM-only, and WavLM LSTM-only, which considers models with the best overall performance, highest improvement, lowest G_{VD} and highest $DeID$, respectively.

In the case of LEV, Wav2vec2 features with LEV result in the best MDD classification performance. For the baseline model without disentanglement, the F1-AVG scores are 0.6830

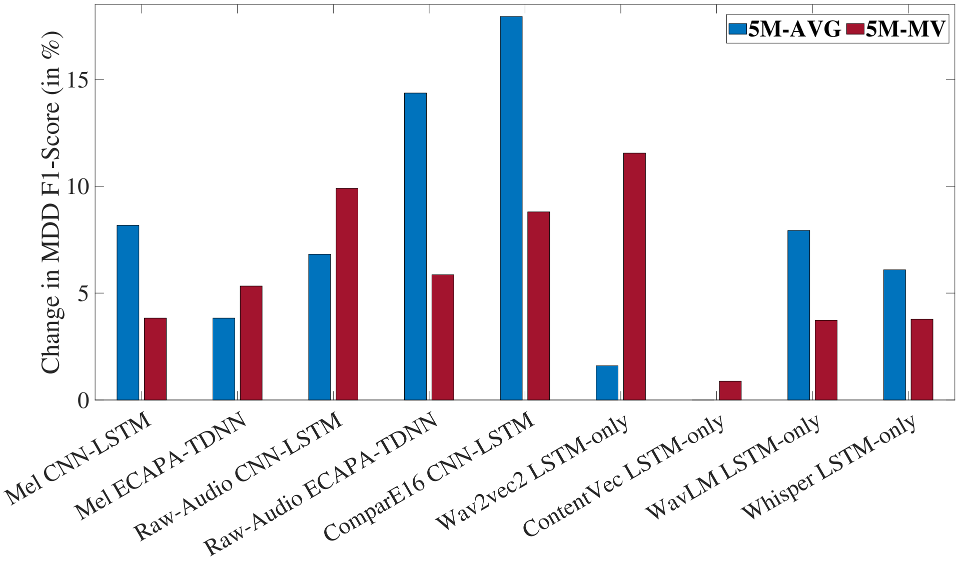
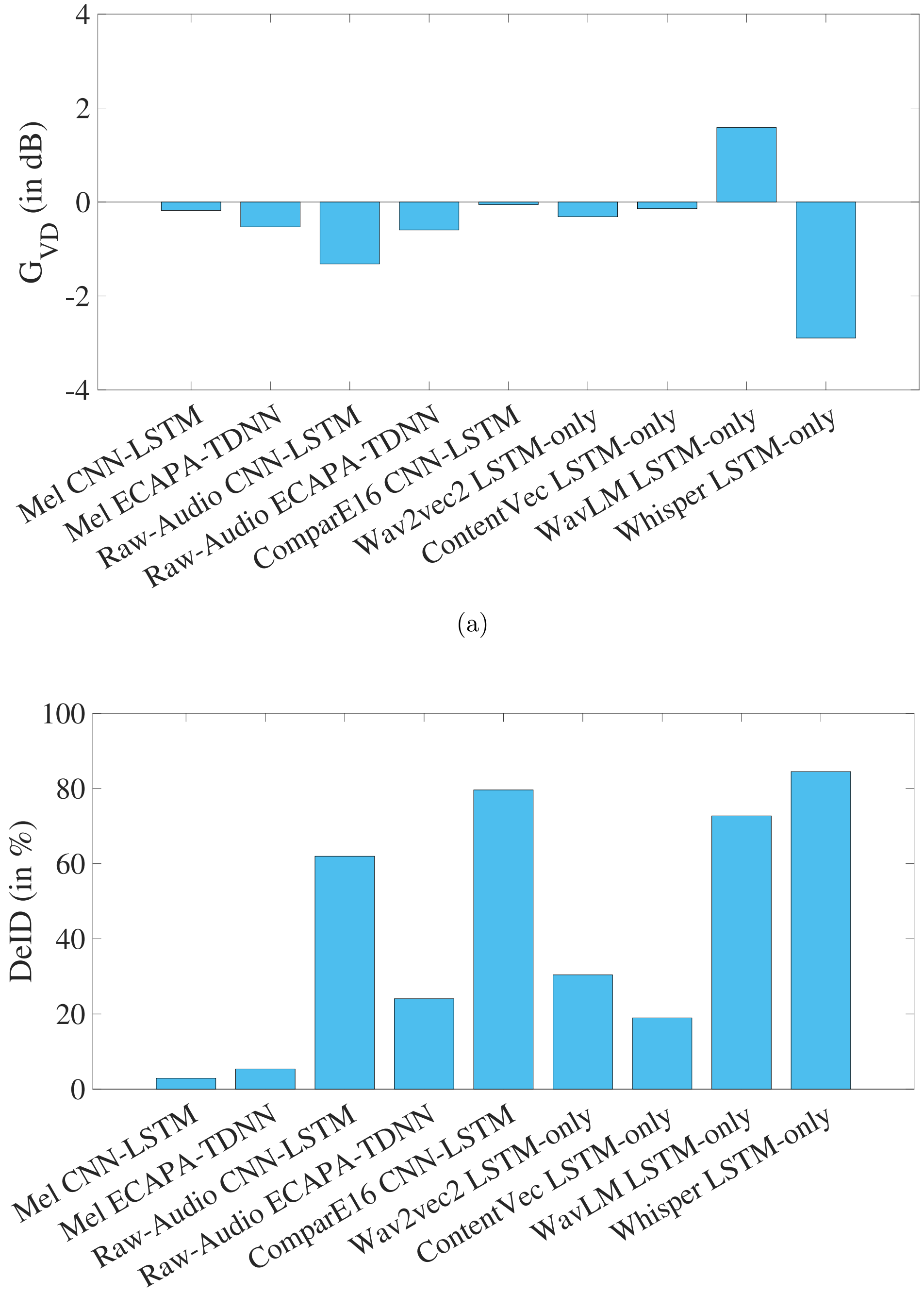


Figure 6.1: Relative improvements, in percentage, in MDD classification F1-Score when speaker disentanglement is applied in the form of LEV. The X-axis of each plot represents the 9 different feature-model combinations. 5M-AVG and 5M-MV refer to the averaging and majority voting aggregation of the 5 models, respectively. A higher value indicates a greater improvement in depression detection after speaker disentanglement is applied.



(a)

(b)

Figure 6.2: (a) G_{VD} in dB and (b) $De - ID$ in %, respectively, for each experiment when speaker disentanglement is applied in the form of LEV. The X-axis of each plot represents the 9 different feature-model combinations.

(5M-AVG) and 0.6830 (5M-MV). When the proposed method with a hyperparameter value of $\lambda = 5e - 3$ is applied, the F1-AVG increases to 0.6939 (5M-AVG) and 0.7619 (5M-MV). For this case, the G_{VD} is -0.3126 dB, and the DeID is 30.41%. A negative G_{VD} further shows that the disentangled speaker representations are less separable than their baseline counterparts.

The highest improvements in MDD detection are observed when CompareE16 features are used, with a 17.94% increase in F1-AVG (5M-AVG), from 0.5791 for the baseline to 0.683 for the proposed method ($\lambda = 2e - 4$). The G_{VD} for this model is -0.0551 dB whereas the DeID is 79.62% indicating a successful speaker-identity masking mechanism but only a small reduction in speaker-separability.

The lowest G_{VD} of -2.589 dB is achieved in LEV when Whisper-base features are used with the LSTM-only model. This feature also has a high DeID of 84.49%. Lastly, similar to ADV, WavLM with LSTM-only and LEV results in the highest G_{VD} of 1.5854 dB but has a DeID of 72.71%. Although this points to a (partially) successful speaker-identity masking method, the disentangled speaker representations are more separable than the baseline embeddings.

6.3.2 Loss Equalization with Cross-Entropy (LECE)

Figures 6.3 and 6.4 show the relative improvement in MDD classification F1-AVG, and the speaker separability metrics G_{VD} (in dB) and $De - ID$ (in %) for each model-feature combination when LECE is applied, respectively. Detailed results, in terms of F1-Score for 5M-AVG and 5M-MV are presented in Tables 6.4 and 6.5, respectively. Speaker-separability results are presented in Table 6.6. Similar to ADV and LEV, this loss function results in improvements in MDD detection across all 9 experiments, with an average improvement in F1-AVG of 8.86%. Unlike the results reported for LEV in Section 6.3.1, a negative G_{VD} is observed in all 9 experiments. Improvements in MDD detection were statistically significant [64] in 4 out of the 9 experiments (relative change obtained with Raw-Audio, ContentVec, WavLM, and Whisper were not statistically significant). Similar to LEV, we

Table 6.1: Results, in terms of F1-Score (5 model logit average - 5M-AVG), for speaker disentanglement through LEV using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	5-Models Logit Average						
				F1-Score			Confusion Matrix			
				F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6081	0.6977	0.5185	15	8	5	7
		Yes (α = 5e-5)	293k	0.6578	0.7556	0.5600	17	6	5	7
	Δ (in %)	-	-	8.17	8.30	8.00	-	-	-	-
	ECAPA-TDNN	No	515k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 5e-2)	529k	0.6830	0.7826	0.5833	18	5	5	7
	Δ (in %)	-	-	3.83	3.57	4.16	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6259	0.7755	0.4762	19	4	7	5
		Yes (α = 1e-3)	459k	0.6686	0.7917	0.5455	19	4	6	6
	Δ (in %)	-	-	6.82	2.09	14.55	-	-	-	-
	ECAPA-TDNN	No	595k	0.6196	0.7391	0.5000	17	6	6	6
		Yes (α=3e-3)	609k	0.7086	0.8085	0.6087	19	4	5	7
	Δ (in %)	-	-	14.36	9.39	21.74	-	-	-	-
CompарE16 (384x130)	CNN-LSTM	No	1.15M	0.5791	0.7234	0.4348	17	6	7	5
		Yes (α = 2e-4)	1.18M	0.6830	0.7826	0.5833	18	5	5	7
	Δ (in %)	-	-	17.94	8.18	34.15	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=5e-3)	3.7M	0.6939	0.8163	0.5714	20	3	6	6
	Δ (in %)	-	-	1.60	4.31	-2.04	-	-	-	-
ContentVec-100 (193x768)	LSTM-only	No	3.6M	**0.7287**	0.7907	0.6667	17	6	3	9
		Yes (α=2e-2)	3.7M	0.7287	0.7907	0.6667	17	6	3	9
	Δ (in %)	-	-	0.00	0.00	0.00	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6429	0.7143	0.5714	15	8	4	8
		Yes (α=2e-3)	3.7M	0.6939	0.8163	0.5714	20	3	6	6
	Δ (in %)	-	-	7.93	14.28	0.00	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6438	0.7660	0.5217	18	5	6	6
		Adv (α = 5e-3)	3.4M	0.6830	0.7826	0.5833	18	5	5	7
	Δ (in %)	-	-	6.09	2.17	11.81	-	-	-	-

Table 6.2: Results, in terms of F1-Score (5 model majority voting - 5M-MV), for speaker disentanglement through LEV using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	5-Models Majority Voting						
				F1-Score			Confusion Matrix			
				F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 5e-5)	293k	0.6830	0.7826	0.5833	18	5	5	7
	Δ (in %)	-	-	3.83	3.57	4.16	-	-	-	-
	ECAPA-TDNN	No	515k	0.7086	0.8085	0.6087	19	4	5	7
		Yes (α = 5e-2)	529k	0.7464	0.8261	0.6667	19	4	4	8
	Δ (in %)	-	-	5.33	2.18	9.53	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6686	0.7917	0.5455	19	4	6	6
		Yes (α = 1e-3)	459k	0.7348	0.8333	0.6364	20	3	5	7
	Δ (in %)	-	-	9.90	5.25	16.66	-	-	-	-
	ECAPA-TDNN	No	595k	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=3e-3)	609k	0.7348	0.8333	0.6364	20	3	5	7
	Δ (in %)	-	-	5.86	7.84	3.41	-	-	-	-
ComparE16 (384x130)	CNN-LSTM	No	1.15M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α = 2e-4)	1.18M	0.7552	0.8182	0.6923	18	5	3	9
	Δ (in %)	-	-	8.80	5.89	12.50	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=5e-3)	3.7M	**0.7619**	0.8517	0.6667	21	2	5	7
	Δ (in %)	-	-	11.55	8.83	14.30	-	-	-	-
ContentVec-100 (193x768)	LSTM-only	No	3.6M	0.7287	0.7907	0.6667	17	6	3	9
		Yes (α=2e-2)	3.7M	0.7351	0.7805	0.6897	16	7	2	10
	Δ (in %)	-	-	0.88	-1.29	3.45	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=2e-3)	3.7M	0.7200	0.8400	0.6000	21	2	6	6
	Δ (in %)	-	-	3.73	8.71	-2.50	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6686	0.7917	0.5455	19	4	6	6
		Adv (α = 5e-3)	3.4M	0.6939	0.8163	0.5714	20	3	6	6
	Δ (in %)	-	-	3.78	3.11	4.75	-	-	-	-

Table 6.3: Speaker separability results, in terms of G_{VD} (in dB) and $DeID$ (in%), for speaker disentanglement through LEV using the DAIC-WOZ dataset. The best G_{VD} and $DeID$ are bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	G_{VD} (in dB)	DeID in (%)
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	-	-
		Yes (α = 5e-5)	293k	-0.1787	2.90
	ECAPA-TDNN	No	515k	-	-
		Yes (α = 5e-2)	529k	-0.5296	5.37
Raw-Audio (61440x1)	CNN-LSTM	No	445k	-	-
		Yes (α = 1e-3)	459k	-1.3189	61.98
	ECAPA-TDNN	No	595k	-	-
		Yes (α=3e-3)	609k	-0.5953	24.06
CompareE16 (384x130)	CNN-LSTM	No	1.15M	-	-
		Yes (α = 2e-4)	1.18M	-0.0551	79.62
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	-	-
		Yes (α=5e-3)	3.7M	- 0.3126	30.41
ContentVec-100 (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=2e-2)	3.7M	-0.1416	18.50
WavLM-base (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=2e-3)	3.7M	1.5854	72.71
Whisper-base (193x512)	LSTM-only	No	3.4M	-	-
		Adv (α = 5e-3)	3.4M	**-2.8950**	**84.49**

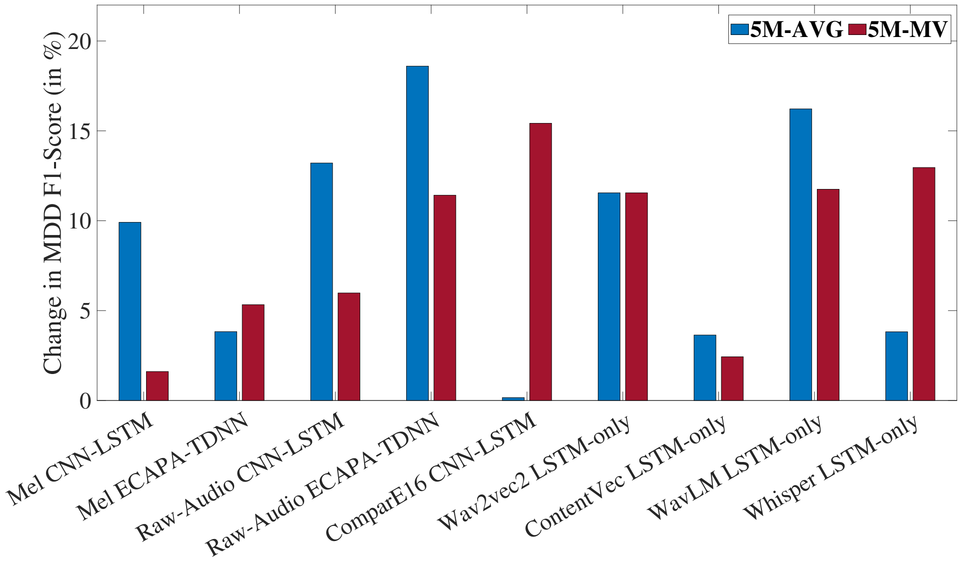
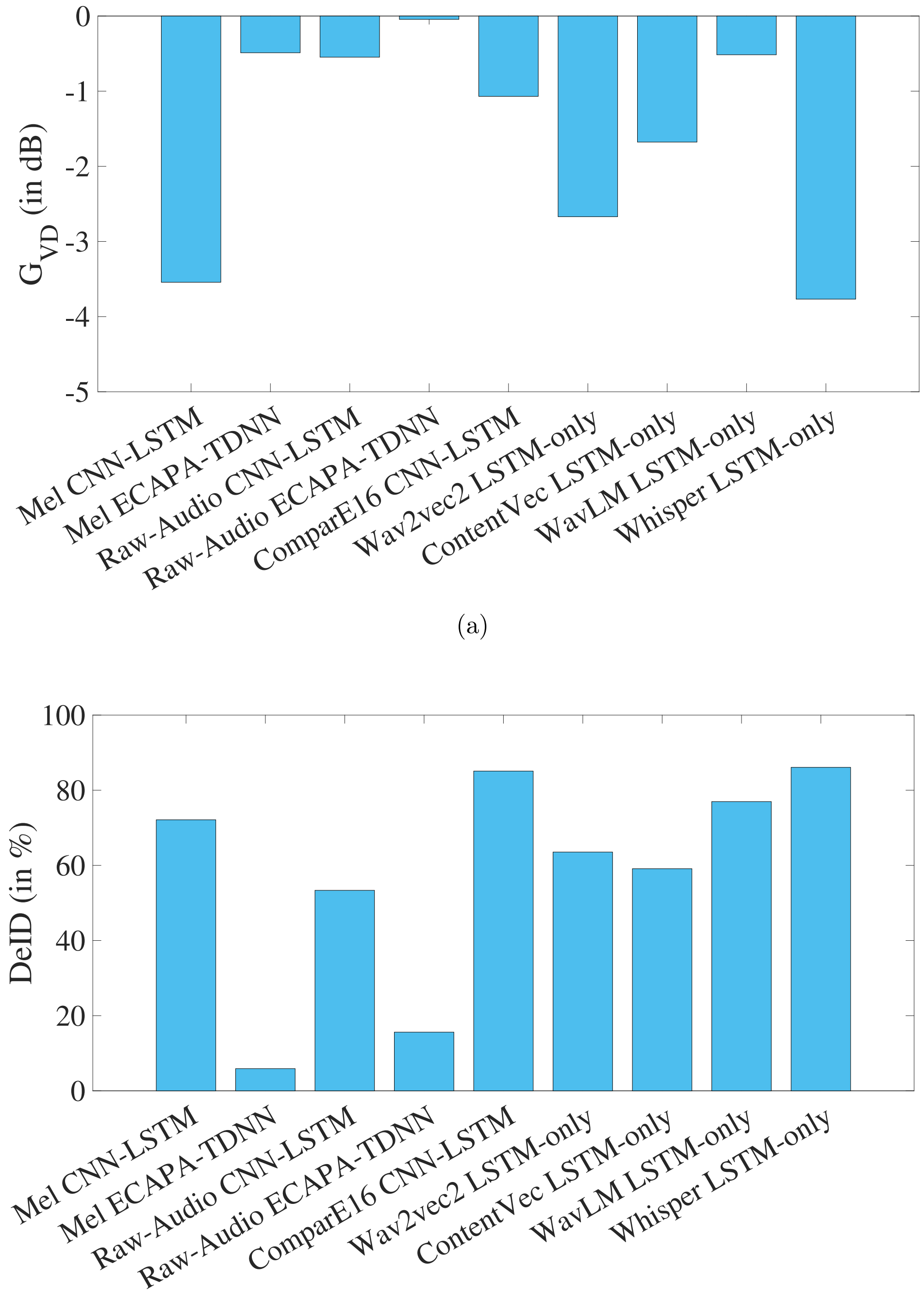


Figure 6.3: Relative improvements, in percentage, in MDD classification F1-Score when speaker disentanglement is applied in the form of LECE. The X-axis of each plot represents the 9 different feature-model combinations. 5M-AVG and 5M-MV refer to the averaging and majority voting aggregation of the 5 models, respectively. A higher value indicates a greater improvement in depression detection after speaker disentanglement is applied.

specifically discuss results for models with best overall performance, highest improvement, lowest G_{VD} and highest $DeID$. For LECE, we discuss results from CompareE16 CNN-LSTM, Raw-Audio ECAPA-TDNN, and Whisper LSTM-only.

CompareE16 features when used with CNN-LSTM features achieved the best MDD classification performance. In the baseline model without disentanglement, the F1-AVG scores are 0.5791(5M-AVG) and 0.6941 (5M-MV). When the proposed method with a hyperparameter value of $\lambda = 1e-7$ is applied, the F1-AVG increases to 0.5800 (5M-AVG) and 0.8011 (5M-MV). For this case, the G_{VD} is -1.0688 dB, and the DeID is 85.10%. A negative G_{VD} along with a high DeID shows that identity has been successfully masked and that the disentangled speaker representations are less separable than the corresponding baseline representations.

The highest improvement in F1-Score is observed when Raw-Audio signals are used to



(a)

(b)

Figure 6.4: (a) G_{VD} in dB and (b) $De - ID$ in %, respectively, for each experiment when speaker disentanglement is applied in the form of LECE. The X-axis of each plot represents the 9 different feature-model combinations.

train the ECAPA-TDNN model. The baseline F1-AVG score improves by 18.60%, from 0.6196 (5M-AVG) to 0.7348. his feature-model combination has the highest G_{VD} of -0.0446 dB with a corresponding DeID of 15.62%. Although this is the highest value for G_{VD} in the LECE experiments, since it is still less than zero the speaker representations after disentanglement are less separable than the corresponding baseline representations.

Similar to LEV, Whisper-base features with the LSTM model resulted in the lowest G_{VD} of -3.767 dB and a DeID of 86.09%.

6.3.3 Loss Equalization with KLD (LEKLD)

Figures 6.5 and 6.6 show the relative improvement in MDD classification F1-AVG, and the speaker separability metrics G_{VD} (in dB) and $De - ID$ (in %) for each model-feature combination when ADV is applied, respectively. Detailed results, in terms of F1-Score for 5M-AVG and 5M-MV are presented in Tables 6.7 and 6.8, respectively. Speaker-separability results are presented in Table 6.9. As seen before in ADV, LEV, and LECE, every experiment leads to an improvement in MDD detection performance with an average improvement in MDD F1-AVG by 7.07% and a negative G_{VD} is 8 out of 9 experiments. Improvements in MDD detection were statistically significant [64] in 7 out of the 9 experiments (relative change obtained with ComparE16 and ContentVec were not statistically significant). For this method, we discuss the results from Whisper LSTM-only, Raw-Audio ECAPA-TDNN, WavLM LSTM-only, and ComparE16 CNN-LSTM to cover models representing the best overall performance, the highest improvement, the lowest G_{VD} and the highest $DeID$.

In the case of LEKLD, the best-performing model is the Whisper LSTM-only model with speaker disentanglement. The baseline F1-AVG of 0.6438 (5M-AVG), 0.6686 (5M-MV) increases by 6.09% and 18.16% to 0.6830 (5M-AVG), 0.7900 (5M-MV), respectively when the proposed method is applied ($\lambda = 1e - 5$). For this model-feature combination, the G_{VD} is -3.93 dB and the corresponding DeID is 69.42%.

Table 6.4: Results, in terms of F1-Score (5 model logit average - 5M-AVG), for speaker disentanglement through LECE using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	5-Models Logit Average						
				F1-Score			Confusion Matrix			
				F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6081	0.6977	0.5185	15	8	5	7
		Yes (α = 4e-1)	293k	0.6684	0.7442	0.5926	16	7	4	8
	Δ (in %)	-	-	9.91	6.66	14.29	-	-	-	-
	ECAPA-TDNN	No	515k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 2e-7)	529k	0.6830	0.7826	0.5833	18	5	5	7
	Δ (in %)	-	-	3.83	3.57	4.17	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6259	0.7755	0.4762	19	4	7	5
		Yes (α = 4e-5)	459k	0.7086	0.8085	0.6087	19	4	5	7
	Δ (in %)	-	-	13.21	4.26	27.82	-	-	-	-
	ECAPA-TDNN	No	595k	0.6196	0.7391	0.5000	17	6	6	6
		Yes (α=3e-5)	609k	0.7348	0.8333	0.6364	20	3	5	7
	Δ (in %)	-	-	18.60	12.75	27.27	-	-	-	-
CompareE16 (384x130)	CNN-LSTM	No	1.15M	0.5791	0.7234	0.4348	17	6	7	5
		Yes (α = 1e-7)	1.18M	0.5800	0.7600	0.4000	19	4	8	4
	Δ (in %)	-	-	0.16	5.06	-8.00	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=5e-5)	3.7M	**0.7619**	0.8571	0.6667	21	2	5	7
	Δ (in %)	-	-	11.55	9.53	14.29	-	-	-	-
ContentVec-100 (193x768)	LSTM-only	No	3.6M	0.7287	0.7907	0.6667	17	6	3	9
		Yes (α=5e-4)	3.7M	0.7552	0.8182	0.6923	18	5	3	9
	Δ (in %)	-	-	3.64	3.48	3.84	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6429	0.7143	0.5714	15	8	4	8
		Yes (α=2e-2)	3.7M	0.7472	0.8627	0.6316	22	1	6	6
	Δ (in %)	-	-	16.22	20.78	10.53	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6438	0.7660	0.5217	18	5	6	6
		Adv (α = 5e-6)	3.4M	0.6684	0.7442	0.5926	16	7	4	8
	Δ (in %)	-	-	3.82	-2.85	13.59	-	-	-	-

Table 6.5: Results, in terms of F1-Score (5 model majority voting - 5M-MV), for speaker disentanglement through LECE using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input				5-Models Majority Voting						
Feature	Model	Speaker	Model	F1-Score			Confusion Matrix			
(Seq.len x Num. of Features)	Architecture	Disentanglement	Parameters	F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 4e-1)	293k	0.6684	0.7442	0.5926	16	7	4	8
	Δ (in %)	-	-	1.61	-1.51	5.82	-	-	-	-
	ECAPA-TDNN	No	515k	0.7086	0.8085	0.6087	19	4	5	7
		Yes (α = 2e-7)	529k	0.7464	0.8261	0.6667	19	4	4	8
	Δ (in %)	-	-	5.33	2.18	9.52	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6686	0.7917	0.5455	19	4	6	6
		Yes (α = 4e-5)	459k	0.7086	0.8085	0.6087	19	4	5	7
	Δ (in %)	-	-	5.98	2.12	11.58	-	-	-	-
	ECAPA-TDNN	No	595k	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=3e-5)	609k	0.7734	0.8511	0.6957	20	3	4	8
	Δ (in %)	-	-	11.42	10.14	13.04	-	-	-	-
CompareE16 (384x130)	CNN-LSTM	No	1.15M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α = 1e-7)	1.18M	**0.8011**	0.8750	0.7273	21	2	4	8
	Δ (in %)	-	-	15.42	13.24	18.18	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=5e-5)	3.7M	0.7619	0.8571	0.6667	21	2	5	7
	Δ (in %)	-	-	11.55	9.53	14.29	-	-	-	-
ContentVec-100 (193x768)	LSTM-only	No	3.6M	0.7287	0.7907	0.6667	17	6	3	9
		Yes (α=5e-4)	3.7M	0.7464	0.8261	0.6667	19	4	4	8
	Δ (in %)	-	-	2.43	4.48	0.00	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=2e-2)	3.7M	0.7756	0.8846	0.6667	23	0	6	6
	Δ (in %)	-	-	11.75	14.48	8.33	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6686	0.7917	0.5455	19	4	6	6
		Adv (α = 5e-6)	3.4M	0.7552	0.8182	0.6923	18	5	3	9
	Δ (in %)	-	-	12.96	3.34	26.91	-	-	-	-

Table 6.6: Speaker separability results, in terms of G_{VD} (in dB) and $DeID$ (in%), for speaker disentanglement through LECE using the DAIC-WOZ dataset. The best G_{VD} and $DeID$ are bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	G_{VD} (in dB)	DeID in (%)
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	-	-
		Yes (α = 4e-1)	293k	-3.5427	72.13
	ECAPA-TDNN	No	515k	-	-
		Yes (α = 2e-7)	529k	-0.489	5.91
Raw-Audio (61440x1)	CNN-LSTM	No	445k	-	-
		Yes (α = 4e-5)	459k	-0.5476	53.56
	ECAPA-TDNN	No	595k	-	-
		Yes (α=3e-5)	609k	-0.0446	15.62
CompareE16 (384x130)	CNN-LSTM	No	1.15M	-	-
		Yes (α = 1e-7)	1.18M	-1.0668	85.10
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	-	-
		Yes (α=5e-5)	3.7M	-2.6701	63.55
ContentVec-100 (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=5e-4)	3.7M	-1.6775	59.11
WavLM-base (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=2e-2)	3.7M	-0.5155	76.98
Whisper-base (193x512)	LSTM-only	No	3.4M	-	-
		Adv (α = 5e-6)	3.4M	**-3.7670**	**86.09**

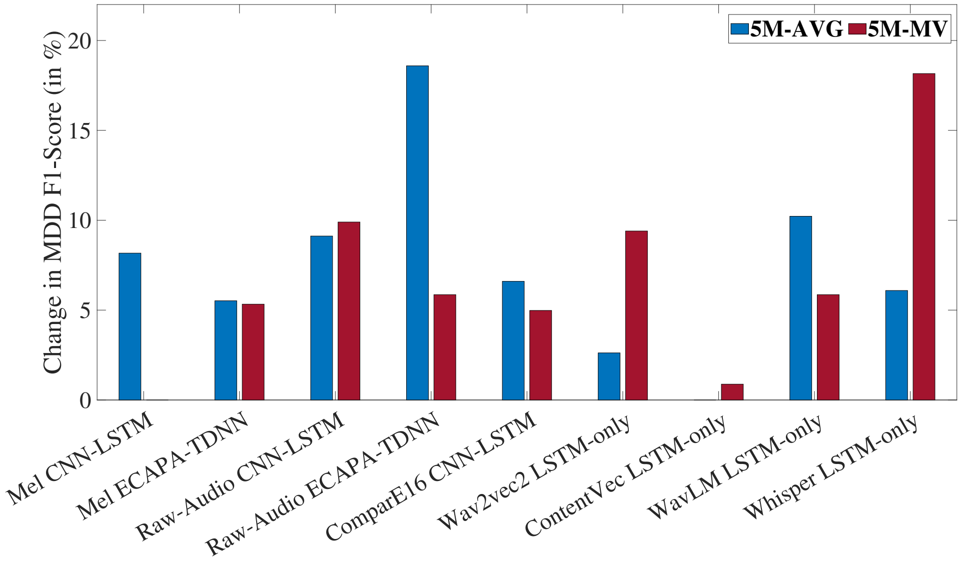
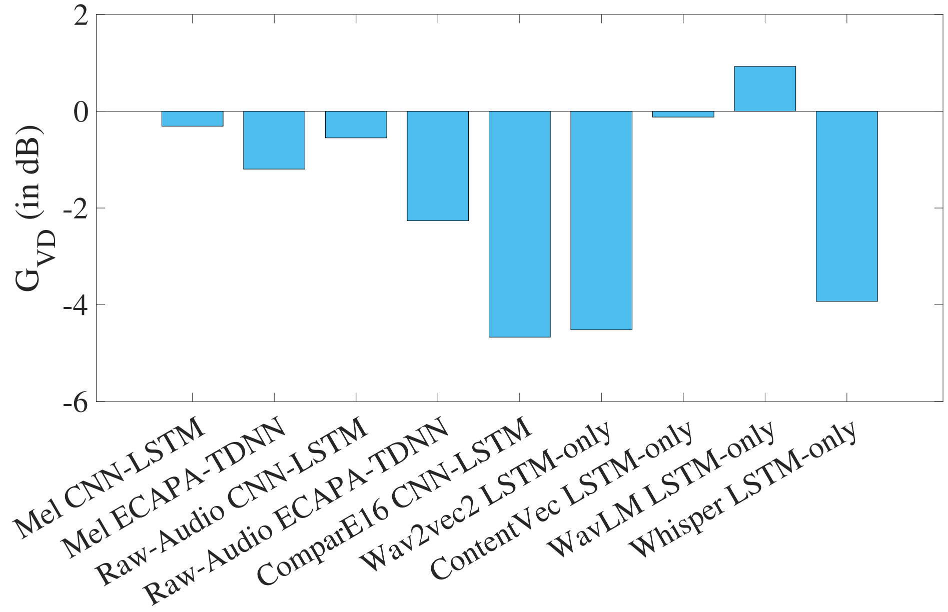
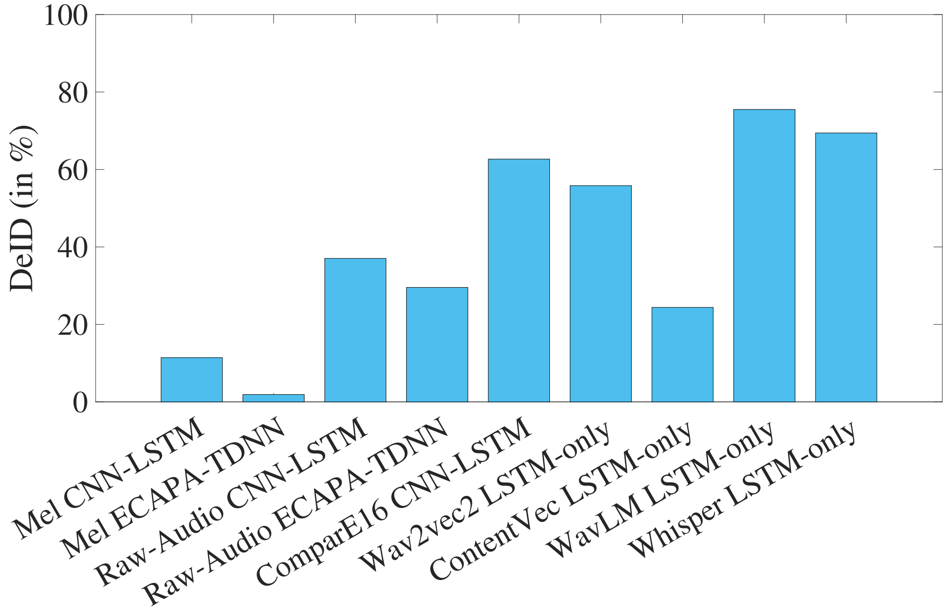


Figure 6.5: Relative improvements, in percentage, in MDD classification F1-Score when speaker disentanglement is applied in the form of LEKLD. The X-axis of each plot represents the 9 different feature-model combinations. 5M-AVG and 5M-MV refer to the averaging and majority voting aggregation of the 5 models, respectively. A higher value indicates a greater improvement in depression detection after speaker disentanglement is applied.



(a)



(b)

Figure 6.6: (a) G_{VD} in dB and (b) $De - ID$ in %, respectively, for each experiment when speaker disentanglement is applied in the form of LEKLD. The X-axis of each plot represents the 9 different feature-model combinations.

Further, the ECAPA-TDNN model trained with Raw-Audio signals achieves the highest improvement in MDD detection with an improvement of 18.59% in F1-AVG (5M-AVG) from 0.6196 for the baseline to 0.7348 for the proposed method ($\lambda = 5e - 3$). The G_{VD} for this model is -2.26 dB and the DeID is 29.56%.

Similar to ADV and LEV, the G_{VD} for WavLM was positive (0.9268 dB). However, the DeID for the same feature was 75%. Again, this shows that LEKLD in this scenario can successfully mask speaker-identity but the disentangled speaker-representations are more separable than the baseline speaker-representations. In contrast, ComparE16 features with the CNN-LSTM model achieved the lowest G_{VD} of -4.66 dB with a DeID of 62.68%.

6.4 Comparison with ADV

After comparing the results of loss equalization-based and adversarial speaker disentanglement methods, we observed that these methods consistently improved depression detection performance while simultaneously degrading speaker identification and separability. Among the proposed methods, the combination of ComparE16 features with a CNN-LSTM model achieved the highest F1-AVG score of 80% for Major Depressive Disorder (MDD) detection when Loss Equalization with Cross-Entropy (LECE) was applied. The second-best F1-AVG score of 79% was achieved by both the Whisper/LSTM-only model with Loss Equalization with KL Divergence (LEKLD) and the Raw-Audio/ECAPA-TDNN model with adversarial training (ADV). The third-best performance, with an F1-AVG score of 0.7756, was obtained using WavLM features with an LSTM-only model and LECE. Among the nine model-feature combinations tested, loss-equalization methods (LEV, LECE, and LEKLD) outperformed the adversarial method (ADV) in seven scenarios, achieved equal performance in one, and showed inferior performance in only one case. These results demonstrate the superiority of loss-equalization methods over adversarial methods for speaker disentanglement in the

Table 6.7: Results, in terms of F1-Score (5 model logit average - 5M-AVG), for speaker disentanglement through LEKLD using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	5-Models Logit Average						
				F1-Score			Confusion Matrix			
				F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6081	0.6977	0.5185	15	8	5	7
		Yes (α = 5e-5)	293k	0.6578	0.7556	0.5600	17	6	5	7
	Δ (in %)	-	-	8.17	8.30	8.00	-	-	-	-
	ECAPA-TDNN	No	515k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 1e-1)	529k	0.6941	0.7727	0.6154	17	6	4	8
	Δ (in %)	-	-	5.52	2.26	9.89	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6259	0.7755	0.4762	19	4	7	5
		Yes (α = 2e-3)	459k	0.6830	0.7826	0.5833	18	5	5	7
	Δ (in %)	-	-	9.12	0.92	22.49	-	-	-	-
	ECAPA-TDNN	No	595k	0.6196	0.7391	0.5000	17	6	6	6
		Yes (α=5e-3)	609k	0.7348	0.8333	0.6364	20	3	5	7
	Δ (in %)	-	-	18.59	12.75	27.28	-	-	-	-
CompareE16 (384x130)	CNN-LSTM	No	1.15M	0.5791	0.7234	0.4348	17	6	7	5
		Yes (α = 1e-2)	1.18M	0.6173	0.6829	0.5517	14	9	4	8
	Δ (in %)	-	-	6.60	-5.60	26.89	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=5e-4)	3.7M	0.7009	0.8462	0.5556	22	1	7	5
	Δ (in %)	-	-	2.62	8.13	-4.75	-	-	-	-
Contentvec-100 (193x768)	LSTM-only	No	3.6M	0.7287	0.7907	0.6667	17	6	3	9
		Yes (α=5e-4)	3.7M	0.7287	0.7907	0.6667	17	6	3	9
	Δ (in %)	-	-	0.00	0.00	0.00	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6429	0.7143	0.5714	15	8	4	8
		Yes (α=5e-1)	3.7M	0.7086	0.8085	0.6087	19	4	5	7
	Δ (in %)	-	-	10.22	13.19	6.53	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6438	0.7660	0.5217	18	5	6	6
		Yes (α = 1e-5)	3.4M	0.6830	0.7826	0.5833	18	5	5	7
	Δ (in %)	-	-	6.09	2.17	11.81	-	-	-	-

Table 6.8: Results, in terms of F1-Score (5 model majority voting - 5M-MV), for speaker disentanglement through LEKLD using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	5-Models Majority Voting						
				F1-Score			Confusion Matrix			
				F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 5e-5)	293k	0.6578	0.7556	0.5600	17	6	5	7
	Δ (in %)	-	-	0.00	0.00	0.00	-	-	-	-
	ECAPA-TDNN	No	515k	0.7086	0.8085	0.6087	19	4	5	7
		Yes (α = 1e-1)	529k	0.7464	0.8261	0.6667	19	4	4	8
	Δ (in %)	-	-	5.33	2.18	9.53	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6686	0.7917	0.5455	19	4	6	6
		Yes (α = 2e-3)	459k	0.7348	0.8333	0.6364	20	3	5	7
	Δ (in %)	-	-	9.90	5.25	16.66	-	-	-	-
	ECAPA-TDNN	No	595k	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=5e-3)	609k	0.7348	0.8333	0.6364	20	3	5	7
	Δ (in %)	-	-	5.86	7.84	3.41	-	-	-	-
CompareE16 (384x130)	CNN-LSTM	No	1.15M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α = 1e-2)	1.18M	0.7287	0.7907	0.6667	17	6	3	9
	Δ (in %)	-	-	4.98	2.33	8.34	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=5e-4)	3.7M	0.7472	0.8627	0.6316	22	1	6	6
	Δ (in %)	-	-	9.40	10.24	8.28	-	-	-	-
Contentvec-100 (193x768)	LSTM-only	No	3.6M	0.7287	0.7907	0.6667	17	6	3	9
		Yes (α=5e-4)	3.7M	0.7351	0.7805	0.6897	16	7	2	10
	Δ (in %)	-	-	0.88	-1.29	3.45	-	-	-	-
WavLM-base (193x768)	LSTM-only	No	3.6M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=5e-1)	3.7M	0.7348	0.8333	0.6364	20	3	5	7
	Δ (in %)	-	-	5.86	7.84	3.41	-	-	-	-
Whisper-base (193x512)	LSTM-only	No	3.4M	0.6686	0.7917	0.5455	19	4	6	6
		Yes (α = 1e-5)	3.4M	**0.7900**	0.8800	0.7000	22	1	5	7
	Δ (in %)	-	-	18.16	11.15	28.32	-	-	-	-

Table 6.9: Speaker separability results, in terms of G_{VD} (in dB) and $DeID$ (in%), for speaker disentanglement through LEKLD using the DAIC-WOZ dataset. The best G_{VD} and $DeID$ are bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	G_{VD} (in dB)	DeID in (%)
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	-	-
		Yes (α = 5e-5)	293k	-0.3079	11.42
	ECAPA-TDNN	No	515k	-	-
		Yes (α = 1e-1)	529k	-1.1953	1.90
Raw-Audio (61440x1)	CNN-LSTM	No	445k	-	-
		Yes (α = 2e-3)	459k	-0.5503	37.05
	ECAPA-TDNN	No	595k	-	-
		Yes (α=5e-3)	609k	-2.2619	29.56
Compar E16 (384x130)	CNN-LSTM	No	1.15M	-	-
		Yes (α = 1e-2)	1.18M	**-4.6687**	62.68
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	-	-
		Yes (α=5e-4)	3.7M	-4.5179	55.83
Contentvec-100 (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=5e-4)	3.7M	-0.1199	24.40
WavLM-base (193x768)	LSTM-only	No	3.6M	-	-
		Yes (α=5e-1)	3.7M	0.9268	**75.49**
Whisper-base (193x512)	LSTM-only	No	3.4M	-	-
		Yes (α = 1e-5)	3.4M	-3.9297	69.42

context of depression detection.

6.5 Experimental Results for the EATD dataset

To evaluate the generalizability of the proposed loss equalization speaker-disentanglement methods to a different language, we applied them to the EATD dataset. Depression detection results for the three loss-equalization disentanglement methods, in terms of F1-Score and Confusion Matrix, are summarized in Tables 6.10 and 6.12, when using Mel-spectrogram and Raw-audio signals as input features, respectively. The corresponding speaker separability results are in Tables 6.11 and 6.13. Similar to the DAIC-WoZ dataset, depression detection performance improves for all 6 experiments with an average improvement of 11.77% for Mel-spectrograms and 10.33% for Raw-audio signals. Improvements in MDD detection were statistically significant in 5 out of the 6 experiments (relative change obtained when Raw-Audio was used as input feature for LEKLD was not statistically significant).

When Mel-spectrograms are used as inputs, LECE achieves the best performance in terms of depression detection with an F1-AVG of 0.5988 which represents an improvement of 15.91% when compared to the baseline setup without disentanglement. In terms of speaker-separability, LECE achieved the best results with a G_{VD} of -2.86 dB and a $DeID$ of 78.81%.

Similar improvements were observed when Raw-audio signals were used as input features. The overall best performance for the EATD dataset was achieved when LECE was applied. In this case, the F1-AVG was 0.7368, outperforming the baseline by 14.57%. The best G_{VD} and $DeID$ were also obtained using LECE: -3.0941 dB and 84.73%, respectively.

Similar to the results observed on the DAIC-WoZ dataset, the performance of loss-equalization methods on this dataset surpassed that of adversarial methods. The combination of Raw-audio signals with a CNN-LSTM model and Loss Equalization with Cluster Em-

Table 6.10: Results, in terms of F1-AVG and Confusion-Matrix, for all loss-equalization based speaker disentanglement methods using the development set of EATD dataset and Mel-Spectrogram as input features. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN, and TP are True Negative, False Positive, False Negative, and True Positive, respectively. The best F1-Score is bold-faced.

Feature-Model	Speaker Disentanglement	Model Parameters	F1-Score			Confusion Matrix			
			F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram CNN-LSTM	No	415k	0.5166	0.8593	0.1739	58	10	9	2
	LEV (α=1e-6)	430k	0.5756	0.8550	0.2963	56	12	7	4
	Δ (in %)	-	11.42	-0.5	70.39	-	-	-	-
	LECE (α=5e-6)	430k	**0.5988**	0.8527	0.3448	55	13	6	5
	Δ (in %)	-	15.91	-0.76	98.27	-	-	-	-
	LEKLD (α=4e-4)	430k	0.5578	0.8657	0.2500	58	10	8	3
	Δ (in %)	-	7.98	0.75	43.76	-	-	-	-

Table 6.11: Results, G_{VD} and DeID, for all speaker disentanglement methods using the development set of EATD dataset and Mel-Spectrogram as input features. The best G_{VD} and $DeID$ are bold-faced.

Feature-Model	Speaker Disentanglement	G_{VD} (in dB)	$DeID$ (in %)
Mel-Spectrogram CNN-LSTM	LEV	-2.3572	58.47
	LECE	**-2.8661**	**78.81**
	LEKLD	-2.0146	64.29

Table 6.12: Results, in terms of F1-AVG and Confusion-Matrix, for all loss-equalization based speaker disentanglement methods using the development set of EATD dataset and Raw-audio signals as input features. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN, and TP are True Negative, False Positive, False Negative, and True Positive, respectively. The best F1-Score is bold-faced.

Feature-Model	Speaker Disentanglement	Model Parameters	F1-Score			Confusion Matrix			
			F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Raw-audio CNN-LSTM	No	445k	0.6431	0.9051	0.3810	62	6	7	4
	LEV (α=2e-3)	456k	0.7052	0.9104	0.5000	61	7	5	6
	Δ (in %)	-	9.66	0.59	31.23	-	-	-	-
	LECE (α=4e-3)	456k	**0.7368**	0.8070	0.6667	63	5	4	7
	Δ (in %)	-	14.57	-10.84	75.12	-	-	-	-
	LEKLD (α=1e-6)	456k	0.6865	0.9444	0.4286	68	0	8	3
	Δ (in %)	-	6.75	4.34	12.49	-	-	-	-

Table 6.13: Results, G_{VD} and DeID, for all loss-equalization based speaker disentanglement methods using the development set of EATD dataset and Raw-audio signals as input features. The best G_{VD} and $DeID$ are bold-faced.

Feature-Model	Speaker Disentanglement	G_{VD} (in dB)	$DeID$ (in %)
Raw-audio CNN-LSTM	LEV	-1.2572	63.44
	LECE	**-3.0941**	**84.73**
	LEKLD	-2.5668	69.22

beddings (LECE) achieved the best performance, with an F1-AVG score of 0.7368. This was followed by the Raw-audio/CNN-LSTM model with adversarial training (ADV) with an F1-AVG of 0.720. For both Mel-Spectrogram and Raw-Audio signal representations, loss-equalization methods consistently outperformed the adversarial method, demonstrating their superiority in speaker disentanglement for depression detection tasks across different datasets and input features.

6.6 Chapter Summary

This chapter introduced loss equalization methods as an alternative to adversarial speaker disentanglement, aiming to overcome its limitations. The proposed methods were evaluated on two datasets in different languages: the English DAIC-WoZ dataset and the EATD dataset. On the DAIC-WoZ dataset, the loss equalization via cross-entropy (LECE) method, combined with ComparE16 features and a CNN-LSTM model, achieved the best depression detection performance with an F1-AVG score of 80%, surpassing the best adversarial F1-AVG score of 79%. Among the nine feature-model combinations tested, loss equalization methods outperformed adversarial speaker disentanglement in seven cases. On the EATD dataset, the combination of Raw-Audio signals, a CNN-LSTM model, and LECE achieved the best performance with an F1-AVG score of 73.68%, outperforming the adversarial training method (ADV), which achieved an F1-AVG score of 72%. These results demonstrate the effectiveness of loss equalization methods in improving depression detection performance across different datasets and languages.

Chapter 7

Unsupervised Speaker Disentanglement

7.1 Introduction

Recent studies on adversarial techniques and loss equalization have made notable progress in improving depression detection performance while reducing reliance on features related to patient identity. However, these approaches still face significant challenges. Firstly, they require speaker labels from patient datasets, compromising the privacy-preserving goals of depression detection systems. Secondly, many methods employ adversarial loss maximization for speaker disentanglement, which, despite its effectiveness, is inherently unstable due to the lack of upper bounds in the adversarial domain objective function [112]. Thirdly, these methods introduce additional parameters to the model training framework, such as adversarial domain prediction layers or reconstruction decoders, which are not essential for the primary task of depression detection. These extra components add complexity without directly contributing to the main objective of depression detection.

This chapter presents a novel speaker disentanglement method addressing the challenges mentioned earlier, drawing inspiration from the growing adoption of unsupervised learning approaches [115].

7.2 Unsupervised Speaker Disentanglement

Recall from Section 5.3, the total loss for speaker disentanglement via adversarial learning can be written as -

$$L_{total-ADV} = L_{MDD} - \alpha \cdot L_{SPK-ADV}, \tag{7.1}$$

Where L_{MDD} is the depression prediction loss and $L_{SPK-ADV}$ is the speaker prediction loss for the adversarial method. α is a hyperparameter controlling the contribution of the adversarial loss to the main loss function where the negative sign indicates that the speaker prediction loss is maximized thereby forcing the model to learn more depression discriminatory features and less speaker discriminatory features. The speaker prediction loss $L_{SPK-ADV}$ is usually the Cross-Entropy loss defined as -

$$L_{SPK-ADV}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \cdot \log(\hat{y}_{ij}), \tag{7.2}$$

where y is the ground-truth speaker label and $\hat{y}$ is the predicted speaker probabilities for N samples and C speakers.

Similarly, the total loss for speaker disentanglement via loss equalization (refer to Section 6.1.1) can be written as:

$$L_{total-LE} = L_{MDD} + \alpha \cdot L_{SPK-LE}, \tag{7.3}$$

where L_{SPK-LE} is the speaker prediction loss equalization loss. Although loss equalization is non-adversarial, it still has a speaker prediction branch to compute the speaker prediction loss.

The aforementioned methods face several major challenges. They require ground-truth speaker labels (y) to achieve disentanglement. The speaker identity disentanglement relies on loss maximization ($-\alpha \cdot L_{SPK-ADV}$), which lacks an upper bound, leading to reduced training stability. Moreover, the speaker prediction branch, used to obtain $\hat{y}$, adds extra model parameters that are not essential for depression detection, making the approach inefficient. In contrast, our proposed approach offers an unsupervised method of speaker disentanglement that addresses these issues. It eliminates the need for speaker labels from patient datasets, avoids loss maximization, and does not introduce additional model parameters. Figure 7.1 illustrates our proposed method.

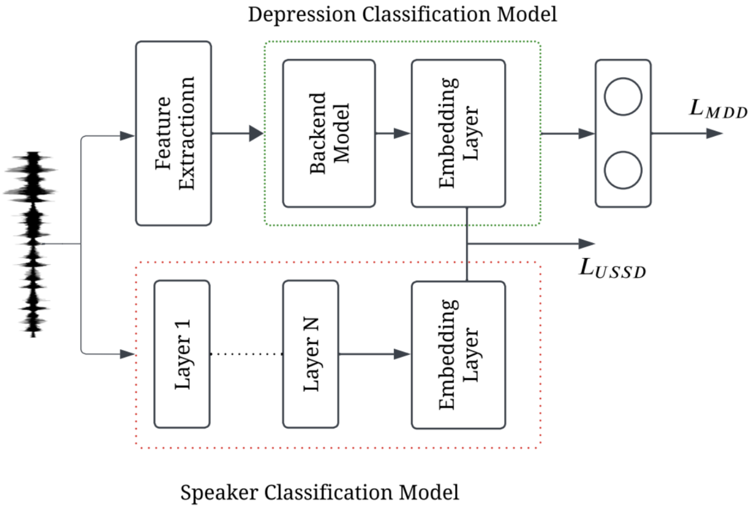


Figure 7.1: The unsupervised speaker disentanglement method (USSD) aims to minimize cosine similarity between latent spaces of depression classification and speaker classification models.

Let's consider two models: a depression classification model (θ_{MDD}) and a speaker classification model (θ_{SPK}). Given a speech input $X \in \mathbb{R}^{N \times F}$, where N represents the batch size and F denotes the number of features, the latent embeddings of these models can be represented as:

$$H_{MDD_X} = \theta_{MDD}(X) \tag{7.4}$$

$$H_{SPK_X} = \theta_{SPK}(X) \tag{7.5}$$

H_{MDD_X} and $H_{SPK_X} \in \mathbb{R}^{N \times D}$, where D represents the embedding size. We then compute the predicted cosine similarity matrix between the two latent space embeddings. This is done by calculating the cosine similarity between every pair of embeddings as follows:

$$Y_{pred_{(i,j)}} = \frac{H_{MDD_{X_i}} \cdot H_{SPK_{X_j}}}{||H_{MDD_{X_i}}|| \cdot ||H_{SPK_{X_j}}||} \tag{7.6}$$

where $1 \leq i, j \leq N$ and $Y_{pred} \in \mathbb{R}^{N \times N}$. The objective of the disentanglement process is to minimize the cosine similarity between the two embedding spaces by enforcing orthogonality between the depression and speaker latent spaces. To achieve this, we specifically set Y_{target} to 0, rather than -1. To enhance convergence during implementation, we incorporate a small noise value, denoted as ϵ [54].

$$Y_{target_{(i,j)}} = 0 + \epsilon, \tag{7.7}$$

$$\epsilon \in U(0, 1e - 8) \tag{7.8}$$

We define the proposed speaker disentanglement loss function L_{USSD} as follows -

$$L_{USSD} = MSE(Y_{pred}, Y_{target}) \tag{7.9}$$

and the total loss as:

$$L_{total-USSD} = L_{MDD} + \alpha \cdot L_{USSD}, \tag{7.10}$$

Minimizing the loss function described in Eq.7.10 drives the model to emphasize learning more discriminatory information related to depression while reducing its focus on speaker-related distinctions. Our proposed method achieves speaker disentanglement through loss minimization, in contrast to ADV (Eq. 7.1). Furthermore, unlike LE (Eq. 7.3), our approach does not require additional parameters for a speaker prediction branch.

A key advantage of our approach is that embeddings from θ_{SPK} can be extracted without the need for speaker labels, making the proposed speaker disentanglement method unsupervised. Furthermore, only the parameters of θ_{MDD} require updating during training. The θ_{SPK} model can remain a pre-trained model with frozen weights, eliminating the need for fine-tuning.

7.3 Experimental Details

In this chapter, we conducted experiments using two datasets: DAIC-WOZ (English) [104], and the EATD [96]. Similar to experiments in Section 6.2, the CONVERGE dataset was not included in the evaluations for this chapter.

In terms of the backend model, we employed three different backend models: 1) a modified version of DepAudioNet [61], 2) an ECAPA-TDNN model, and 3) an LSTM-only model. For the English dataset, we evaluated the performance using Mel-Spectrogram, Raw-Audio, CompарE16 and Wav2vec2 input features, while for the Mandarin dataset, we focused on Mel-Spectrogram and Raw-Audio signals only. All model parameters such as number of layers, number of channels of convolutional filters, stride, kernel size, etc are the same as those described in Section 5.4

7.4 Results and Discussion

7.4.1 USSD versus Baseline for DAIC-WOZ

Figures 7.2 and 7.3 show the relative improvement in MDD classification F1-AVG, and the speaker separability metrics G_{VD} (in dB) and $De - ID$ (in %) for each model-feature combination when USSD is applied, respectively. Detailed results, in terms of F1-Score for 5M-MV are presented in Table 7.1. Speaker-separability results are presented in Table 7.2. From Figure 7.2, it can be observed that every experiment leads to an improvement in MDD detection performance with an average improvement in MDD F1-AVG by 8.2%. The highest improvement was 11.81%, achieved with ComparE16 features and the LSTM-only model. The smallest improvement was 3.8%, observed with Mel-Spectrogram features and the CNN-LSTM model. Improvements in MDD detection were statistically significant [64] in 4 out of the 6 experiments (relative change obtained with Raw-Audio signals were not statistically significant). The G_{VD} is negative for all six experiments. In terms of $DeID$, ComparE16 input features with CNN-LSTM model resulted in the highest DeID of 92.87% and Mel-Spectrograms with ECAPA-TDNN models resulted in the lowest DeID of 5.97%.

7.4.2 USSD versus Baseline for EATD

The generalizability of the proposed approach is evaluated by applying the method on the EATD dataset. The results, summarized in Table 7.3 demonstrate that applying ADV to the CNN-LSTM model trained on Mel-Spectrograms leads to a 5.2% increase in F1-AVG, from 0.5166 to 0.5660, with a G_{VD} of -2.756dB and a DeID of 84.5%. When Raw-audio is used as the input feature, similar improvements in depression detection are observed. The F1-AVG for MDD prediction increases by 12.4%, from 0.6430 for the baseline model to 0.7238 for the proposed method ($\lambda = 1e - 6$), with a G_{VD} of -2.457dB and a DeID of 85.9%.

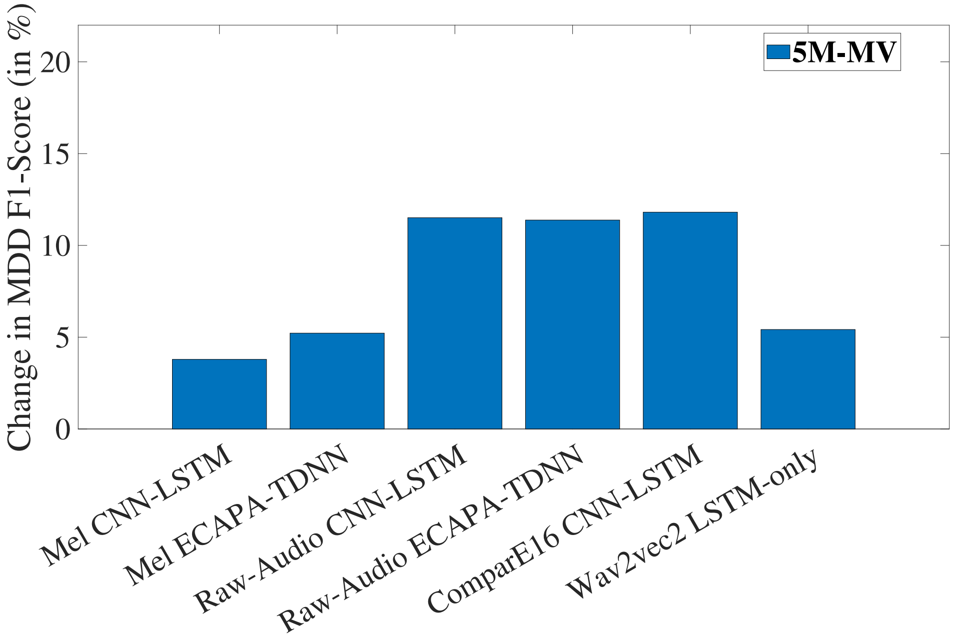
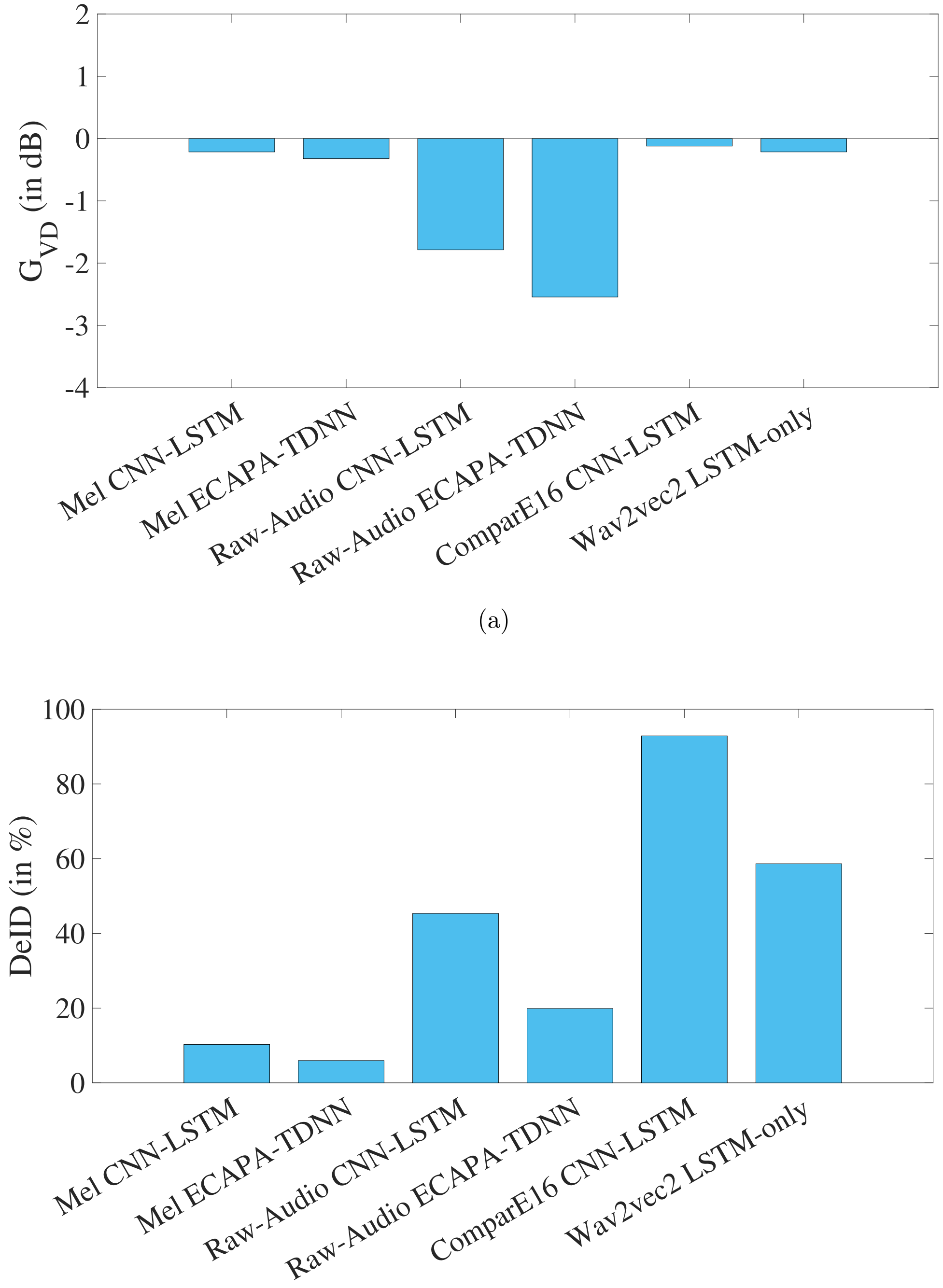


Figure 7.2: Relative improvements, in percentage, in MDD classification F1-Score when speaker disentanglement is applied in the form of USSD. The X-axis of each plot represents the 6 different feature-model combinations. 5M-MV refer to the averaging and majority voting aggregation of the 5 models, respectively. A higher value indicates a greater improvement in depression detection after speaker disentanglement is applied.



(a)

(b)

Figure 7.3: (a) G_{VD} in dB and (b) $De - ID$ in %, respectively, for each experiment when speaker disentanglement is applied in the form of USSD. The X-axis of each plot represents the 9 different feature-model combinations.

Table 7.1: Results, in terms of F1-Score (5 model majority voting - 5M-MV), for speaker disentanglement through USSD using the development set of the DAIC-WOZ dataset. The highlighted row (Δ) for each feature-model configuration indicates the relative change in performance of that model without disentanglement versus our proposed method. TN, FP, FN and TP stands for True Negative, False Positive, False Negative and True Positive, respectively. The best F1-Score is bold-faced.

Input				5-Models Majority Voting						
Feature	Model	Speaker	Model	F1-Score			Confusion Matrix			
(Seq.len x Num. of Features)	Architecture	Disentanglement	Parameters	F1(Avg)	F1(ND)	F1(D)	TN	FP	FN	TP
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	0.6578	0.7556	0.5600	17	6	5	7
		Yes (α = 1e-4)	280k	0.683	0.783	0.583	18	5	5	7
	Δ (in %)	-	-	3.79	3.63	4.11	-	-	-	-
	ECAPA-TDNN	No	515k	0.7086	0.8085	0.6087	19	4	5	7
		Yes (α = 5e-3)	515k	0.746	0.826	0.667	19	4	4	8
	Δ (in %)	-	-	5.22	2.16	9.57	-	-	-	-
Raw-Audio (61440x1)	CNN-LSTM	No	445k	0.6686	0.7917	0.5455	19	4	6	6
		Yes (α = 3e-4)	445k	0.746	0.826	0.667	20	3	5	7
	Δ (in %)	-	-	11.51	4.33	22.27	-	-	-	-
	ECAPA-TDNN	No	595k	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α=5e-5)	595k	0.773	0.851	0.696	20	3	5	7
	Δ (in %)	-	-	11.38	10.13	13.09	-	-	-	-
CompareE16 (384x130)	CNN-LSTM	No	1.15M	0.6941	0.7727	0.6154	17	6	4	8
		Yes (α = 2e-5)	1.15M	**0.776**	0.885	0.667	17	6	3	9
	Δ (in %)	-	-	11.82	14.53	8.38	-	-	-	-
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	0.6830	0.7826	0.5833	18	5	5	7
		Yes (α=4e-5)	3.6M	0.720	0.840	0.600	22	1	6	6
	Δ (in %)	-	-	5.42	7.33	2.86	-	-	-	-

Table 7.2: Speaker separability results, in terms of G_{VD} (in dB) and $DeID$ (in%), for speaker disentanglement through USSD using the DAIC-WOZ dataset. The best G_{VD} and $DeID$ are bold-faced.

Input Feature (Seq.len x Num. of Features)	Model Architecture	Speaker Disentanglement	Model Parameters	G_{VD} (in dB)	DeID in (%)
Mel-Spectrogram (120x40), (120x80)	CNN-LSTM	No	280k	-	-
		Yes (α = 1e-4)	280k	-0.2147	10.29
	ECAPA-TDNN	No	515k	-	-
		Yes (α = 5e-3)	515k	-0.3228	5.97
Raw-Audio (61440x1)	CNN-LSTM	No	445k	-	-
		Yes (α = 3e-4)	445k	-1.787	45.35
	ECAPA-TDNN	No	595k	-	-
		Yes (α=5e-5)	595k	**-2.5441**	19.90
CompareE16 (384x130)	CNN-LSTM	No	1.15M	-	-
		Yes (α = 2e-5)	1.15M	-0.1211	**92.87**
Wav2Vec2.0-base (200x768)	LSTM-only	No	3.6M	-	-
		Yes (α=4e-5)	3.6M	-0.2143	58.65

Table 7.3: Results, in terms of F1-AVG, Confusion-Matrix, G_{VD} and DeID, for speaker disentanglement through USSD using the development set of EATD dataset. TN, FP, FN, and TP are True Negative, False Positive, False Negative, and True Positive, respectively. The best F1-Score is bold-faced.

				Confusion Matrix					
Feature-Model	Speaker Disentanglement	# Params	F1-AVG	TN	FP	FN	TP	G_{VD} (in dB)	DeID (in %)
Mel-Spectrogram	No	415k	0.5166	58	10	9	2	-	-
CNN-LSTM	USSD ($\alpha = 2e-4$)	415k	0.5660	56	12	7	4	-2.756	84.5
Raw-Audio	No	445k	0.643	62	6	7	4	-	-
CNN-LSTM	USSD ($\alpha = 4e-5$)	445k	**0.7238**	62	6	5	6	-2.457	85.9

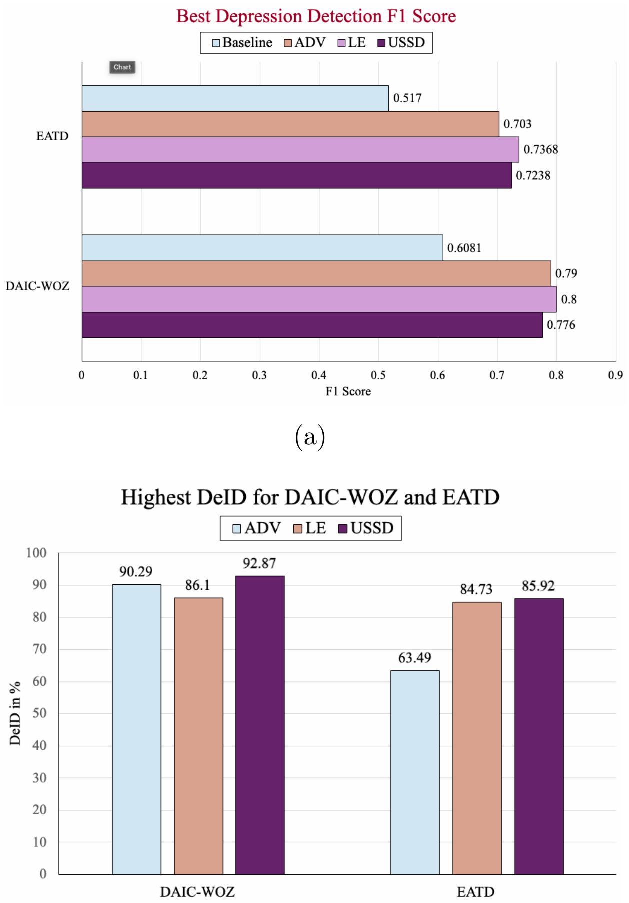
7.4.3 USSD versus Other Speaker Disentanglement Methods

The USSD method demonstrates competitive performance when compared to established techniques like Adversarial (ADV) and Loss Equalization (LE) approaches. The best F1 scores achieved by USSD (0.776) are comparable to those of ADV (0.79) and LE (0.8), indicating that our method maintains high accuracy in depression detection. Notably, USSD outperforms these methods in terms of speaker de-identification (DeID), showing superior capability in reducing speaker-specific information. This is particularly advantageous for preserving patient privacy in clinical applications. A key strength of USSD lies in its applicability to scenarios where speaker labels are unavailable in the training data. This feature makes USSD especially valuable in real-world settings where obtaining speaker labels for the training data may be challenging or impractical.

7.4.4 Chapter Summary

The proposed method introduced in this chapter aims to reduce the cosine similarity between the latent spaces of two models: one for depression detection and another for speaker classification. By operating at the embedding level, we eliminate the need for speaker labels in patient datasets, enhancing privacy protection. We reformulate the training process into a loss minimization framework, overcoming the unboundedness issues associated with adversarial methods. Our approach achieves efficiency by utilizing speaker classification models solely as embedding extractors, without retraining or fine-tuning. Additionally, this strategy avoids the need for domain prediction or reconstruction, resulting in a more streamlined model with fewer parameters compared to previous approaches.

To validate the efficacy of our proposed method, we conducted comprehensive experiments. The results demonstrate its superiority over baseline models lacking speaker disentanglement in depression detection tasks. Moreover, our approach achieves performance comparable to



(a)

(b)

Figure 7.4: Comparison of USSD and other speaker disentanglement methods in terms of(a) best depression detection F1 and (b) privacy attribute DeID score for DAIC-WOZ and EATD datasets.

adversarial and loss equalization methods. We evaluated the framework across multiple input features and backend models, establishing its generalizability to diverse architectures.

This approach yields superior performance compared to baseline models without disentanglement. Notably, using ComparE16 features with a CNN-LSTM model and the English dataset, we achieved an F1-Score of 0.776, outperforming the baseline by 11.7%. Using the Mandarin EATD dataset and Raw-Audio signals, a 12.4% improvement was observed. When compared to other speaker disentanglement methods, our Unsupervised Speaker Space Disentanglement (USSD) achieves comparable performance in depression detection while demonstrating improved results in speaker de-identification (DeID) - for example, using the DIAC-WOZ dataset USSD achieves a DeID of 92% compared to ADV's 90.29% . These findings highlight the effectiveness of our method in balancing accurate depression detection with enhanced privacy protection in scenarios where speaker labels are unavailable for the training data.

Chapter 8

Summary and Future Work

In recent years, the field of detecting depression through speech analysis has experienced substantial growth. This innovative approach leverages advancements in machine learning and signal processing to identify subtle vocal cues that may indicate depressive symptoms. As a non-invasive and potentially cost-effective method, speech-based depression detection holds promise for early diagnosis and monitoring of mental health conditions. However, significant challenges remain.

Data scarcity presents one such challenge in the field of speech-based depression detection, hampering the development and validation of robust detection models. Obtaining large-scale, high-quality datasets of speech samples from individuals with depression is inherently difficult due to several factors. Ethical considerations necessitate careful protocols to ensure participant well-being and informed consent, adding complexity to data collection efforts. Accurate labeling of speech samples requires professional clinical diagnoses, which are both time-consuming and costly to obtain on a large scale. Privacy concerns surrounding mental health data often make potential participants hesitant to contribute, further limiting data availability.

Another critical issue that has emerged recently affecting speech-based depression detection is privacy concerns surrounding speaker identity as the collection and analysis of voice data raise questions about data security, consent, and the potential for misuse of sensitive personal information. Voice data inherently contains unique bio-metric information that can serve as an acoustic fingerprint, potentially allowing for the identification of individuals even from short audio samples. This raises significant privacy risks, as voice recordings collected for depression analysis could inadvertently compromise a person's anonymity. The distinctiveness of voice characteristics means that even if traditional identifiers are removed, re-identification remains a possibility through voice matching techniques. This risk is particularly concerning in the context of mental health data, where confidentiality is paramount.

In this dissertation, to address the data-scarcity problem, a novel data-augmentation method called frame rate-based data augmentation (FrAUG) is proposed and presented in Chapter 4. Prior to this chapter, the database and features used are presented in Chapter 2 and the models and evaluation metrics are described in Chapter 3.

FrAUG effectively tackles the issue of limited training data by generating new samples through varying frame-width and frame-shift parameters during feature extraction. This method provides models with different time-frequency resolutions without modifying vocal tract or voice source related parameters and hence preserves acoustic information that may be important for MDD modeling purposes.

To address privacy concerns arising from the use of speaker identity information, the dissertation investigated several speaker disentanglement techniques. Five distinct methods were proposed and compared: adversarial SID-loss maximization (ADV) in Chapter 5, SID-loss equalization with variance (LEV), with cross-entropy (LECE),and with KL divergence (LEKLD) in Chapter 6 and Unsupervised Speaker Disentanglement via cosine similarity minimization(USSD) in Chapter 7.

8.1 Discussion

The research presented in this dissertation represents a significant step forward in the development of accurate and privacy-preserving speech-based depression detection systems. By addressing the dual challenges of data scarcity and privacy concerns, this work paves the way for more reliable and ethically sound deployment of these systems in clinical settings. The FrAUG technique offers a powerful solution to the perennial problem of limited training data in mental health applications. By generating diverse and informative training samples without altering critical depression-related information, FrAUG enables the development of more robust and generalizable models. This advancement is particularly crucial in the context of mental health, where data collection can be challenging and time-consuming.

The various speaker disentanglement methods proposed in this dissertation, particularly the USSD approach, represent a significant leap forward in balancing the need for accurate depression detection with the imperative of protecting patient privacy. These methods consistently demonstrated improvements in depression detection performance while simultaneously enhancing privacy attributes, as evidenced by improved speaker de-identification (DeID)scores. This dual achievement is particularly noteworthy, as it addresses one of the primary concerns hindering the widespread adoption of speech-based mental health screening tools.

The consistent improvements observed across different datasets, languages, and model architectures underscore the robustness and versatility of the proposed approaches. This generalizability is crucial for the practical implementation of these techniques in diverse clinical settings and populations. Moreover, the success of these methods in both English and Mandarin contexts suggests their potential applicability to a wide range of linguistic and cultural backgrounds.

The research presented in this dissertation opens up new avenues for exploring the intricate relationship between speech characteristics and mental health states. The ability to disentangle speaker-specific information from depression-related features not only enhances privacy but also provides a clearer window into the acoustic markers of depression. This improved understanding could lead to more targeted and personalized interventions in the future.

In conclusion, this dissertation makes substantial contributions to the field of speech-based depression detection, addressing critical challenges of data scarcity and privacy-preservation and paving the way for more widespread and responsible use of these technologies in clinical practice. The proposed methods offer a promising foundation for future research and development in this important area of mental health diagnostics.

8.2 Future Work

While this dissertation has made significant strides in advancing speech-based depression detection, several promising directions for future research remain. These directions aim to further enhance the accuracy, robustness, and clinical utility of the proposed methods, as well as explore their potential applications in broader mental health contexts.

One important area for future investigation is the integration of speech-based features with other modalities, such as text, facial expressions, or physiological signals. Preliminary work in this direction shows promising results of score-level fusion between speaker-disentangled audio models and Word2vec-based text models [86], suggesting that there is significant potential in multi-modal approaches. Future work should explore more sophisticated ways to combine text and speech modalities, potentially developing joint embedding spaces that capture complementary information from both sources while maintaining privacy. Additionally, investigating the use of advanced natural language processing techniques, such as transformers, in conjunction with the proposed speaker disentanglement methods could lead to even more accurate and robust depression detection systems.

Expanding the cross-cultural validation of the proposed methods is also an important consideration. While this dissertation has shown promising results in both English and Mandarin contexts, further evaluation across a wider range of languages and cultural backgrounds would ensure the generalizability and effectiveness of these approaches across diverse populations. Additionally, exploring the adaptability of the proposed methods to other mental health conditions is another important area of research. Investigating whether similar approaches can be effective in detecting conditions such as anxiety disorders or bipolar disorder could significantly expand the impact of this work in the broader field of mental health diagnostics.

Enhancing the interpretability of the disentangled models represents another valuable direction for future work. Developing techniques to provide clinicians with more insights into the specific aspects of speech that contribute to depression detection could greatly enhance the clinical utility of these systems. This could involve creating visualizations or explanations of the most salient acoustic features associated with depression, helping to bridge the gap between machine learning outputs and clinical decision-making. In this regard, developing adaptive disentanglement methods that can dynamically adjust the degree of speaker disentanglement based on specific clinical requirements or privacy regulations could enhance the flexibility and applicability of these systems in various healthcare settings.

Finally, the implementation of these methods in real-time systems for continuous monitoring and early intervention represents a challenging but potentially highly impactful direction for future work. This would involve optimizing the computational efficiency of the proposed algorithms and developing strategies for handling streaming audio data in privacy-preserving ways.

By pursuing these future research directions, the field of speech-based depression detection can continue to advance, ultimately leading to more accurate, privacy-preserving, and clinically valuable diagnostic tools. These advancements have the potential to significantly impact mental health care, enabling earlier detection, more personalized treatment, and improved outcomes for individuals suffering from depression and other mental health conditions.

Bibliography

[1] U Rajendra Acharya, Vidya K Sudarshan, Hojjat Adeli, Jayasree Santhosh, Joel EW Koh, and Amir Adeli. Computer-aided diagnosis of depression using eeg signals. *European neurology*, 73(5-6):329–336, 2015.

[2] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan. Effectiveness of Voice Quality Features in Detecting Depression. In *Proc. Interspeech 2018*, pages 1676–1680, 2018.

[3] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Alan McCree, and Abeer Alwan. Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification. In *Interspeech 2020*, pages 4318–4322, 2020.

[4] Sharifa Alghowinem, Roland Goecke, et al. Detecting depression: a comparison between spontaneous and read speech. In *ICASSP*, pages 7547–7551. IEEE, 2013.

[5] Nancy JC Andreasen et al. Linguistic analysis of speech in affective disorders. *Archives of General Psychiatry*, 33(11):1361–1367, 1976.

[6] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proc. CVPR*, pages 5297–5307, 2016.

[7] Alexei Baevski et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[8] Andrew Bailey et al. Gender bias in depression detection using audio features. In *2021 29th EUSIPCO*, pages 596–600. IEEE, 2021.

[9] Andrew Bailey and Mark D Plumbley. Raw audio for depression detection can be more robust against gender imbalance than mel-spectrogram features. *arXiv preprint arXiv:2010.15120*, 2020.

[10] Sweta Bhadra and Chandan Jyoti Kumar. An insight into diagnosis of depression using machine learning techniques: a systematic review. *Current Medical Research and Opinion*, 38(5):749–771, 2022.

[11] Suhas Bn and Saeed Abdullah. Privacy sensitive speech analysis using federated learning to assess depression. In *ICASSP*, pages 6272–6276. IEEE, 2022.

[12] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[13] Carlos Busso, Murtaza Bulut, et al. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[14] Carlos Busso et al. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2016.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[16] N Chinchor. Muc-4 evaluation metrics in proc. of the fourth message understanding conference 22–29, 1992.

[17] Karol Chlasta et al. Automated speech-based screening of depression using deep convolutional neural networks. *Procedia Computer Science*, 164:618–628, 2019.

[18] Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, and Jarek Krajewski. Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 970–974. IEEE, 2014.

[19] Nicholas Cummins, Stefan Scherer, et al. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[20] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech*, pages 3830–3834, 2020.

[21] Yazheng Di et al. Using i-vectors from voice features to identify major depressive disorder. *Journal of Affective Disorders*, 288:161–166, 2021.

[22] Yazheng Di, Joel Mefford, Elior Rahmani, Jinhan Wang, Vijay Ravi, Aditya Gorla, Abeer Alwan, Tingshao Zhu, and Jonathan Flint. Genetic association analysis of human median voice pitch identifies a common locus for tonal and non-tonal languages. *Communications Biology*, 7(1):540, 2024.

[23] Yazheng Di, Elior Rahmani, Joel Andrew Mefford, Jinhan Wang, Vijay Ravi, Aditya Gorla, Abeer Alwan, Kenneth S Kendler, Tingshao Zhu, and Jonathan Flint. Unraveling the associations between voice pitch and major depressive disorder: A multisite genetic study. *medRxiv*, pages 2024–10, 2024.

[24] S Pavankumar Dubagunta et al. Learning voice source related information for depression detection. In *ICASSP*, pages 6525–6529. IEEE, 2019.

[25] Sri Harsha Dumpala, Katerina Dikaios, Sebastian Rodriguez, Ross Langley, Sheri Rempel, Rudolf Uher, and Sageev Oore. Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. *Scientific Reports*, 13(1):11155, 2023.

[26] Sri Harsha Dumpala et al. Significance of speaker embeddings and temporal context for depression detection. *arXiv preprint arXiv:2107.13969*, 2021.

[27] Sri Harsha Dumpala et al. Sine-Wave Speech and Privacy-Preserving Depression Detection. In *Proc. SMM21, Workshop on Speech, Music and Mind 2021*, pages 11–15, 2021.

[28] Sri Harsha Dumpala, Sebastian Rodriguez, Sheri Rempel, Mehri Sajjadian, Rudolf Uher, and Sageev Oore. Detecting depression with a temporal context of speaker embeddings. *Proc. AAAI SAS*, 2022.

[29] Sri Harsha Dumpala, Rudolf Uher, Stan Matwin, Michael Kiefte, and Sageev Oore. Sine-wave speech and privacy-preserving depression detection. In *Proc. SMM21, Workshop on Speech, Music and Mind*, volume 2021, pages 11–15, 2021.

[30] José Vicente Egas-López, Gábor Kiss, Dávid Sztahó, and Gábor Gosztolya. Automatic assessment of the degree of clinical depression from speech using x-vectors. In *ICASSP*, pages 8502–8506. IEEE, 2022.

[31] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. 18th ACM-MM*, pages 1459–1462, 2010.

[32] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of ACM Multimedia*, pages 1459–1462, 2010.

[33] Yue Fan, JW Kang, et al. Cn-celeb: a challenging chinese speaker recognition dataset. In *ICASSP*, pages 7604–7608. IEEE, 2020.

[34] Kexin Feng and Theodora Chaspari. Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels. In *IEEE-EMBS ICBHI*, pages 01–07. IEEE, 2022.

[35] Daniel Joseph France et al. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[36] Yaroslav Ganin et al. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[37] Daniel Garcia-Romero and Carol Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech 2011*, pages 249–252, 2011.

[38] Itai Gat et al. Speaker normalization for self-supervised speech emotion recognition. *arXiv preprint arXiv:2202.01252*, 2022.

[39] Larry S Goldman, Nancy H Nielsen, Hunter C Champion, and American Medical Association Council on Scientific Affairs. Awareness, diagnosis, and treatment of depression. *Journal of General Internal Medicine*, 14(9):569–580, 1999.

[40] Ian Goodfellow, Jean Pouget-Abadie, et al. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[41] Amir Harati, Elizabeth Shriberg, et al. Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus. In *ICASSP*, pages 7273–7277. IEEE, 2021.

[42] Lang He and Cui Cao. Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, 83:103–111, 2018.

[43] Wei-Ning Hsu et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[44] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[45] Tzu-hsien Huang, Jheng-hao Lin, and Hung-yi Lee. How far are we from robust voice conversion: A survey. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 514–521. IEEE, 2021.

[46] Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, page 21, 2013.

[47] Spencer L James et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858, 2018.

[48] Alexander Johnson, Kevin Everson, Vijay Ravi, Anissa Gladney, Mari Ostendorf, and Abeer Alwan. Automatic dialect density estimation for african american english. In *Interspeech 2022*, pages 1283–1287, 2022.

[49] Kenneth S Kendler, Steven H Aggen, and Michael C Neale. Evidence for multiple genetic factors underlying dsm-iv criteria for major depression. *JAMA psychiatry*, 70(6):599–607, 2013.

[50] Kurt Kroenke, Tara W Strine, et al. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

[51] Manoj Kumar, Tae Jin-Park, et al. Designing neural speaker embeddings with meta learning, 2020.

[52] Haoqi Li et al. Speaker-invariant affective representation learning via adversarial training. In *ICASSP*, pages 7144–7148. IEEE, 2020.

[53] Lantian Li, Ruiqi Liu, et al. Cn-celeb: multi-genre speaker recognition, 2020.

[54] Lu-Qiao Li, Kai Xie, Xiao-Long Guo, Chang Wen, and Jian-Biao He. Emotion recognition from speech with stargan and dense-dcnn. *IET Signal Processing*, 16(1):62–79, 2022.

[55] Yun Li, S Shi, et al. Patterns of co-morbidity with anxiety disorders in chinese women with recurrent major depression. *Psychological medicine*, 42(6):1239–1248, 2012.

[56] Shih Cheng Liao, Chien Te Wu, Hao Chuan Huang, Wei Teng Cheng, and Yi Hung Liu. Major depression detection from eeg signals using kernel eigen-filter-bank common spatial patterns. *Sensors*, 17(6):1385, 2017.

[57] Zhenyu Liu, Huimin Yu, Gang Li, Qiongqiong Chen, Zhijie Ding, Lei Feng, Zhijun Yao, and Bin Hu. Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection. *Frontiers in Neuroscience*, 17:1141621, 2023.

[58] Daniel M Low et al. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116, 2020.

[59] Samuel D Lustgarten et al. Digital privacy in mental healthcare: current issues and recommendations for technology use. *Current opinion in psychology*, 36:25–31, 2020.

[60] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[61] Xingchen Ma, Hongyu Yang, et al. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42, 2016.

[62] Rigel Mahmood and Bishad Ghimire. Automatic detection and classification of alzheimer's disease from mri scans using principal component analysis and artificial neural networks. In *IWSSIP*, pages 133–137. IEEE, 2013.

[63] Colin D Mathers and Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442, 2006.

[64] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. Publisher: Springer.

[65] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[66] Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proc. 6th AVEC*, pages 43–50, 2016.

[67] A Nilsonne. Speech characteristics as indicators of depressive illness. *Acta Psychiatrica Scandinavica*, 77(3):253–263, 1988.

[68] Paul-Gauthier Noé, Jean-François Bonastre, Driss Matrouf, N. Tomashenko, Andreas Nautsch, and Nicholas Evans. Speech Pseudonymisation Assessment Using Voice Similarity Matrices. In *Proc. Interspeech 2020*, pages 1718–1722, 2020.

[69] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[70] Sarala Padi, Dinesh Manocha, and Ram D Sriram. Multi-window data augmentation approach for speech emotion recognition. *arXiv preprint arXiv:2010.09895*, 2020.

[71] Anastasia Pampouchidou, Panagiotis G Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions of Affective Computing*, 10(4):445–470, 2017.

[72] Soo Jin Park, Amber Afshan, Zhi Ming Chua, and Abeer Alwan. Using voice quality supervectors for affect identification. In *Interspeech*, pages 157–161, 2018.

[73] Daniel Povey, Arnab Ghoshal, et al. The kaldi speech recognition toolkit. In *ASRU*. IEEE Signal Processing Society, 2011.

[74] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *ICML*, pages 18003–18017. PMLR, 2022.

[75] Lawrence Rabiner and Ronald Schafer. *Theory and applications of digital speech processing.* Prentice Hall Press, 2010.

[76] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[77] Srinivasan Ramakrishnan. Recognition of emotion from speech: A review. *Speech Enhancement, Modeling and Recognition–Algorithms and Applications*, 7:121–137, 2012.

[78] Barkha Rani. I-vector based depression level estimation technique. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 2067–2071, 2016.

[79] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[80] Vijay Ravi, Ruchao Fan, Amber Afshan, Huanhua Lu, and Abeer Alwan. Exploring the Use of an Unsupervised Autoregressive Model as a Shared Encoder for Text-Dependent Speaker Verification. In *Proc. Interspeech*, pages 766–770, 2020.

[81] Vijay Ravi, Yile Gu, Ankur Gandhe, Ariya Rastrow, Linda Liu, Denis Filimonov, Scott Novotney, and Ivan Bulyko. Improving accuracy of rare words for rnn-transducer through unigram shallow fusion. *arXiv preprint arXiv:2012.00133*, 2020.

[82] Vijay Ravi, Soo Jin Park, et al. Voice quality and between-frame entropy for sleepiness estimation. *Interspeech*, 2019.

[83] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement. In *Proc. Interspeech*, pages 3338–3342, 2022.

[84] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. Fraug: A frame rate based data augmentation method for depression detection from speech signals. In *ICASSP*, pages 6267–6271. IEEE, 2022.

[85] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. *Computer Speech & Language*, 86:101605, 2024.

[86] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. A privacy-preserving unsupervised speaker disentanglement method for depression detection from speech. In *CEUR workshop proceedings*, volume 3649, page 57. NIH Public Access, 2024.

[87] Emna Rejaibi, Ali Komaty, et al. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107, 2022.

[88] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107, 2022.

[89] Fabien Ringeval et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th AVEC*, 2019.

[90] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9. ACM, 2017.

[91] Atefeh Safayari and Hamidreza Bolhasani. Depression diagnosis by deep learning using eeg signals: A systematic review. *Medicine in Novel Technology and Devices*, 12:100102, 2021.

[92] Afef Saidi, Slim Ben Othman, and Slim Ben Saoud. Hybrid cnn-svm classifier for efficient depression detection system. In *2020 4th International Conference on Advanced Systems and Emergent Technologies*, pages 229–234, 2020.

[93] Michelle Hewlett Sanchez et al. Using prosodic and spectral features in detecting depression in elderly males. In *Interspeech*, pages 3001–3004, 2011.

[94] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Interspeech*, volume 8, pages 2001–2005. ISCA, 2016.

[95] Nadee Seneviratne, James R. Williamson, Adam C. Lammert, Thomas F. Quatieri, and Carol Espy-Wilson. Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression. In *Proc. Interspeech*, pages 4551–4555, 2020.

[96] Ying Shen et al. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. *arXiv preprint arXiv:2202.08210*, 2022.

[97] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[98] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[99] David Snyder, Daniel Garcia-Romero, et al. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333. IEEE, 2018.

[100] Douglas Sturim, Pedro A. Torres-Carrasquillo, Thomas F. Quatieri, Nicolas Malyska, and Alan McCree. Automatic detection of depression in speech using Gaussian mixture modeling with factor analysis. In *Proc. Interspeech*, pages 2981–2984, 2011.

[101] BN Suhas et al. Privacy sensitive speech analysis using federated learning to assess depression. *ICASSP*, 2022.

[102] W Ter Smitten, MH and Smeets, RMW and Van den Brink. Composite international diagnostic interview (CIDI), version 2.1. *Amsterdam: World Health Organization*, pages 343–345, 1998.

[103] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al. The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, 74:101362, 2022.

[104] Michel Valstar, Jonathan Gratch, et al. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

[105] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *CSTR*, 2016.

[106] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion. In *Proc. Interspeech*, pages 1344–1348, 2021.

[107] Dong Wang, Yanhui Ding, Qing Zhao, Peilin Yang, Shuping Tan, and Ya Li. ECAPA-TDNN Based Depression Detection from Clinical Speech. In *Proc. Interspeech*, pages 3333–3337, 2022.

[108] Jinhan Wang, Vijay Ravi, and Abeer Alwan. Non-uniform speaker disentanglement for depression detection from raw speech signals. *arXiv preprint arXiv:2306.01861*, 2023.

[109] Jinhan Wang, Vijay Ravi, Jonathan Flint, and Abeer Alwan. Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals. In *Proc. Interspeech*, pages 2018–2022, 2022.

[110] Jinhan Wang, Vijay Ravi, Jonathan Flint, and Abeer Alwan. Speechformer-ctc: Sequential modeling of depression detection with speech temporal classification. *Speech Communication*, 163:103106, 2024.

[111] Wen Wu, Chao Zhang, and Philip C Woodland. Self-supervised representations in speech-based depression detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[112] Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. *NIPS*, 34:26523–26535, 2021.

[113] Mingke Xu, Fan Zhang, et al. Speech emotion recognition with multiscale area attention and data augmentation. In *ICASSP*, pages 6319–6323. IEEE, 2021.

[114] Le Yang, Dongmei Jiang, and Hichem Sahli. Feature augmenting networks for improving depression severity estimation from speech signals. *IEEE Access*, 8:24033–24045, 2020.

[115] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb:

Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[116] Ying Yang et al. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150, 2012.

[117] Yufeng Yin et al. Speaker-invariant adversarial domain adaptation for emotion recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 481–490, 2020.

[118] William WK Zung. A self-rating depression scale. *Archives of general psychiatry*, 12(1):63–70, 1965.

[119] Lishi Zuo and Man-Wai Mak. Avoiding dominance of speaker features in speech-based depression detection. *Pattern Recognition Letters*, 173:50–56, 2023.