UNIVERSITY OF CALIFORNIA

Los Angeles

Towards Better and Privacy-Preserving Speech Modeling for Depression Detection

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

by

Jinhan Wang

2024

ABSTRACT OF THE DISSERTATION

Towards Better and Privacy-Preserving Speech Modeling for Depression Detection

by

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

Automatic depression detection systems based on speech signals have recently garnered significant attention. Depression modeling from speech signals, however, faces three challenges. The first challenge is data scarcity. The second, is the risk of privacy exposure in depression detection systems. The third, is the lack of consideration of non-uniformly distributed depression patterns within speech signals. In this dissertation, we address these challenges so that better and privacy-preserving speech-based depression detection systems are built.

To address the data scarcity issue, we propose a modified Instance Discriminative Learning (IDL) pre-training method to enable the model to extract augment-invariant and instance-spread-out embeddings from pre-training tasks using out-of-domain unlabeled data. The pre-trained model is then used for initialization of the downstream model, DepAudioNet, and fine-tuned for depression detection tasks. We investigate different augmentation techniques and instance sampling strategies in the pre-training stage. Specifically, we propose a novel sampling strategy, Pseudo-Instance-based Sampling (PIS), to further reveal the correlation between depression characteristics with the underlying acoustic units.

Second, to address the privacy-preservation issue, we propose a novel non-uniform speaker disentanglement (NUSD) adversarial learning framework to disentangle speaker-identity information from depression characteristics. The approach utilizes idiosyncratic behaviors of different layers of detection models, and varies the adversarial disentanglement strength of different model components. The method shows that depression detection can be done without an over-reliance on speaker-identity features. More importantly, we found that attenuating more speaker information in the Feature Extraction (FE) module yields better performance than assigning the disentanglement weights uniformly.

Third, to address the non-uniformity of depression patterns in speech signals, we propose a novel framework. The framework, Speechformer-CTC, models dynamically varying depression characteristics within speech segments using a Connectionist Temporal Classification (CTC) objective function without the necessity of input-output alignment. Two novel CTC-label generation policies, namely the Expectation-One-Hot and the HuBERT policies, are proposed and incorporated in objectives at various granularities. Additionally, experiments using Automatic Speech Recognition (ASR) features are conducted to demonstrate the compatibility of the proposed method with content-based features. Our findings show that depression detection can benefit from modeling non-uniformly distributed depression patterns and the proposed framework can potentially be used to determine significant depressive regions in speech utterances.

Experiments show that the proposed techniques achieve state-of-the-art performance and are validated for both English and Mandarin Chinese.

The dissertation of Jinhan Wang is approved.

Lin Yang

Gregory J. Pottie

Jonathan Frederic R. Flint

Abeer A. Alwan, Committee Chair

University of California, Los Angeles

2024

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not be possible without the support of many people. First and foremost, I would like to sincerely thank my doctoral advisor, Professor Abeer Alwan. She has always been a great mentor and my model of a researcher, shedding light on my academic journey and guiding me toward my goal. With her consistent support and under her supervision, I was able to explore unknown directions of the field of speech processing and contribute to the community. More importantly, I have learned to be a better researcher and a good person.

I am immensely grateful to the members of my Ph.D. committee. Thanks to Prof. Gregory Pottie, and Prof. Lin Yang from the University of California, Los Angeles (UCLA), Department of Electrical and Computer Engineering, and Prof. Jonathan Flint, from UCLA, Department of Psychiatry and Biobehavioral Sciences, for their invaluable suggestions and comments on my dissertation. Special thanks to Prof. Flint and his group members for their assistance in the CONVERGE data collection on depression research.

It is a great honor to work with amazing mentors during my internships. Many thanks to Dr. Xiaosu Tong, Dr. Jinxi Guo, Dr. Long Chen, and Dr. Di He for mentoring me at Amazon Alexa AI Acoustic Modeling team. I've learned a lot from discussions with them, especially a differential insight from academic and industrial perspectives. Their consistent support and suggestions keep directing and encouraging me all the time even after the internship. Many thanks to Dr. Andreas Stolcke for guiding me through my large language model project during my third-time internship at Amazon. The study won't be possible without his instruction and suggestions. I am grateful to Amazon mentors, Tobias Menne, Aparna Khare, Anirudh Raju, Pranav Dheram, Minhua Wu, and Venkatesh Ravichandran for their valuable comments and suggestions about my research. It is a great pleasure to collaborate with them. I also want to express my gratitude to Dr. Myungjong Kim and Dr. Oluwatobi Olabiyi, for their support during my current internship at Nvidia.

My sincere thanks go to my former and current labmates at the Speech Processing and Auditory Processing Laboratory (SPAPL). I've spent most of my time with them and obtained countless inspiration and encouragement from them. Their consistent help is one of the most crucial factors for me to arrive at the current stage. Thanks to Amber, Gray, AJ, Vishwas, Balaji, Eray, Vijay, Yunzheng and Ruchao. Special thanks to Yunzheng, with whom I collaborated on our first academic paper, and who helped me a lot when I was struggling. Sincere thanks to Ruchao Fan for literally everything I've got help from him throughout these years. He is like an older brother who happens to know everything in this domain and can always come up with a solution. Many of my studies wouldn't be possible without the discussion and inspiration from him.

Special thanks to Vijay Ravi, who almost spent his entire Ph.D career with me and tolerated many of my emotional moments throughout our collaboration. I am truly thankful for having his company through this long journey.

I would like to extend my thanks to the administrative staff of the Department. Thanks to Deeona Columbia, Ylena Requena, Tatyana Vigil, Jacquelyn Trang, Lorena Sanchez, Yadira Trejo, and Jose Cano, for their support of my various requests throughout my Ph.D time at UCLA.

I am grateful to have amazing roommates, Xiao Han and Qiao Mi, during my time in Los Angeles. Their company made me feel not alone abroad. Special thanks to Xiao Han, with whom I spent 4 years together and who took care of me on many things. I also would like to express my thanks to my brother, Mingzhe Lv, for his consistent support for me which helps me go through lots of sleepless nights.

Last but not least, I want to thank my father, Hongfu Wang, and mother, Yanjie Yu, for their encouragement and support throughout this period. I wouldn't be the person nowadays without them. Thanks to them for taking care of the whole family these years so I can bury myself in research.

VITA

2015–2019	Bachelor of Engineering, Automation, Beijing Jiaotong University, Beijing, China
2017–2019	Bachelor of Electrical Engineering, University of Minnesota, Twin Cities, Minneapolis, Minnesota
2019–2021	M.S. Electrical and Computer Engineering, University of California, Los Angeles, Los Angeles, California
2021	Applied Scientist, Intern, Amazon, Los Angeles, California
2022	Applied Scientist, Intern, Amazon, Seattle, Washington
2023	Applied Scientist, Intern, Amazon, Sunnyvale, California
2024	Deep Learning Intern - Speech AI, Nvidia, Santa Clara, California

PUBLICATIONS

Jinhan Wang, Vijay Ravi, Jonathan Flint and Abeer Alwan, ”Speechformer-CTC: Sequential Modeling of Depression Detection with Speech Temporal Classification”, Speech Communication, 2024

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pravan Dheram, Di He, Minhua Wu, Andreas Stolcke and Venkatesh Ravichandran, ”Turn-taking and Backchannel Prediction with Acoustic and Large Language Model Fusion”, ICASSP 2024

Vijay Ravi, **Jinhan Wang**, Jonathan Flint, and Abeer Alwan, ”Enhancing Accuracy and Privacy in Speech-based Depression Detection through Speaker Disentanglement” in Computer Speech & Language, 2024

Jinhan Wang, Vijay Ravi and Abeer Alwan, ”Non-uniform Speaker Disentanglement for Depression Detection From Raw Speech Signals”, Interspeech 2023

Jinhan Wang, Vijay Ravi, Jonathan Flint and Abeer Alwan, ”Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals”, Interspeech 2022

Jinhan Wang, Xiaosu Tong, Jinxi Guo, Di He and Roland Maas, ”VADOI: Voice-activity-detection Overlapping Inference for End-to-end Long-form Speech Recognition”, ICASSP 2022

Jinhan Wang, Yunzheng Zhu, Ruchao Fan, Wei Chu and Abeer Alwan, ”Low Resource German ASR with Untranscribed Data Spoken by Non-native Children – Interspeech 2021 Shared Task SPAPL System”, Interspeech 2021

CHAPTER 1

Introduction

1.1 Motivation

Major Depressive Disorder (MDD) is one of the most severe chronic mental health disorders, characterised by persistent depressed mood, loss of interest in social activities, lack of energy, and in several cases, thoughts of self-harm or even suicide [LJD14]. Around 4.4% of the world's population is estimated to suffer from depression [Org17]. According to the World Health Organization, MDD may become the leading cause of global disease burden by 2030 [CPJ11].

Conventionally, diagnosing depression is conducted by self-reported questionnaires [KS09]. Some notable ones are Patient Health Questionnaire depression scale (PHQ) [KS09], Composite International Diagnostic (CIDI) [KAM98], and Hamilton Depression Rating Scale (Ham-D) [BRS04]. These questionnaires include rating questions related to depressed mood, insomnia, and somatic symptoms, etc. The self-reported survey can be further used during the clinical interview with an expert doctor for reference. However, there are many subjective factors, which might cause mis-diagnosis, such as inaccurately-reported depressive symptoms, irregular patient history or the expertise of the doctor [KYK21]. As a consequence, these conventional depression evaluation protocols result in high diagnostic error [Mit10]. Therefore, there is an urgent need to develop an automatic depression detection system that can aid doctors in diagnosing depression with the help of artificial intelligence.

1.2 Automatic Depression Detection from Speech Signals

There have been extensive studies focusing on automatic depression detection from biosignals, such as voice, video, and electroencephalogram (EEG), etc. [HEJ19, PMT15, WLG15, ASA15]. Among those approaches, automatic depression detection based on speech signals draws more attention because speech signals can be an effective clinical marker for depression [CS15, LDG14]. Recently, depression detection through verbal cues has gained substantial traction, largely due to the ease of collecting speech data and the rapid advances in neural network-based modeling methods. Previous research has extensively examined speech-based depression detection from various perspectives, encompassing diverse acoustic features (vocal source [Dub19, KBB23], voice quality [AG18], vocal tract articulators [CVS17]), model types (spatial [YDX23], Recurrent Neural Networks (RNNs) [MY16], self-attention [ZLC20]), data augmentation (Generative Adversarial Network (GAN) [YJS20], FrAUG [RWF22a]) and backend modeling techniques (transfer learning [WWY22], self-supervised pre-training [ZWD21, WRF22]). We will introduce relevant background in this section.

1.2.1 System Description

Automatic depression detection is formulated as a classification task for depression vs non-depression or multi-level depression severity detection. Some studies also investigated regression models to predict severity scores. In this dissertation, we specifically focus on classifying an individual as depressed or non-depressed, as shown in Figure 1.1. Given an utterance from an individual, acoustic features are obtained by passing the input speech into a feature extractor. The extracted features are then fed into the model for depression detection. The final decision is made by either thresholding a single confidence score or choosing the larger score between the two classes.

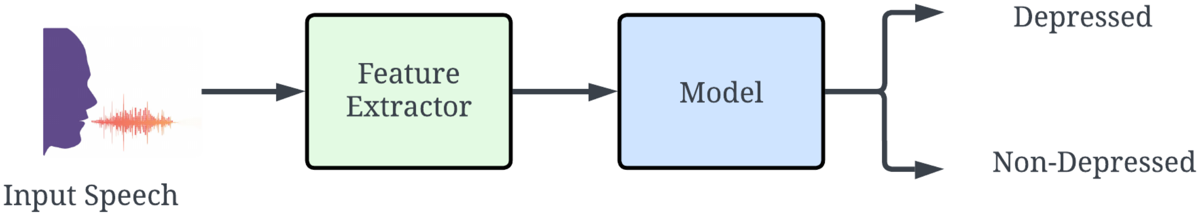


Figure 1.1: A schematic diagram of a speech-based automatic depression detection system.

1.2.2 Acoustic Features

Numerous studies [CS15, WWL23, AG18, Low11, LH15] have shown that the speech of individuals suffering from depression can be characterised by one or more of the following acoustic cues: slow, uniform, monotonous, and hesitant voice. Among various acoustic features that have been investigated as bio-markers are voice source features, which capture voice production information, such as jitter, shimmer, and Harmonic-to-noise ratio (HNR); these features are shown to have a high correlation with depression [ASS14, LML10]. In [Yan12], vocal prosody features, especially switching pause (pause duration between speaker utterances), were found to have great inter and intrapersonal characteristics related to depression severity. Moreover, since changes in speech motor control are proven to be associated with depression, corresponding formant features that represent dominant spectrum components and contain resonance properties of the vocal tract were also used for depression detection in a gender-dependent manner [CVS17]. Cepstral features, like Mel-Frequency Cepstral Coefficients (MFCC), are also widely used as input features across multiple frameworks for depression detection [RKM22, Sha20, HCM18]. Specifically, in [TTN18], the second coefficient of MFCC, which represents the energy difference of frequency around 2k to 3k Hz, has been shown to have significant discriminative characteristics between depressive and non-depressive speech.

1.2.3 Model Architecture

Automatic depression detection systems were previously built based on architectures like Gaussian Mixture Models [WQH13], logistic regression [AG18], Naive Bayes and Random Forest [LHL16]. With the advancements in Deep Neural Networks (DNNs), more studies have increasingly adopted DNN-based architectures, showcasing improved performance compared to previous machine learning approaches [MY16, RKM22, LDL19]. Notably, Convolutional Neural Networks (CNNs), RNNs, and their variants have gained prominence in the research community. DepAudioNet [MY16], which consists of CNN and Long-Short-Term-Memory (LSTM) layers, has proven effective in depression detection, and it is widely used as a baseline model [RWF22a, RWF22b, WRA23, RWF24b, RWF24a]. In [HEJ20], a dilated CNN is proposed and coupled with full vocal tract coordination features to leverage its potential in determining depression states. [ZLD21] integrates a multi-head attention mechanism with LSTM to emphasize key temporal information, enhancing the performance of depression detection tasks.

Recently, Transformer-based approaches have gained more attention for their great adaptability across various domains and superior performance. In [LLL22], the Transformer encoder is cascaded with a CNN to capture long-range dependencies. In [YDX23], a parallel CNN-Transformer architecture to simultaneously capture local knowledge and temporal sequential information was proposed. More recently, a new framework, Speechformer, was proposed by modifying multi-head self-attention into a speech-based variant with hierarchical granularities (Frame, Phoneme, Word or Utterance) based on speech pronunciation structure [CXX22]. It was shown that the Speechformer model achieved superior performance compared to its Transformer counterpart.

1.3 Data Scarcity Challenge

Depression corpora suffer from the data scarcity problem, which hinders the generalizability of automatic detection systems. One possible solution is applying data augmentation to simulate acoustic variability among the depression/non-depression groups. But conventional augmentation techniques might alter the depression status, such as pitch, formants, and speed. etc. [CS15]. Some augmentation techniques were validated to perform well on depression detection [Fen22, RWF22a], but this topic is still in a preliminary stage without extensive experimental verification. Another solution is to use pre-trained models which we will describe in this section. Alternatively, to provide a better generalizability of the detection model, utilizing unlabelled data through Unsupervised-Pre-Training (UPT) is another solution. We introduce two UPT methods in Chapter 3.

1.3.1 Contrastive Predictive Coding

Contrastive Predictive Coding (CPC) [OLV18] is a universal unsupervised learning method in extracting high-level representations by optimizing the probabilistic contrastive loss in predicting the future autoregressively. Many UPT methods extended from CPC have shown their great performance in speech-related tasks, such as Wav2vec [SBC19], Wav2vec2.0 [Bae20], Vq-Wav2vec [BSA19]. However, here we focus on CPC as it's the basic framework design without the requirement for specific architecture like its extensions.

Given an input sequence $[x_1, ..., x_t, ...x_T]$, a non-linear encoder g_{enc} maps the input sequence into a latent representation sequence $[z_1, ..., z_t, ...z_T]$. Then a sequence of context latent representation $[c_1, ..., c_t, ...c_T]$ is obtained by autogressively mapping $z_{\leq t}$ through a RNN module g_{ar}. Therefore, each context latent representation c_t has aggregate information from $z_{\leq t}$. For prediction, instead of predicting the k-step future observations x_{t+k} directly, like what Autoregressive Predictive Coding (APC) does, CPC models the future prediction density ratio $f_k(x_{t+k}, c_t)$ over a group of N samples X, where only one sample drawn from $P(x_{t+k}|c_t)$

is positive, and the rest ($N-1$ samples) are negative, and drawn from some pre-defined distribution. For each future time step k, the density ratio function can be modeled using a simplified log-bilinear model. As a consequence, the model is optimized to distinguish the only true positive sample from a group of candidates with $N-1$ distractors using a InfoNCE loss. This objective can be formulated as:

$$\begin{aligned} L_N &= -\mathbb{E}_{\mathbb{N}}[log\frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}] \\ f_k(x_{t+k}, c_t) &= exp(z_{t_k}^T W_k c_t) \end{aligned} \tag{1.1}$$

where W_k is the linear projection of the prediction for each time step k. In implementation, InfoNCE loss is aggregated over multiple future time steps k from 1 to max_step.

CPC provides high tractability thanks to the negative sampling mechanism. By controlling the sampling policy, we can manipulate the model to be capable of distinguishing specific attributes. For example, by drawing negative samples strictly from different speakers, CPC can simultaneously capture speaker and content-related information.

1.3.2 Speech SimCLR

Speech SimCLR [JLC20] is speech-variant version of its image-domain counterpart SimCLR [CKN20]. In general, SimCLR methods train the model to maximize the agreement between data samples augmented in different ways in an unsupervised fashion. Similar to the CPC method where a contrastive loss is applied, SimCLR, however, collapses each sample into a fix-dimensional embedding and computes the contrastive loss mutually without the necessity of an autoregressive operation. This setup enables the model to focus more on the global attributes and improves the robustness of the model toward perturbation.

Given a sample x, two augmentations are applied to generate a correlated pair of samples x_i and x_j. These two samples then pass through a neural network module in parallel, for example, a Transformer encoder, to generate two latent representations h_i and h_j. An aggregation operation is then used to collapse the representation to be fixed in dimension z_i and z_j for contrastive loss computation. This operation is typically modeled using an average pooling operation followed by a projection operation. For a batch of N samples, with respect to a sample i, consider (i, j) to be the positive pair and the rest $2(N-1)$ to be negative pairs, the model is trained to optimize the constrastive loss defined as:

$$
\begin{aligned}
L_{i,j} &= -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{k \neq i} exp(sim(z_i, z_k)/\tau)} \\
sim(z_i,z_j) &= \frac{z_i^T z_j}{||z_i||||z_j||} \\
\mathbb{I}_{k \neq i} &= 1 \text{ if } k \neq i \text{ else } 0
\end{aligned}
\tag{1.2}
$$

where the *sim* function is the cosine similarity between two vectors which can also be substituted into other functions, and τ is the pre-defined temperature factor to scale the similarity score.

1.4 Privacy Concern of Depression Detection Systems

Despite the extensively developed speech-based depression detection systems and applications, one inevitable concern is the risk of privacy exposure when people are using such systems. According to a survey regarding patients' attitude towards speech-based mental health smartphone application [BNG21], more than 25% of patients with a mental health disorder indicated that they will object to the application or will only consider using it if privacy is guaranteed. In addition, the ratio of absolute objection increases to 49% if they were informed that the application will record their speech automatically instead of manually. As a consequence, privacy concerns will hinder the deployment of automatic depression detection systems in both clinical and daily use cases.

One notable type of malicious hack in machine learning that may lead to a privacy breach is the Membership Inference Attack (MIA) [HSS22]. MIAs aims to infer whether a specific data sample is included in the dataset for model training. In the case of speech-based automatic depression detection, attackers can query the system using an individual's utterance and discover if that individual has a depression disorder. Such an attack leads to confidentiality violation with respect to speaker identity.

Thus, as a privacy-preserving automatic depression detection system, acoustic features extracted from either the feature encoder or the intermediate layers should be distinguishable with respect to depression states but not to speaker identity.

1.5 Non-uniform Depression Patterns

Aside from the challenges mentioned, depression patterns may not be uniformly distributed in an utterance, implying that not every frame, phoneme, or sub-word may equally convey depression (or non-depression) related information [SCP17,NLT21,WLZ21]. Previous research has unveiled that depression severity affects different regions of speech segments disparately, considering both time and frequency intervals [NLT21]. This influence is evident in word choices [CGS23] and variations in vowels between individuals with depression and those without [YLC23].

A straightforward way of solving this problem is by training human experts to manually label each audio segment and provide more fine-grained depression status labels either at the frame, phoneme, or word level instead of just one label per speaker. This process is expensive, time-consuming, and may introduce annotator bias because of the subjective labeling process [KYK21]. As a consequence, previous modeling approaches make the implicit assumption of "label extension", i.e., each chopped segment or even frames share the same depressive or non-depressive label as the overall label of the corresponding speaker. This assumption may lead to sub-optimal performance, as it fails to account for the non-uniformly distributed depression intervals as mentioned earlier.

1.6 Dissertation Overview

This dissertation addresses three challenges in the development of automatic depression detection systems. First, we mitigate the data scarcity issue in depression detection from speech signals using unsupervised pretraining techniques. We propose a modified UPT method (Instance Discriminative Learning) framework to learn prior knowledge from pretraining datasets. Second, we propose a non-uniform speaker disentanglement method to better preserve speaker's privacy through adversarial training while simultaneously improve detection performance. It should be noted that anoymization of the speech-based depression detection system is especially crucial for clinical deployment. Lastly, a novel framework, Speechformer-CTC is introduced to explicitly model non-uniform depression characteristic in the temporal domain.

The remainder of this dissertation is organized as follows. Chapter 2 describes datasets and features used throughout the dissertation. Chapter 3 introduces the novel UPT method, Instance Discriminative Learning, to address the data scarcity challenge in depression detection. Chapter 4 includes the adversarial training frameworks to tackle the privacy concerns of depression detection systems. Chapter 5 presents a way of modeling non-uniform depression characteristics across segments with the proposed framework, Speechformer-CTC. Chapter 6 concludes the dissertation and provides some directions for future research.

CHAPTER 2

Databases and Features

2.1 Databases

2.1.1 DAIC-WOZ

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [VG16] is an English dataset containing audio, text, and video of interviews collected by a virtual interviewer from 189 (male and female) participants. Each participant conducts a PHQ [KS09] questionnaire where a score sum greater or equal to 10 indicates depression. The audio files have durations ranging from $7 \sim 33$ min (16 min on average), with a non-depression vs depression ratio of 3:1 and total duration $\sim$50 hours, sampled at 16kHz. Segments are extracted from the responses of speakers using the provided time stamps. The dataset is split into train, validation, and test sets with a ratio of $107\colon\ 35\colon\ 47$, following the official split. Sessions with errors are properly handled manually based on the provided documentation.

2.1.2 CONVERGE

The Mandarin dataset used is part of China, Oxford, and Virginia Commonwealth University Experimental Research Genetic Epidemiology (CONVERGE) project [LS12]. Subjects were interviewed by trained interviewers using a computerized assessment system. The diagnoses of depressive disorder were conducted by expert clinicians according to the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) criteria [Dia94]. The dataset is split into CONVERGE1 and CONVERGE2 based on the date of collection. The

CONVERGE1 subset comprises 7959 female speakers (4217 ND vs. 3742 D) with a total duration of ~436 hours, sampled at 16kHz. Responses for each participant are segmented into multiple audio clips by annotators. For training, validation, and testing the model performance, the CONVERGE1 dataset is split into 60%, 20%, and 20%, respectively, without speaker overlap. CONVERGE2 is a subset sampled in 8kHz for replication study purposes and to test the proposed methods' performance in an out-of-domain scenario. CONVERGE2 contains 1189 female speakers (699 ND vs 490 D) with a total duration of ~71 hours. Note that two CONVERGE datasets have different recording conditions.

2.2 Features

2.2.1 Mel-spectrogram Features

Mel-spectrograms are the most widely applied feature sets in depression detection from speech signals [Fen22, CXX22, CXX23]. After the Short-Time Fourier Transform (STFT), each frame's frequency content is filtered by a set of Mel-filters, characterized by a human perceptual frequency scale with narrower low-frequency filter bandwidths than the high-frequency ones. Their reasonable capability in representing spectral envelopes makes them a standard feature set in various speech tasks, such as Automatic Speech Recognition (ASR) [WZF21, FAA21, FCC23], Speaker Verification [JZW18, TD23, VLM14], and Speech Emotion Recognition (SER) [MCL21, DWL19].

In this dissertation, mel-spectrogram features with different window and hop sizes are extracted depending on the backend models in different studies.

2.2.2 Raw Audio

Aside from extracting hand-crafted features like mel-spectrograms, using raw audio directly as input is an emerging trend in the speech community [RWF24a, FSA24, PC15, ODZ16].

With the advancement of deep neural networks, models are not only built to conduct specific tasks given extracted features but also able to adaptively extract task-specific features using a learnable feature encoder [HBT21,Bae20,SBC19], which is usually a convolution module. Some notable studies show that intermediate features extracted from learnable feature encoders are more representative than mel-spectrograms for depression detection [RWF24a, Bai21].

2.2.3 High-level Features

High-level features, defined as features extracted from models pre-trained on some tasks, are extensively explored in various speech tasks. These models are usually pre-trained on much larger and more diverse datasets, and frozen in downstream tasks. It is expected that prior knowledge learned from pre-training tasks can help the models extract better representations than those obtained from statistical functions. Here we introduce two related and most popular pre-training frameworks.

2.2.3.1 HuBERT

Hidden unit BERT (HuBERT) [HBT21] is a self-supervised pre-training framework which utilizes an intermediate clustering phase to generate pseudo-labels for masked prediction loss optimization. Suppose we have a sequence of raw speech signals, the feature encoder projects the raw input audio into a sequence of representations $Z = [z_1, ..., z_t, ...z_T]$. The feature encoder usually consists of several convolution layers. The representations are randomly masked with a pre-defined ratio, where the masked time steps are denoted as a set M. The backbone model f, which is a Transformer encoder module, maps the partially masked representation sequence into label spaces for cross-entropy loss calculation with respect to the generated pseudo-label $Y = [y_1, ..., y_t, ...y_T]$. Pseudo-labels are generated through an acoustic unit discovery process k using a clustering operation, which is usually a K-means model. The cluster model can be fit using either Mel-frequency Cepstral Coefficients (MFCCs) extracted from raw speech signal or intermediate HuBERT model layer's output. The HuBERT loss can be formulated as:

$$
\begin{aligned}
L_{HuBERT} &= \alpha L_{masked} + (1-\alpha) L_{unmasked} \\
&= -\alpha \sum_{t \in M} logP(y_t|Z) - (1-\alpha) \sum_{t \notin M} logP(y_t|Z)
\end{aligned}
\tag{2.1}
$$

where α is the weight factor controlling the loss contribution from masked and unmasked time steps. In implementation, multiple sets of pseudo-labels, for example, generated from a K-means model with different numbers of clusters, can be used simultaneously and losses are summed up. Unless otherwise specified, we use a 24-layer HuBERT-large model and 12-th layer's output as the extracted high-level HuBERT features for the English dataset.

2.2.3.2 Whisper

Whisper [RKX23] models are pre-trained with weak transcription supervision on large and diverse domains of corpora (680k hours) through multi-task training, and have been shown to have great robustness. The multi-task training is accomplished by manipulating the input tokens to the decoder, such as specifying whether the task is transcription or translation across different languages. The unified model setup results in great ASR performance. A single model shows great robustness in multilingual scenarios without the need for fine-tuning.

Whisper is an encoder-decoder Transformer model with 80-channel log-mel-spectrogram as input features. In this dissertation, Whisper-medium with 24 layers are selected for the Mandarin datasets because of its superior ASR performance in Mandarin compared to other model sizes.

CHAPTER 3

Mitigating Data Scarcity with UPT in Depression Detection

In this chapter, we attempt at achieving better speech modeling of depression by mitigating data scarcity through an unsupervised pre-training (UPT) technique which was first introduced in [Wan22b].

3.1 Motivation

UPT is shown to be effective in various tasks given limited data, such as low-resource Automatic Speech Recognition (ASR) [HBT21, FAA21], and Speech Emotion Recognition (SER) [LYL21, JLC20]. For SER tasks, Contrastive Predictive Coding (CPC) was applied by separating positive examples from negative examples through InfoNCE (Noise-Constrastive Estimation) minimization [LYL21]. A different UPT technique, Speech SimCLR, was proposed to optimize contrastive loss between samples that are augmented [JLC20]. There have been a few studies that focused on UPT for depression detection. Most of these approaches use UPT models for feature extraction [SEG18, ZWD21, SNR22]. However, it's highly likely that the extracted features might lose important information for depression classification, since the pre-trained model is not optimized for depression detection. Hence, there is a domain discrepancy problem which might negatively affect performance. Inspired by a successful UPT method, Instance Discriminative Learning (IDL) [YZY19], for image classification tasks, we propose a modified IDL approach for depression detection based on speech signals to

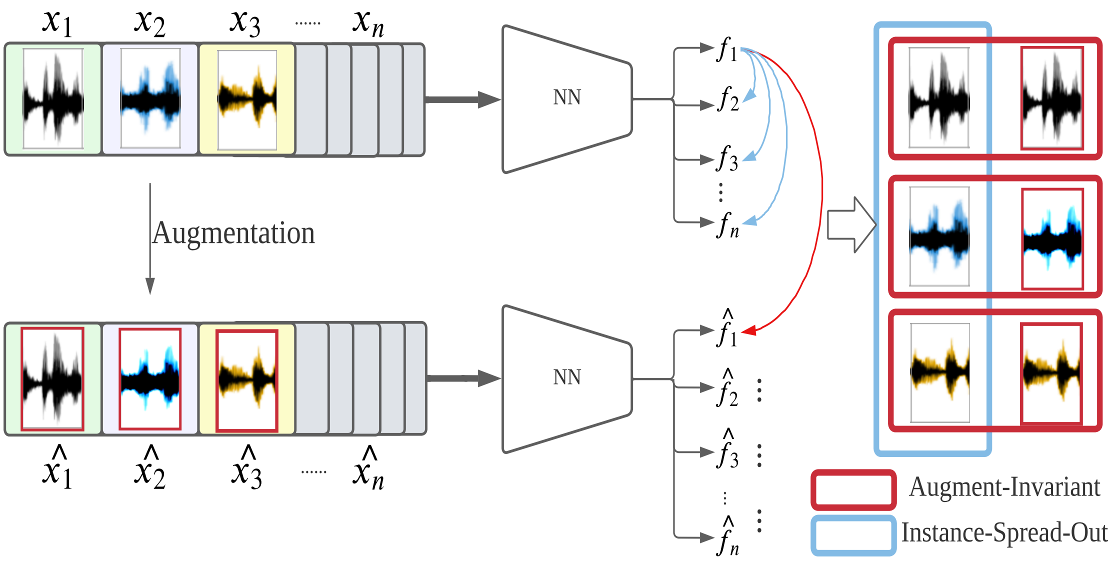


Figure 3.1: A schematic diagram of IDL pre-training. x_i indicates an original instance and $\hat{x}_i$ represents its augmented version, $i = 1, 2, ..., n$, where n is the batch size. f_i and $\hat{f}_i$ denote embeddings of x_i and $\hat{x}_i$. Red and blue blocks are augment-invariant and instance-spread-out computations, respectively.

extract augment-invariant and instance-spread-out embeddings in the pre-training stage. Then, the pre-trained model is used to initialize the downstream model instead of freezing it to be a feature extractor. Various augmentation techniques are applied, and Time Masking is found to yield the best performance. We also investigate different sampling strategies for pre-training and find that the method that works the best preserves some speaker information. Additionally, we propose a new sampling strategy, Pseudo Instance-based Sampling (PIS), to boost instance-spread-out characteristics.

3.2 Method

The schematic diagram of IDL [YZY19] is shown in Figure 3.1. The assumption of IDL pre-training is that, the embedding of an instance and the embedding of its augmented object should be invariant, that is close in the latent space, and embeddings of different instances should be spread-out, that is far away from each other in the latent space. The embedding here is the output of the pre-trained model. In our case, all utterances are divided into segments with equal length, and each segment is taken as a distinct instance in pre-training. For each batch, n segments are selected with pre-defined sampling strategies which are introduced in Section 3.2.2. For each segment x_i in the batch, augmentation is applied and we obtain $\hat{x_i}$. Embeddings, denoted by f_i and $\hat{f}_i$, are obtained by feeding the original segment and the augmented segment into the Neural Network (NN) module, respectively. Assume a hyper-parameter τ, to obtain an augment-invariant embedding, the probability that the augmented instance $\hat{x_i}$ being classified as instance x_i is maximized, and is defined as:

$$P(x_i|\hat{x_i}) = \frac{exp(< f_i, \hat{f}_i > /\tau)}{\sum_{k=1}^{n} exp(< f_k, \hat{f}_i > /\tau)} \tag{3.1}$$

As another objective of the pre-training, embeddings should also be instance-spread-out, where the probability that an instance x_j being classified as another instance x_i, $j \neq i$, in a batch is minimized, and is defined as:

$$P(x_i|x_j) = \frac{exp(< f_i, f_j > /\tau)}{\sum_{k=1}^{n} exp(< f_k, f_j > /\tau)}, j \neq i \tag{3.2}$$

Here, probabilities are calculated as the ratios of exponential inner products of two embeddings scaled by τ. In all experiments, τ is empirically set to be 10. Note that all equations in this section are similar to those introduced in [YZY19].

We assume that the probability of an instance x_i being recognized as a different instance x_j is independent for $j \neq i$. For each instance x_i, the joint probability that its augmented instance $\hat{x_i}$ can be recognized as x_i and other instances x_j cannot be recognized as x_i is denoted as:

$$P_i = P(x_i|\hat{x_i}) \prod_{j \neq i} (1 - P(x_i|x_j)) \tag{3.3}$$

The NN module is then optimized by minimizing the average negative log-likelihood over a batch. Hence, the loss is :

$$L_{IDL} = -\frac{1}{n}(\sum_{i} logP(x_i|\hat{x_i}) + \sum_{i}\sum_{j\neq i} log(1 - P(x_i|x_j))) \tag{3.4}$$

3.2.1 Augment-Invariant Embeddings

The first goal of IDL pre-training is to learn augment-invariant embeddings, corresponding to Equation 3.1. Augmentation techniques are applied to instances in the batch during pre-training. However, for speech signals, effects of augmentation methods on depression status have not been fully explored. To mitigate potential negative effects of some augmentation methods that might change acoustic correlates of depression, such as pitch, formants, etc. [CS15], augmentation methods used in this work are carefully chosen to be additive noise and volume perturbation at the signal level [Ma19]. For comparison, Vocal Tract Length Perturbation (VTLP) [JH13], as a powerful augmentation technique but might change formant features, is also applied. All signal-level augmentations are conducted using open-source nlpaug package [Ma19]. Here signal-level augmentation refers to operations applied directly on the raw waveform before feature extraction. In addition, to better analyze the effects of augmentation in the time and frequency domains separately, TM (time masking), FM (frequency masking) and SpecAugment (as a combination of TM and FM along with time-warping) [PCZ19] are applied at the feature level, using mel-spectrograms.

3.2.2 Instance-Spread-Out Embeddings

3.2.2.1 Distinct speaker-based and Random Sampling

Equation 3.2 results in the embeddings being instance-spread-out. Depending on different sampling strategies, the pre-training task has different implicit tendencies. In distinct speaker-based sampling (DS), it is guaranteed that each segment in a batch is from a distinct speaker. Hence, while the model is trying to spread out embeddings from different instances, it is also optimized to classify different speakers. In other words, speaker information might be preserved in the embedding. The other sampling strategy is random sampling (RS), where segments in a batch are not constrained to be chosen from distinct speakers. If two samples in a batch are from the same speaker, they will still be classified as two instances. Therefore, setting the RS strategy might not preserve speaker discriminative information as well as DS in the embedding.

3.2.2.2 Pseudo Instance-based Sampling (PIS)

Inspired by HuBERT [HBT21], pseudo labels generated by clustering algorithms can reveal implicit correlation between hidden representations and underlying acoustic units in ASR tasks. Embeddings trained from IDL might also have such a non-trivial correlation with depression status. The correlation can provide distinguishable characteristics across instances which may help with depression classification. Therefore, PIS is proposed to sample instances in a batch according to pseudo labels assigned by clustering algorithms instead of being speaker-dependent. Let $\mathbf{X}$ denote all segments $\mathbf{X} = [x_1, x_2, ..., x_n]$. At the first stage, the model is pre-trained using IDL with DS sampling strategy. Then, embeddings, $\mathbf{F} = [f_1, f_2, ...f_n]$, are obtained by feeding $\mathbf{X}$ into the best pre-trained model. Embeddings are clustered using a simple k-means model with C cluster centroids. Corresponding pseudo labels, $\tilde{\mathbf{Y}} = [\tilde{y_1}, \tilde{y_1}, ..., \tilde{y_n}]$, are assigned as cluster centroid variables, where $\tilde{y_i} = 1, 2, ..., C$. At the second stage, instead of sampling instances in a batch with DS, each instance is sampled from a distinct cluster to guarantee all samples in a batch have distinct pseudo labels. Therefore, C is pre-determined to be the batch size in the second stage to guarantee all samples in a batch are from different clusters.

3.3 Experimental Setup

For both pre-training and downstream datasets, we use 40 dimensional mel-spectrograms as input features, extracted with the Librosa library [MRL15] every 32ms with a 64ms Hanning window. Experiments are conducted using PyTorch [PGM19]. Two datasets, DAIC-WOZ (English) and CONVERGE1 subset from CONVERGE (Mandarin), as described in Chapter 2, are used in this study.

3.3.1 Pre-training Datasets

Librispeech-960 [PCP15] is used as the pre-training dataset for downstream tasks on DAIC-WOZ. It is one of the largest publicly available speech corpora in English (960 hours) with mostly reading style speech. Two subsets, training and validation, are partitioned with the ratio of 9:1 without speaker overlap, and the validation set is used to choose the best pre-trained model.

CN-Celeb [FK20] is chosen as the pre-training dataset for classification experiments conducted on CONVERGE1. The dataset contains more than 130,000 utterances in Mandarin from 1000 Chinese celebrities across multiple genres (270 hours). Training and validation subsets are partitioned in the same way as Librispeech.

3.3.2 DepAudioNet

The back-end model we apply is DepAudioNet proposed in [MY16, BP20] for depression detection from speech signals, as shown in Figure 3.2. 40-dimensional mel-spectrogram features first pass through a one-dimensional convolution layer with a kernel size of 3 to capture short-term characteristics. Batch normalization is enabled after convolution. A max-pooling layer with a kernel size of 3 follows the convolution layer to capture mid-term features, with a dropout factor of 0.05 and an activation function as ReLU. Two LSTM layers and a fully connected (FC) layer activated by a Sigmoid with a hidden size of 128 are

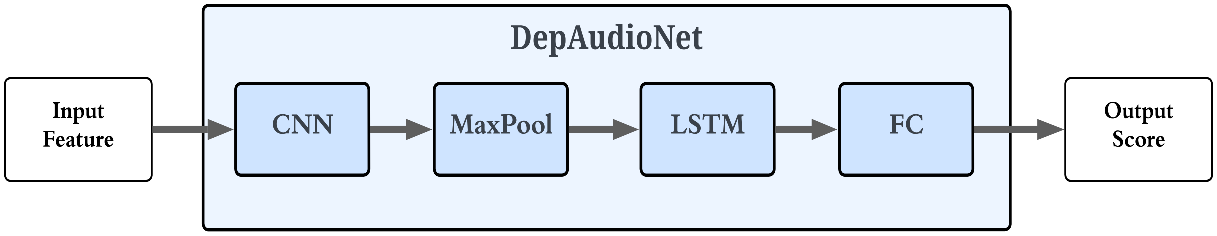


Figure 3.2: A schematic diagram of DepAudioNet Model [MY16]

concatenated to generate binary predictions.

For experiments on DAIC-WOZ, the model is pre-trained on the Librispeech-960 corpus. Utterances from each speaker are combined and segmented into multiple fixed length segments with 120 frames each. The model is trained for 100 epochs with a batch size of 20. The learning rate is set to 1e-3, and the decay factor is set to 0.9 every two epochs. The model with the smallest validation loss is chosen to initialize the downstream model. For the depression classification task, to mitigate length variation and class imbalance, random cropping and random sub-sampling are applied [MY16]. Each utterance is randomly cropped into a fragment with the same length as the shortest utterance, to mitigate any influence from longer utterances. Then, each fragment is segmented with a fixed window length of 120 frames. A new subset is generated by sampling an equal number of depression and non-depression segments randomly without replacement. To fully utilize as many samples as possible, five individual models are trained using five randomly selected subsets for 100 epochs each, with a learning rate of 1e-3, and a decay factor of 0.9 every two epochs. Final predictions are obtained by averaging probabilities of the five models.

For CONVERGE1 experiments pre-trained on CN-Celeb, random cropping and sub-sampling are disabled because the dataset (CONVERGE1) is balanced. One model is trained using all segments. The learning rate is empirically set to 1e-2. Other experimental configurations are the same as DAIC-WOZ experiments. All experiments are evaluated by

Table 3.1: Performance of IDL in terms of average F1-scores with different augmentation methods using distinct speaker-based sampling (DS) on DAIC-WOZ and CONVERGE1 evaluation sets. Baseline is the experiments without pre-training. ND and D stands for non-depression and depression, respectively. In the Augmentation column, TM stands for Time-Masking, FM is Frequency-Masking, and SpecAug stands for SpecAugment. VTLP denotes Vocal Tract Length Perturbation. Noise and Volume stands for noise perturbation and volume perturbation, respectively. The best F1-scores are boldfaced. * indicates that the change in performance is not statistically significant.

Exp	Augmentation	DAIC-WOZ			CONVERGE1		
		F1-avg	F1-ND	F1-D	F1-avg	F1-ND	F1-D
Baseline	NA	0.6083	0.8000	0.4167	0.7228	0.7101	0.7355
IDL (DS) - Feature Level Aug	TM	**0.6458**	0.8116	0.4800	**0.7412***	0.7503	0.7321
	FM	0.6103	0.7761	0.4444	0.7256	0.7343	0.7168
	SpecAug	0.6189	0.8378	0.4000	0.7316	0.7432	0.7200
IDL (DS) - Signal Level Aug	VTLP	0.5428	0.6984	0.3871	0.7197	0.7378	0.7015
	Noise	0.6083	0.8000	0.4167	0.7293	0.7537	0.7048
	Volume	**0.6458**	0.8116	0.4800	0.7257	0.7453	0.7063

macro-average F1-score between depressed and non-depressed speakers.

3.4 Results and Discussion

3.4.1 Comparison of Different Augmentation Methods

To investigate the effects of augment-invariant characteristics, different augmentation methods used in pre-training are compared while fixing the sampling strategy to be DS. F1-scores for the baseline system without pre-training, and IDL pre-training with various augmentation methods on DAIC-WOZ and CONVERGE1 evaluation sets are shown in Table 3.1. The improvements in the F1-scores are statistically significant compared with the baseline unless otherwise specified. In the feature-level augmentation experiments, the effect of time and frequency perturbation on depression status can be analyzed by perturbing the spectrogram in time or frequency. Table 3.1 shows that TM gives the best performance with a relative F1-avg improvement of 6.16% on DAIC-WOZ and 2.55% on CONVERGE compared with the baseline system. FM is the worst among the three with a relative improvement of 0.3% and 0.39% on DAIC-WOZ and CONVERGE1, respectively. Applying SpecAugment achieves an intermediate performance with a relative improvement of 1.74% and 1.22%, respectively. The results suggest that, augmenting the signal in the spectral domain, such as frequency masking, might result in loss of depression-specific information. The worst performance is observed for IDL with VTLP as the augmentation method, which also proves the hypothesis that modifications of spectral domain parameters can negatively affect depression classification. Because the spectrum is less affected by noise and volume perturbations compared with VTLP, moderate improvement can be observed on both datasets using these two types of techniques. Combination of augmentation methods are evaluated but performances are worse than using a single method. Note that the improvements on DAIC-WOZ is more significant than on CONVERGE1, and this could be due to two factors. The first is the fact that the CONVERGE dataset is balanced while DAIC-WOZ is not. The second reason could be that the labels in DAIC-WOZ is based on subjects' self-assessments, while the labels of the CONVERGE1 dataset reflected clinical diagnosis.

3.4.2 Sampling Strategy Comparison

The instance-spread-out characteristics of the proposed pre-training method are explored by setting different sampling strategies with TM augmentation, since TM is the best augmentation technique found experimentally (as reported in Section 3.4.1). F1-avg scores on DAIC-WOZ and CONVERGE1 evaluation sets using different sampling strategies are shown in Table 3.2.

In Table 3.2, we observe that setting the batch sampling strategy to be DS yields relative improvements of 13.64% and 4.79% on DAIC-WOZ and CONVERGE1, respectively, compared with using RS. Unlike RS, DS might preserve speaker information because embeddings of two speakers' instances are optimized to be spread-out, this observation implies speaker-discriminative information might be crucial in determining depression status. An ablation study is conducted in Section 3.4.3 to prove that speaker information is preserved using DS.

Table 3.2: F1-avg scores of Baseline and unsupervised pre-training methods on DAIC-WOZ and CONVERGE1. SS stands for sampling strategy. Best Results are boldfaced.

F1-avg	SS	DAIC-WOZ	CONVERGE1
Baseline	-	0.6083	0.7228
CPC [OLV18]	-	0.6167	0.7371
Speech SimCLR [JLC20]	-	0.6258	0.7288
IDL - Feature Level (TM)	RS	0.5683	0.7073
	DS	0.6458	0.7412
	PIS	**0.6834**	**0.7435**

In the IDL-PIS experiments, TM is also chosen for augmentation. PIS can further improve depression classification performance compared with DS, with relative improvements of 5.82% on DAIC-WOZ and 0.31% on CONVERGE. Improvements demonstrate that pseudo-labels generated by the clustering model provide a high correlation with depression status.

As a comparison with other UPT methods, experiments with CPC [OLV18] and Speech SimCLR [JLC20] without reconstruction loss are conducted and reported in Table 3.2. Results show that CPC and Speech SimCLR can perform better than the baseline system but not as well as the proposed IDL method.

3.4.3 Ablation Study on Speaker Classification

We have shown that sampling instances in a batch using the DS sampling strategy during pre-training can help with the downstream depression classification task. To prove that speaker

Table 3.3: F1-avg scores and Speaker Classification Accuracy (Spk Cls Acc) of Baseline and unsupervised pre-training on DAIC-WOZ. w/o ft stands for no fine-tuning. SS stands for sampling strategy.

Exp	SS	F1-avg	Spk Cls Acc(%)
Baseline	NA	0.6083	21.98
IDL - Feature Level (TM)	DS (w/o ft)	0.3472	68.26
	DS	0.6458	72.16
	RS	0.5683	59.45
	PIS	0.6834	71.60

information is preserved in embeddings optimized by pre-training, speaker classification tasks with embeddings generated from downstream models as input are conducted on the DAIC-WOZ test set using a simple Support Vector Machine (SVM) classifier [CV95]. 30% of the segments constitute the speaker classification test set.

Table 3.3 shows that, for IDL pre-training using TM with DS on Librispeech, without fine-tuning on the DAIC-WOZ, a 68.26% speaker classification accuracy is achieved. The corresponding F1-score of 0.3472 is reasonable since the model hasn't been tuned for the depression task. The accuracy increases to 72.16% after fine-tuning. This improvement can be explained by in-domain dataset adaptation through the downstream task. As a comparison, IDL-RS only achieves 59.45% speaker classification accuracy. The relative speaker classification accuracy degradation of 17.61% from DS to RS proves that speaker information are better preserved using DS. Additionally, using PIS achieves a comparable speaker classification accuracy and an improved F1-score on depression classification compared with DS. This observation reveals that depression-specific characteristics can be preserved in embeddings trained using PIS, along with some speaker information.

3.5 Summary

In this chapter, to deal with the data scarcity challenge, a modified UPT approach, IDL, is proposed to learn augment-invariant and instance-spread-out embeddings for depression detection tasks on DAIC-WOZ and CONVERGE1 datasets. IDL uses the framework proposed for image classification [YZY19], except for preprocessing speech signals, augmentation and sampling strategies. Different augmentation techniques are compared in terms of augment-invariant characteristics. Results show that TM yields the best performance among all augmentation methods. For learning instance-spread-out embeddings, different sampling strategies, DS and RS are investigated and compared. Results show that preserving speaker information in the embedding using DS might help with depression classification. We also propose a new sampling strategy, PIS, to generate pseudo labels based on clustering, to reveal a deeper correlation between embeddings and depression status. Compared with the baseline without pre-training, the proposed approach, PIS, achieves significant improvements in the detection of depression on DAIC-WOZ and CONVERGE1 datasets.

CHAPTER 4

Privacy-preserving Depression Detection

In this chapter, we present a novel approach to disentangle speaker information from depression characteristics through adversarial training, to address privacy concerns in depression detection. This study was first introduced in [WRA23].

4.1 Motivation

In Chapter 3, we have shown that preserving some speaker-identity characteristics can help with depression detection. This observation is consistent with previous approaches that achieve better detection performance using speaker information [DWL21, Ega22, Dum22]. As a hypothesis, we suspect that the depression space and speaker space have some overlap. Thus, capturing distinguishable speaker characteristics can help with detecting depression status. Subsequently, one important question that needs to be answered is, can we attenuate those components of speaker characteristics that are irrelevant to depression status to preserve patients' privacy?

Although the field of privacy-preserving depression detection is relatively new, few previous studies have attempted to endeavour in this direction. Among them, Federated learning [Suh22] and Sine-wave speech [Dum21] are notable examples of such methods. Although these methods are promising, their application to low-resource depression detection from speech signals is still in its early stages, and results in significant performance loss [Suh22].

More recently, [Li20, Gat22, Yin20] proposed to remove speaker-related information from speech signals using adversarial learning for speech emotion recognition. We refer to this approach as uniform speaker disentanglement (USD) where the whole model is trained with the same adversarial loss. Despite the promising results of USD in detecting depression, as reported in [Rav22], the model has certain limitations that can impede its performance. One such limitation is the lack of consideration for the interactions between different layers of the model, and the relationship between the tasks being performed and the intermediate representations. For example, recent research has shown that different layers of a model capture information differently [Che22]. It is, therefore, possible that some layers capture more depression information and less speaker information or vice versa, and applying speaker disentanglement to all the layers uniformly may result in sub-optimal performance.

In this chapter, we hypothesize that speaker-related information encoded by different layers of a model is idiosyncratic, both in terms of quantity and quality, where some layers may encode more or fewer speaker characteristics than other layers, some of which may not be relevant for depression detection. Assigning a higher penalty to such layers during adversarial training can improve overall model performance. Hence, we propose a novel non-uniform speaker disentanglement method (NUSD) that regulates the proportion of speaker disentanglement applied to different model layers and shows that NUSD outperforms USD.

4.2 Background: Uniform Speaker Disentanglement

As the research introduced in this chapter is extended from the previously proposed USD study [Rav22], we briefly introduce the USD methodology.

USD [Rav22] minimizes the prediction loss for the primary task and maximizes the loss of the auxiliary task. In the context of depression detection with speaker disentanglement, the primary task is depression detection, and the auxiliary task is speaker identification (SID). The final objective is the sum of depression detection loss and speaker identification loss scaled by a pre-defined negative factor. By training the model in this adversarial manner, we are forcing the model to make correct predictions about depression status without identifying the speaker.

4.3 Non-uniform Speaker Disentanglement

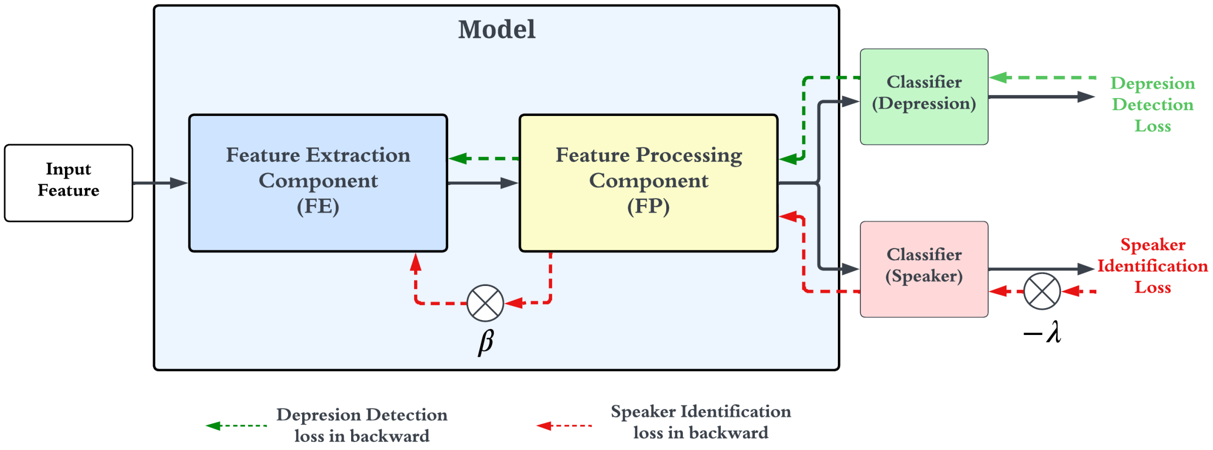

Model
Input Feature
Feature Extraction Component (FE)
Feature Processing Component (FP)
Classifier (Depression)
Classifier (Speaker)
Depresion Detection Loss
Speaker Identification Loss
β
$-\lambda$
Depresion Detection loss in backward
Speaker Identification loss in backward

Figure 4.1: Block diagram representing non-uniform speaker disentanglement of speaker and depression characteristics.

To address the limitation of not considering differential layer behaviours of the network in the USD approach, we propose a non-uniform speaker disentanglement (NUSD) technique. The NUSD framework is as shown in Figure 4.1. Similar to USD, we have the primary depression detection task and the auxiliary SID task. The final objective function is defined as

$$L_{NUSD} = L_{Dep} - \lambda(L_{SID}) \tag{4.1}$$

where L_{Dep} is the depression-detection loss and λ controls how much of the SID loss, L_{SID} contributes to the total loss, L_{NUSD}. Conventionally, L_{Dep} is Binary Cross Entropy loss, and L_{SID} is multi-class Cross-Entropy loss. A higher value of λ indicates a greater adversarial cost during training.

To accommodate the differential layer behaviour of the model, the loss gradients of the auxiliary task can be split into multiple components based on model layers and unlike USD, loss maximization can be applied differently to each component thereby allowing for varying levels of disentanglement to be applied to different layers. In this study, we split the gradients into two components: the feature extraction component (FE) composed of the initial layers and the feature processing component (FP) made up of the final layers as detailed in Section 4.4.3.1. This decision is made based on the observation from [Che22] that, initial layers have greater activation weight in speaker-related tasks. The layer separation policy used in this study is based on our preliminary observation of depression and speaker separability across layers, such that the layer with significant separability change is selected to separate FE and FP components.

Let the trainable parameters of a model be denoted as θ_{ALL}, the trainable parameters of these components can be represented as θ_{FE} and θ_{FP} for the FE and the FP layers, respectively. When the SID loss is backpropagated to FE layers, it is further multiplied by a controllable scalar β. This can be written as -

$$\frac{\partial L_{SID}}{\partial \theta_{ALL}} = [\frac{\partial (\beta L_{SID})}{\partial \theta_{FE}}, \frac{\partial (L_{SID})}{\partial \theta_{FP}}] \tag{4.2}$$

By changing the factor β, NUSD can regulate adversarial disentanglement of different layers of the model differently. For example, if $\beta < 1$, then the FP layers are penalized more than the FE layers during adversarial training and vice-versa. Conversely, if $\beta = 1$, then NUSD is equivalent to USD.

During the optimizer's update step, the model's parameters are updated as follows:

$$\begin{aligned} \theta_{FE} &= \theta_{FE} + \alpha(\lambda\beta\frac{\partial L_{SID}}{\partial \theta_{FE}} - \frac{\partial L_{Dep}}{\partial \theta_{FE}}) \\ \theta_{FP} &= \theta_{FP} + \alpha(\lambda\frac{\partial L_{SID}}{\partial \theta_{FP}} - \frac{\partial L_{Dep}}{\partial \theta_{FP}}) \end{aligned} \tag{4.3}$$

where α is the learning rate. The negative term (positive sign) for the speaker gradient in Eq. 4.3 ensures that the model maximizes L_{SID} while simultaneously optimizing L_{Dep} thereby partially disentangling speaker identity and depression status.

Table 4.1: Architecture details of the ECAPA-TDNN Model. [InC,OutC,K,S,P,D] are in-channels, out-channels, kernel, stride, padding and dilation, respectively.

Layer Name	**InC,OutC,K,S,P,D**
Input Layer	[1,128,1024,512,0,1]
SE-Res2-1	[128,128,3,1,2,2]
SE-Res2-2	[128,128,3,1,3,3]
SE-Res2-3	[128,128,3,1,4,4]
Feature aggregation	-
Concat-Conv	[384,384,1,1,0,1]
AttentiveStatsPool	[384,768,-,-,-,-]
Embedding Layer	[768,128,-,-,-,-]
Speaker Prediction Layer	[128,107,-,-,-,-]
Depression Prediction Layer	[128,1,-,-,-,-]

4.4 Experimental Setup

4.4.1 Database and Pre-processing

We use DAIC-WOZ to evalute the effectiveness of our techniques. The models were trained using raw-audio features as input, with similar pre-processing steps as those used to address data imbalance in [MY16, Bai21, Rav22, Wan22b]. The training data was pre-processed with random cropping and sampling, where each utterance was randomly cropped to the length of the shortest utterance and segmented into multiple 3.84s segments (equivalent to 61440 raw-audio samples). The experimental configurations are identical to those mentioned in Chapter 3.

4.4.2 Models

4.4.2.1 ECAPA-TDNN

The Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) [Des20] architecture extends temporal attention-based pooling to channel dimension, such that each self-attention score represents the importance of a frame given the channel. Additionally, to build a global channel inter-dependencies, 1-dimensional Squeeze-Excitation (SE) block is introduced. Each channel representation is multiplied by a weighting factor obtained from passing the statistical mean over time into a sub-neural-network module. The SE block is concatenated after several (dilated) convolutional layers with input residual to build a SE-Res2Block. In addition, multi-layer information exploitation is achieved by passing all preceding outputs from the SE-Res2Blocks and the initial convolutional module into the next block through residual connections.

In contrast to previous studies [Wan22a] that use spectrograms or MFCCs as inputs, the ECAPA-TDNN model is trained using raw-audio signals. To accommodate raw-audio speech signals as inputs and avoid overfitting the model to a small training dataset, the ECAPA-TDNN model architecture was modified (see Table 4.1). Specifically, the kernel and stride of the input convolution layer, the number of channels in the intermediate layers and the dimensions of the prediction layers were modified.

4.4.2.2 DepAudioNet

This model employed a CNN-LSTM architecture as proposed in [MY16] with implementation based on [Bai21]. To accommodate the raw-audio input feature type, two 1-D Convolution layers followed by two unidirectional LSTM layers were used. Lastly, the depression and speaker prediction layers were fully connected layers with 107 output dimensions for speaker labels.

4.4.3 Experiments

4.4.3.1 USD and NUSD

In the USD experiments, both models share the same adversarial weight λ across all layers. In the NUSD experiments, the FE layers are weighted with $\beta\lambda$ and the FP layers with λ. We consider the input layer and three SE-Res2blocks of the ECAPA-TDNN model as FE. The feature aggregation layer, Concat-Conv layer, attention layer, the fully connected embedding layer and the prediction layers are considered as the FP layers. Similarly, for DepAudioNet, the first 2 convolutional layers are FE with the two LSTM layers along with the prediction layers as FP. β and λ values are empirically chosen.

4.4.3.2 Speaker Identification Experiments

To investigate how speaker disentanglement affects speaker-identity, we conduct an SID experiment. This involves training a support vector classifier (SVC) using either the embedding layer output from the ECAPA-TDNN model or the hidden representation of the last LSTM layer from the DepAudioNet model. During the experiment, the SVC training embeddings are obtained from the baseline model without speaker disentanglement, while the evaluation embeddings are taken from the model with or without speaker disentanglement.

4.4.3.3 Layer-wise Generalized Discrimination Value Analysis

Because the proposed method regulates the magnitude of adversarial disentanglement applied to different components of the models, we investigate layer-wise behaviour of the models with and without NUSD. This is accomplished with Generalized Discrimination Value (GDV) [Sch21] analysis. Previously, GDV has been proposed as a metric to evaluate the separability of specific representations with respect to various class and data labels. In this paper, we employ GDV to measure speaker and depression-separability of individual layer

Table 4.2: Depression detection performance for DepAudioNet (D1-D3) and ECAPA-TDNN (E1-E3) based on F1-avg, F1-ND, F1-D, and SID accuracy using the DAIC-WOZ dataset. The symbol ' ↑' and ' ↓' indicate a higher or lower value is better, respectively. The best results are highlighted in bold.

Model Architecture	**Input Feature**	**Disentanglement Method**	**Model Parameters**	**F1-avg** ↑	**F1-ND** ↑	**F1-D** ↑	**SID Accuracy** ↓
DepAudioNet [Bai21] ($D1$)	Raw-Audio	None	445k	0.6259	0.7755	0.4762	10.04%
DepAudioNet [Rav22] ($D2$)	Raw-Audio	USD	459k	0.6830	0.7826	0.5833	8.91%
DepAudioNet ($D3$)	Raw-Audio	NUSD	459k	0.7086	0.8085	0.6087	8.05%
ECAPA-TDNN ($E1$)	Raw-Audio	None	595k	0.6329	0.7273	0.5385	42.33%
ECAPA-TDNN ($E2$)	Raw-Audio	USD	609k	0.7086	0.8085	0.6087	9.38%
ECAPA-TDNN ($E3$)	Raw-Audio	NUSD	609k	**0.7349**	**0.8333**	**0.6364**	**4.68%**
Δ ($E3$ vs $E2$) in %	-	-	-	3.70	2.80	4.55	-50.11

outputs for the models in consideration. The prediction layers are excluded in this analysis and GDV values are sign-flippped, such that a higher value stands for a better separability.

4.5 Results and Discussion

Experimental results are shown in Table 4.2, which is divided into two main parts (D for DepAudioNet, and E for ECAPA-TDNN) depending on the backend model in use. Methods are compared using speaker-level F1-scores for the depressed (F1-D), the non-depressed (F1-ND) classes, and their unweighted (macro) average (F1-avg).

4.5.1 Speaker Disentanglement

The DepAudioNet model ($D1$), trained on raw-audio, achieves an F1-avg score of 0.6259, whereas the proposed ECAPA-TDNN model ($E1$), also trained on raw audio signals and without speaker disentanglement, achieves an F1-avg score of 0.6329, demonstrating a 1.12% improvement.

When USD is applied to the DepAudioNet Model ($D2$), its performance improves by 9.12% from 0.6259 to 0.6830 (λ = 3e$-$4). Furthermore, when the ECAPA-TDNN model is trained using USD ($E2$), it achieves an impressive F1-avg score of 0.7086 (λ = 3e$-$3), outperforming $E1$ by 11.96%. Along with a significant increase in depression detection performance, there is a decrease in SID accuracy of 11.2% (from 10.04% to 8.91%) and 77.8% (from 42.33% to 9.38%) for $D2$ and $E2$, respectively.

Next, we apply NUSD to the DepAudioNet and ECAPA-TDNN models, and label the resulting best-performing models as $D3$ and $E3$, respectively. Model $D3$ achieves an F1-avg of 0.7086 ($\beta = 5$, λ = 4e$-$4), an increase of 13.21% over $D1$ and 3.75% over $D2$, while only marginally reducing the speaker classification accuracy to 8.05%. The overall best-performing model is $E3$, which achieves an F1-avg of 0.7349 ($\beta = 5$, λ = 8e$-$6), outperforming the corresponding baseline models $E1$ and $E2$ by 16.12% and 3.7%, respectively while simultaneously reducing the speaker classification accuracy to 4.68%. These results imply that applying NUSD can be an effective way to enhance depression classification performance while reducing SID performance. All performance improvement due to disentanglement are statistically significant.

4.5.2 The Effect of β

In order to study the impact of the hyper-parameter β on model performance, we conducted a series of experiments using different values of β ranging from 10 to 0.1. The F1-avg scores were plotted as a function of β for both models (Figure 4.2). Our analysis revealed two key observations: Firstly, for both models, NUSD ($\beta = 5$) consistently outperformed USD ($\beta = 1$), indicating that a non-uniform manner of adversarial training can be beneficial for performance.

Secondly, we observed a trend in both DepAudioNet and ECAPA-TDNN models wherein higher values of β produced better results up to $\beta = 5$. This finding suggests that assigning a higher penalty to the initial layers than to the final layers during adversarial training can

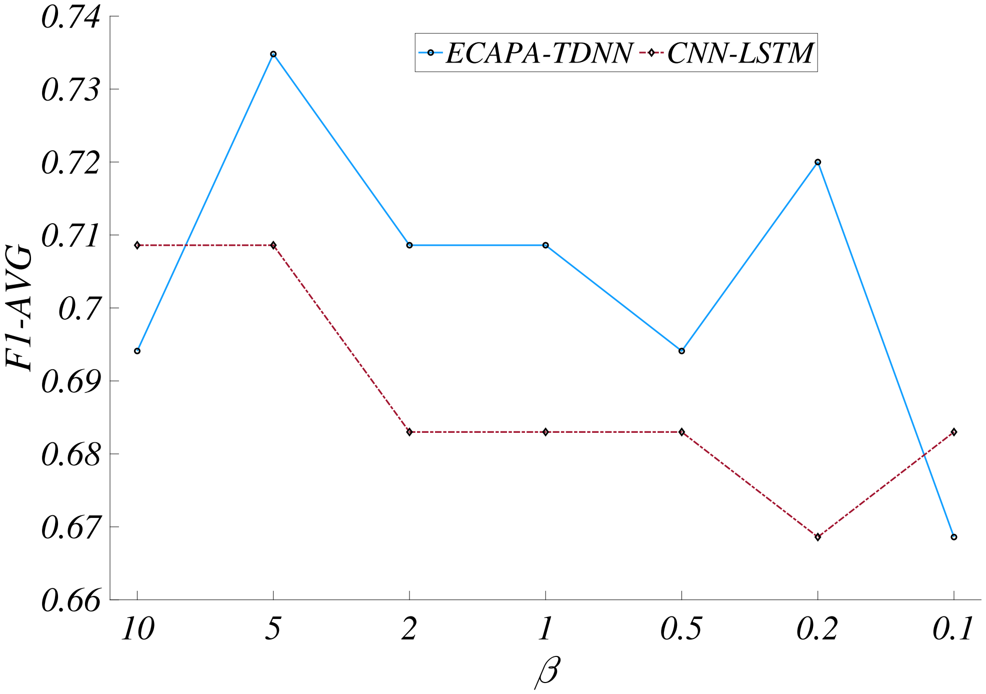


Figure 4.2: A plot of F1-avg versus NUSD β values for the ECAPA-TDNN and the DepAudioNet CNN-LSTM model. Best viewed in color.

improve model performance. One possible explanation is that assigning a higher weight to penalize FE layers in NUSD leads to a more effective suppression of speaker-specific feature extraction that may not be too relevant to the primary task of depression detection, compared to assigning the same weight to both FE and FP layers as in USD. Although this is a consequential outcome and holds true for depression detection using the DAIC-WOZ dataset, further investigation is required to verify that the framework generalizes to other domains.

4.5.3 Layer-wise GDV Analysis

Depression and speaker separability of individual layers of the $E1$, $E2$ and $E3$ were analyzed using the GDV scores (Figure 4.3). Overall, these plots offer valuable insights into the behavior of the model and shed light on how the proposed method affects the separation of depression and speaker features within the model. The main outcomes are as follows: firstly, for speaker-separability, NUSD had the lowest GDV scores among the three methods in all layers except in the embedding layer (0.664 for USD vs. 0.688 for NUSD) showing that NUSD was better at speaker disentanglement than USD in the FE layers and comparable to USD in the FP layers.

Secondly, for the depression separability, we observe that NUSD has a significantly better separability profile throughout the model than the USD and the baseline counterparts. These findings support our hypothesis that speaker information encoded by different layers of the model is distinctive and non-uniform speaker disentanglement, which exploits this characteristic of model-behavior, leads to better depression detection.

4.6 Summary

In this chapter, we introduce a novel framework, NUSD, to address privacy concerns in depression detection through adversarial training. Compared to the USD method [RWF22b],

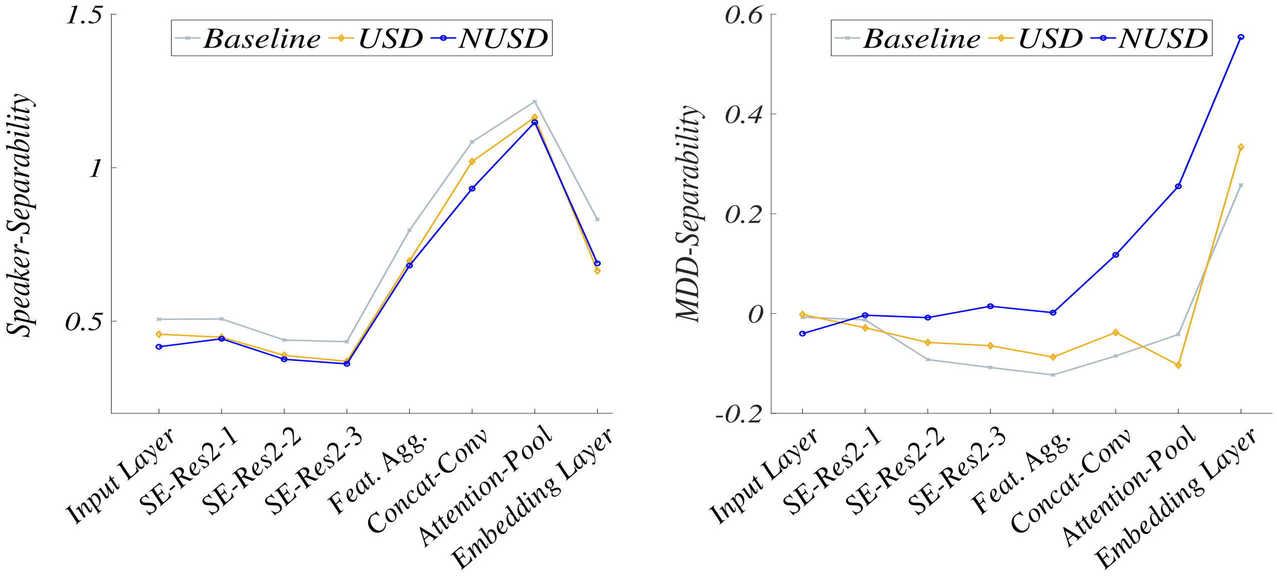


Figure 4.3: Plots of layer-wise speaker (left) and depression (right) separability GDV scores of the ECAPA-TDNN model. Three models are analyzed - baseline without speaker disentanglement ($E1$), USD ($E2$), and NUSD ($E3$). X-axis represents the layers of the ECAPA-TDNN model. Best viewed in color.

the proposed NUSD considers differential information encoding behavior across different layers of the model, and adjusts the weighting of the adversarial loss between the two portions of the model: the FE and the FP components. The proposed NUSD approach achieves better performance on the DAIC-WOZ dataset compared to the baseline system without disentanglement and the USD method while simultaneously lowering SID accuracy. We analyze the behavior of the model layers using a class separability framework, finding that a higher adversarial weight to the FE layers more effectively suppresses speaker information than USD, leading to a better encoding of depression information and performance improvement. These findings suggest that our approach leads to better model performance with improved speaker disentanglement. More importantly, we prove that speech-based depression detection can be done without over-reliance on speaker-identity features.

CHAPTER 5

Improving Depression Detection through Speech Temporal Modeling

In this chapter, we address non-uniformly distributed depression patterns within speech signals through the proposed Speechformer-CTC framework. This study was first introduced in [WRF24].

5.1 Motivation

Several studies have attempted to leverage the non-uniformly distributed speech patterns in depression detection tasks. In [WLZ21], the authors use a multi-channel convolutional layer to generate a 3D feature map with different temporal spans at the frame level. An attention module is incorporated to enable the model to automatically determine valid and invalid frames. In [LNZ22], a long-term global information embedding (GIE) is proposed to re-weight each frame output from the LSTM module, allowing frames with more significant depression cues to be emphasized through attention function. Additionally, in [NLT21], the authors introduce the Time-Frequency Attention (TFA) and merge it with the Squeeze-and-Excitation component to emphasize timestamps, frequency bands, and frequency channels related to depression.

Non-uniform temporal modeling of depression characteristics is still in its early stages. To the best of our knowledge, all previous studies addressing this variability in depression patterns rely on the attention mechanism in an "unsupervised" manner. However, a potentially more effective method involves assigning pseudo-label sequences to segments, allowing sequential modeling of depression detection tasks in a ”supervised” manner. This approach could enhance the identification of temporal regions with highly correlated verbal depression cues.

The approach for pseudo-label generation for sequential modeling of speech has been investigated in Speech Emotion Recognition (SER), another domain that can benefit from modeling the non-uniformity in a given speech utterance [WLL24, LLZ23, LB23]. Prior studies include [LT15, CHR18, HRC18, CP17], where a CTC objective [1] is used to reformulate the SER task into a sequence-to-sequence task.

Building upon the achievements in SER, notably in pseudo-label generalization and optimizing models via CTC-based methods, we present a novel framework, Speechformer-CTC (Connectionist Temporal Classification), for depression detection. This framework aims to capture the non-uniformly distributed patterns of depression using generated CTC-labels through sequential modeling methodology. Two novel CTC-label generation policies are proposed, namely the One-Hot policy, and the HuBERT policy, and the effectiveness of our proposed method is evaluated at various granularities, including the frame level, phoneme level, word level, and utterance level.

5.2 Background

5.2.1 Speechformer

Given the variability in the temporal characteristics of depression in speech signals [KBB23, ZBZ19, MSH20, ZGH22], it is necessary to apply the proposed method at various stages and examine the significance of each stage separately. Therefore, Speechformer is selected as the foundation model for our study.

In [CXX22], two key ideas were introduced based on the Transformer model to facilitate

[1] Please refer to 5.2.2 for CTC details.

the modeling of the speech signal structure and improve computational efficiency:

1) Hierarchical Merge Operation: The study assumes that the speech signal is structured as *frame* → *phoneme* → *word* → *utterance*, progressing gradually from local to global scales in the temporal domain [CXX22]. Consecutive stages are interconnected through merging blocks to aggregate finer-grained representations into coarser-grained representations. It allows the model to capture task-related information across multiple granularities.

2) Speechformer block (SF-block) with Speech-based Multi-head Self-Attention (Speech-MSA): Speech-MSA differs from the conventional Transformer by constraining the attention computation within small-scope windows that contain only several adjacent time steps. This attention scale constraint significantly reduces computational complexity. The local span at each stage is manually selected through statistical analysis of speech signals.

5.2.2 Connectionist Temporal Classification

CTC was originally proposed as a method for labeling unsegmented sequence data for sequence-to-sequence tasks [GFG06]. Its capability to automatically align different lengths of input and output sequences makes it a standard approach for Automatic Speech Recognition (ASR) tasks [FCC21, FCC23, FWG23]. Here, we briefly provide a mathematical derivation of CTC.

Denote an input sequence $\mathbf{X} = [x_1, ..., x_t, ...x_T]$ of length T, and the corresponding target sequence $\mathbf{Y} = [y_1, ..., y_m, ...y_M]$ of length M. Let $\mathbf{Z} = [z_1, ..., z_t, ...z_T]$ be the output of the model where z_t is mapped from the input x_t at time step t. Denote the original label set as L (which is vocabulary/alphabet in ASR), CTC introduces the blank token $Null$ for loss computation and extends the label set L to $L' = \{L, Null\}$, i.e. $z_t \in L'$. During loss calculation, $Null$ and repeated tokens are removed through a collapse function β. Therefore, correctly predicted Zs are defined as those where $Z = \beta^{-1}(Y)$. Since the collapse function β is a many-to-one mapping, multiple Zs can be mapped to the groundtruth target sequence Y.

As a result, CTC loss is defined as the summation of all valid Zs negative log-probabilities, which is formulated as:

$$L_{CTC} = -log \sum_{Z \in \beta^{-1}(Y)} \prod_{t=1}^{T} P(z_t|X) \tag{5.1}$$

5.3 Method

The proposed method is outlined in four stages. First, we describe the architecture of Speechformer-CTC, followed by a preliminary experiment that shows the limitation of the Naive-One-Hot policy. Next, we explain two CTC-label generation policies: the Expectation-based One-Hot policy (E-One-Hot), and the HuBERT policy. Lastly, we discuss the fusion of the content features.

5.3.1 Speechformer-CTC

Before introducing the proposed Speechformer-CTC framework (shown in Figure 5.1), the backbone Speechformer model [CXX22] is described.

The Speechformer model consists of alternately concatenated SF-blocks and merging blocks, where input features are transformed from the frame level (F) to the utterance level (U) through the phoneme level (P) and word level (W). In the first module of Speechformer (SF-block-F), the frame-level input features (e.g. Log-Mel-spectrogram), denoted as X_F with a length of T_F, are transformed into latent representations $\hat{X}_F$ with the same length T_F. The SF-block-F is followed by an $F \rightarrow P$ merging block that aggregates the transformed frame-level representation $\hat{X}_F$ into a phoneme-level representation X_P with a length of T_P through adaptive average pooling followed by a linear transformation. Similar operations in the subsequent stages transform speech representations to specific granularities: the SF-block-P, SF-block-W, and SF-block-U generate $\hat{X}_P$, $\hat{X}_W$, $\hat{X}_U$, respectively. The SF-block-P is followed by a $P \rightarrow W$ merging and the SF-block-W is followed by a $W \rightarrow U$ merging, that

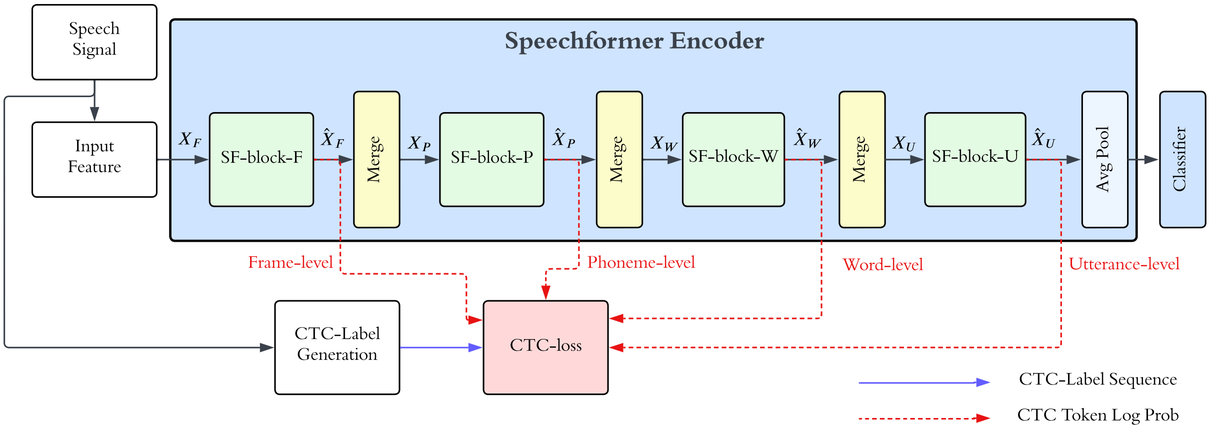


Figure 5.1: A block diagram of the proposed Speechformer-CTC model.

Table 5.1: Merge Scales of the Speechformer model. F, P, W, and U are $frame$, $phoneme$, $word$ and $utterance$. // stands for floor division.

Stage	Merge Scale (ms)	Merge Scale (step)	Description
F → P	∼50	m1 = 50 // hop1	Min length of phoneme
P → W	∼250	m2 = 250 // hop2	Min length of word
W → U	∼1000	m3 = 1000 // hop3	Max length of word
Note: hop2 = m1hop1, hop3 = m2hop2, hop1: feature extraction hop			

aggregate corresponding input sequences to obtain X_W and X_U, respectively. The merging scale for each block is the same as that proposed in the original study [CXX22], as described in Table 5.1. The last average pooling layer aggregates the output of SF-block-U, $\hat{X}_U$ (length T_U), into an embedding vector to perform utterance-level loss calculation and prediction.

The Speechformer model was designed to model speech signals by hierarchically aggregating structural speech components. However, for the speech classification task, the model optimizes a cross-entropy (CE) objective function for each utterance. This can limit the potential of this architecture because a global pooling operation from a sequence of features to a single vector might lose important local temporal characteristic information at finer granularities that may be relevant for depression identification. Therefore, the model's capability in detecting depression can benefit from considering those temporal variations explicitly.

To model the non-uniform distribution of depression characteristics across local temporal regions, we propose to incorporate a sequence-to-sequence objective in aligning different levels of representations with a self-defined, pseudo-label sequence. However, the primary challenge with such an approach is generating an appropriate pseudo-label sequence for an utterance in the sequence-to-sequence-based depression detection task. As depression datasets only provide ground-truth labels at the speaker level, generating pseudo-label sequences manually for each utterance is necessary. The pseudo-labels need to be generated based on some assumptions and prior knowledge, such that they can represent the dynamic depression states at specific stages. Even with a hypothetical intelligent pseudo-label sequence generation policy, the generated sequence is highly unlikely to be aligned with input representation sequences, which makes it difficult to train a sequence-to-sequence model in an alignment-based manner [WZF21, FAA21]. However, for such a non-aligned sequence-to-sequence task, CTC objective optimization is an ideal candidate [GFG06, FWG23]. We propose a novel framework that embeds the CTC objective function into the Speechformer model at various stages aiming to automatically align different levels of representations with generated

CTC-label sequences[2].

Before describing the proposed CTC-label generation policy, let us define the framework objective. We define the generated CTC-label sequence as Y_{ctc}, groundtruth speaker level depression label of the utterance X as y, and final singular classifier output as $\tilde{y}$. The objective function is then formulated as:

$$\begin{aligned}
L_{Dep} &= -y \cdot log(\tilde{y}) - (1-y) \cdot log(1-\tilde{y}) \\
L_{CTC} &= -log\, P(Y_{ctc}|\hat{X}_s) \\
&= -log \sum_{Z \in \beta^{-1}(Y_{ctc})} P(Z|\hat{X}_s) \\
&= -log \sum_{Z \in \beta^{-1}(Y_{ctc})} \prod_{t=1}^{T_s} P(z_t|\hat{X}_s)\ s \in \{F, P, W, U\} \\
L_{total} &= L_{Dep} + \alpha L_{CTC}
\end{aligned} \tag{5.2}$$

where L_{Dep} is utterance-level CE loss, s stands for the stage where the CTC loss function is applied, and α is the weight factor that controls the contribution of the CTC loss towards the final loss. In our experiments, all loss terms are averaged over a batch. The CTC loss is additionally averaged over target sequence Y_{ctc} to accommodate different target lengths. We apply the CTC objective on each stage separately to investigate the effect of different granularities for depression state alignment.

Compared to previous SER studies [LT15, CHR18, CP17], we keep the CE loss L_{Dep} in the final objective function for two reasons: 1) Only applying CTC loss on a specific stage will make the model ignore coarser-grained, global information that may contain relevant depression information, and 2) Compared to emotional attributes, depression-related attributes tend to be relatively longer and contain various local patterns, such as rapid emotion transition [WWY22], voiced/unvoiced regions [MSH20], or different vowels [FC23]. Therefore, preserving L_{Dep} can avoid overfitting the model to only some local speech patterns.

[2] CTC-label refers to the generated pseudo-label

At each stage, the output of the SF-block (before the merging block) is selected as the output sequence for CTC loss computation. This is done to ensure that the representations have been processed through Speech-MSA modules for in-stage feature transformation but have not undergone merging to be transformed into the representations of the next stage.

5.3.2 Preliminary Experiments using Naive-One-Hot Policy

As a preliminary experiment, we apply the previously proposed CTC objective-based sequential modeling approach, known for its effectiveness in SER tasks [LT15, HRC18, CP17], to depression detection. In this experiment, the CTC-label generation method involves the extension of the groundtruth label $y \in \{0,1\}$(0: non-depression class or ND; 1: depression class or D) into a CTC-label sequence Y_{ctc} with identical entries. For a depression sample, the Y_{ctc} generated via label-extension will be $\{1, 1, ...1\}$ with length M.

Although individuals who are depressed may have utterances with unevenly distributed depression patterns, we assume, based on SER studies [LT15, CHR18, HRC18], that speech utterances of longer duration tend to contain more regions of interest. Therefore, we set the CTC target sequence length M to be proportional to input length and denote this method of generating CTC-label sequence as the "Naive-One-Hot" policy, which is described as follows:

$$Y_{ctc} = \begin{cases} 0^M & y = 0 \\ 1^M & y = 1 \end{cases} \qquad (5.3)$$

$$\text{where } M = \frac{length(\hat{X}_s)}{k} \quad s \in \{F, P, W, U\}$$

The variable k (ratio of input to output length M) is empirically chosen to be 3 to maintain a reasonable number of valid paths during CTC optimization. Considering the optimization process of CTC, the task is to automatically align the input speech signal into M isolated D/ND regions.

The Speechformer model [CXX22] is selected as the baseline, and for the Naive-One-Hot experiment, the proposed Speechformer-CTC framework (Section 5.3.1) is used. Experiments are conducted on the DAIC-WOZ dataset [VG16] with 128-dimensional log-melspectrograms as input features, evaluated using F1-scores. Detailed configurations are described in Section 5.4.

Table 5.2: F1-scores with and without Naive-One-Hot policy on the DAIC-WOZ dataset.

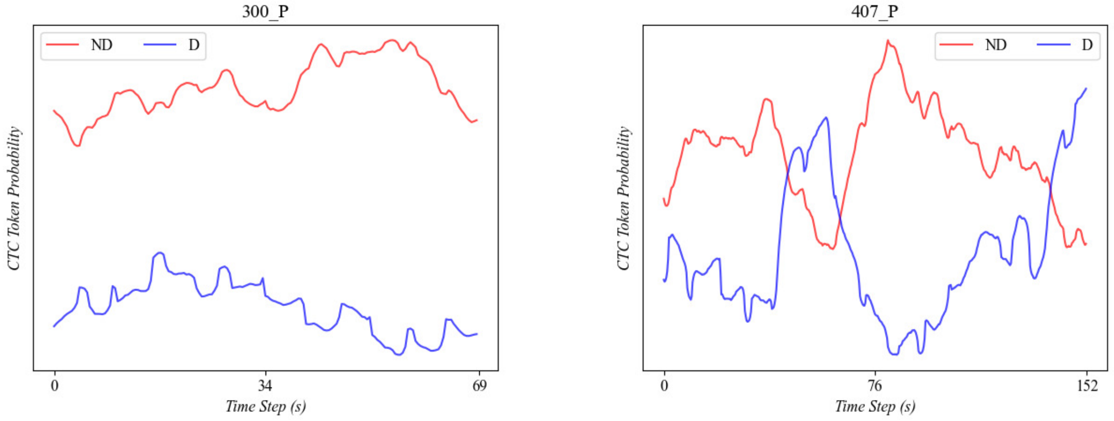
Naive-One-Hot	F1-score		
	Avg	ND	D
✗	0.7142	0.8358	0.5926
✓	0.7631	0.8387	0.6875

Table 5.2 shows that F1-scores are improved by incorporating Naive-One-Hot policy, with a 6.85% relative improvement on F1-avg over the baseline. This result verifies that explicitly modeling depression temporal variation improves detection performance.

The CTC scores for ND and D classes are plotted for four speakers in Figures 5.2 and 5.3, respectively.

As shown in Figures 5.2(a) and 5.3(a), some individuals have consistently higher correctly-predicted token scores throughout the entire session. Patients with such evident differences between D and ND scores throughout the session justify approaches that do not incorporate sequential modeling. This is because any audio clip segment from these sets of individuals exhibits significant discriminative depression characteristics. However, the persistent contrast between D/ND scores is not always observed. In Figure 5.2(b), we observe that, for a non-depressive individual, certain regions exhibit higher D class scores compared to ND class scores. Similarly, Figure 5.3(b) displays a comparable pattern, where only half of the session demonstrates higher D class scores than ND class scores for a depressive individual.

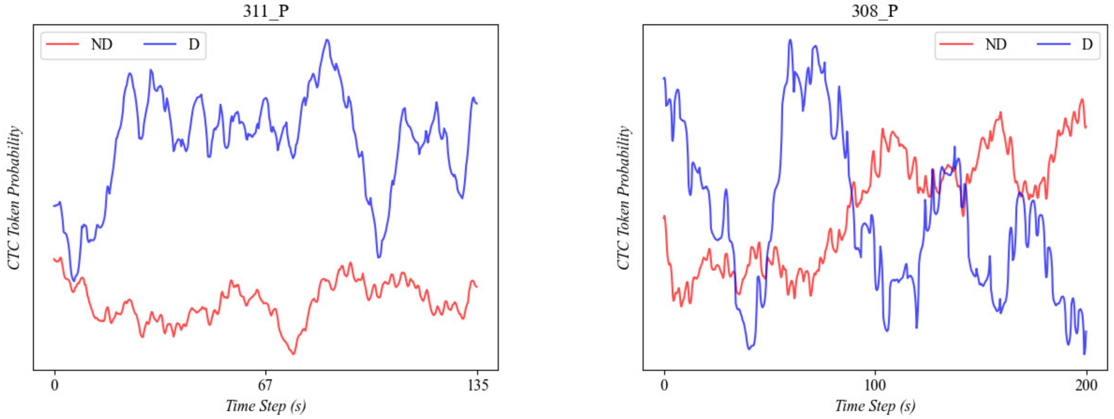
These findings suggest two things - 1) depression patterns can manifest in a non-uniform manner, and 2) the density of depression states differs among speakers, where density



(a)

(b)

Figure 5.2: Visualization of the CTC token probabilities on two ND individuals. (a) 300 (b) 407.



(a)

(b)

Figure 5.3: Visualization of the CTC token probabilities on two D individuals. (a) 311 (b) 308.

refers to the ratio of significant depression-related regions compared to the entire segment. Consequently, depression detection can benefit from sequential modeling. However, setting the CTC-label sequence length to be proportional to the input length, i.e., assuming constant depression density, could result in sub-optimal performance.

To overcome this challenge, we propose a novel label-generation policy called the E-One-Hot (Expectation-based One-Hot policy) that can tackle the varying depression state densities. Further, we utilize the HuBERT models [HBT21] to generate more descriptive label sequences using latent embedding cluster centroids. Lastly, we investigate the effects of applying non-uniform modeling at various granularities and explore the complementarity between the proposed method and content-based ASR features (fine-tuned HuBERT features for English and Whisper [RKX23] features for Mandrain).

5.3.3 Expectation-based One-Hot Policy (E-One-Hot)

As observed in the preliminary experiments (Section 5.3.2), even for depressed individuals, some cases have majority speech regions classified as non-depression. It is suspected that speech signals for some patients suffering from depression may carry significant depression-related attributes within a relatively minor but dense region across the sample, such as when being asked specific questions. To accommodate the above-mentioned variations in depression density, we propose an Expectation-based One-Hot CTC-label generation policy, denoted as the E-One-Hot label policy.

The E-One-Hot policy shares a similar label extension mechanism as the Naive-One-Hot policy does, where all entries in the generated sequence share identical values depending on the groundtruth label y. However, in contrast to earlier methods of setting M proportional to input length, a random length uniformly drawn from 1 to half of the input length is selected as the CTC-label sequence length M. The longer the CTC label length, the fewer the valid paths, making the CTC loss more aggressive. For example, when the length is chosen to be half of the input length, the model is trained to make CTC prediction as $[..., y, Null, y, ...Null, y, Null, ...]$, where groundtruth label y and $Null$ tokens occur alternately. On the contrary, when the length is selected to be 1, we assume depression characteristics are present over only one region, filling with $Null$ state before and after this region. Repeatedly and randomly choosing different values for M for every sample at every epoch of model training makes the proposed method equivalent to optimizing the expectation of the CTC loss with respect to label sequence lengths. The E-One-Hot method can be written as:

$$
\begin{aligned}
M_i &\sim uniform(1, \frac{length(\hat{X}_{s,i})}{2}) \\
Y_{ctc,i} &= \begin{cases} 0^{M_i} & y = 0 \\ 1^{M_i} & y = 1 \end{cases} \\
L_{CTC,i} &= -logP(Y_{ctc,i}|\hat{X}_{s,i}) \\
\\
L_{CTC} &\approx -\mathbb{E}_{Y_{ctc}}[logP(Y_{ctc}|\hat{X}_s)] \\
\\
&\text{where } s \in \{F, P, W, U\}
\end{aligned}
\tag{5.4}
$$

where $\mathbb{E}_{Y_{ctc}}$ stands for the expectation function with respect to Y_{ctc}, and i stands for sample index.

5.3.4 HuBERT Policy

The One-Hot label generation methods, naive and Expectation-based, use a similar label extension approach, in that all label sequences share the same entry as the groundtruth label, varying only in how the sequence lengths are determined. This 3-class (1, 0, $Null$) classification setup only guides the model to distinguish salient vs non-salient regions without leveraging more or less descriptive and representative characteristics related to depression throughout the sentences.

Inspired by the HuBERT study [HBT21], pseudo-labels generated by clustering algorithms can reveal implicit correlations between hidden representations and underlying acoustic units. For example, [LGW23] demonstrates that fine-tuning a HuBERT model can provide frame-level pseudo-emotion labels for SER, aiding in distinguishing emotional/non-emotional frames. Additionally, [WRF22] revealed that latent embeddings clustered into different groups have superior depression discriminative characteristics. These insights suggest that intermediate embeddings from HuBERT models can provide finer depression-related labels. Therefore, in this work, we propose to utilize the HuBERT model to generate the CTC-label sequences, referred to as the HuBERT policy.

The raw-audio input to a HuBERT model is transformed into a feature sequence (output of layer number 12) and it is denoted as $\mathbf{A} = [a_1, ..., a_n, ...a_{N_F}]$, where N_F is frame-level feature length. Depending on the stage at which the HuBERT label will be used, the feature $\mathbf{A}$ are merged through average pooling operations following merging scales presented in 5.1. The resulting feature sequence lengths are denoted by N_P, N_W, and N_U, respectively. Next, two separate K-means models are used to generate cluster centroid IDs, one using all D features and the other using all ND features. As the generated centroids follow a 0-index manner, to differentiate the D and ND classes, the CTC labels for the D class are shifted upward by a bias factor equal to the number of centroids. The corresponding cluster centroid IDs as $\mathbf{C} = [c_1, ..., c_n, ...c_{N_s}]$, where $s \in \{F, P, W, U\}$. We then define a function γ to remove repetitive tokens from the centroid-based sequences for each sample. The final $\gamma(\mathbf{C})$ is expected to be a descriptive label sequence representing depression patterns in latent spaces.

$$
\begin{aligned}
A &= HuBERT(Raw\,Audio) \\
K_e &\sim f(A_e, k)\text{where } e \in \{D, ND\} \\
C_i &= \begin{cases} K_{ND}(A_i) & y_i = 0 \\ K_D(A_i) & y_i = 1 \end{cases} \\
Y_{ctc,i} &= \begin{cases} \gamma(C_i) & y_i = 0 \\ \gamma(C_i) + k & y_i = 1 \end{cases}
\end{aligned}
\tag{5.5}
$$

where f is the K-means fitting function with k cluster centroids, Here, i is sample index and A_e denotes all features belonging to the e class.

An additional advantage of the HuBERT policy over the One-Hot policies is, unlike One-Hot label generation where length M is determined solely based on input sequence length, the HuBERT policy does not require a rule-based label length mapping function, but instead uses the length of $\gamma(C)$ directly. Moreover, since CTC loss can only be computed when the label length is shorter or equal to the input length, the function γ guarantees that this restriction is satisfied.

5.3.5 Content Features

Previous research has shown that text modality is effective in depression detection [HHK19, AGG18]. However, even in the textual domain, depression characteristics can be non-uniformly distributed. For example, it has been shown that some words carry higher depression-related signals than others [CGS23]. Since text modeling is not always feasible for depression corpora due to the lack of transcriptions, we propose to use features extracted from pre-trained ASR models, believed to be representative of content information.

5.4 Experiments

The effectiveness of the proposed methods is shown on DAIC-WOZ (English), CONVERGE1 and CONVERGE2 (Mandarin). Detailed experimental setups of dataset configurations (Table 5.3), acoustic features, model, and training/evaluation scheme are presented in this section.

Table 5.3: Summary of datasets. D and ND stand for depression and non-depression, respectively.

	DAIC-WOZ	CONVERGE1	CONVERGE2
Language	English	Mandarin	Mandarin
Number of Participants	189	7959	1189
D/ND	56/133	3742/4217	490/699
Gender	M/F	F	F
Sampling Rate (Hz)	16000	16000	8000
Total Duration (Hours)	50 (patient 25)	436	71
Number of Segments	32k	300k	65k
D/ND	10k/22k	219k/82k	45k/20k

5.4.1 Features and Embeddings

Log-melspectrogram (log-mel) features are selected as the acoustic features for a fair comparison with the Speechformer study [CXX22]. The window and hop sizes are set to 25ms and 10ms, respectively. Prior to feature extraction, all audio files are resampled to 16kHz, particularly for the CONVERGE2 dataset. Unless specifically mentioned, 128-dimensional log-mel features are used as input features for the Speechformer-CTC model.

Content-related features are embeddings expected to provide acoustic and text information. Due to the fact that CONVERGE datasets do not have transcriptions, in experiments where content-related features are explored, embeddings extracted from pre-trained ASR models are used as input features. For the DAIC-WOZ dataset, the HuBERT-large model [HBT21], pre-trained on 60k hours of Libri-light [KRZ20] and fine-tuned on 960 hours of Librispeech [PCP15], is used to extract the 1024-dimensional embeddings with hop size of 20ms. Extraction is performed using the fairseq toolkit [OEB19]. Regarding the CONVERGE datasets, Whisper [RKX23] is selected as the information extractor, where only the encoder is used. Being the state-of-the-art (SOTA) model for ASR tasks trained with large-scale multilingual and multi-task datasets with supervision, the Whisper model demonstrates good performance on Mandarin ASR tasks. Extracted Whisper embedding vectors have a dimension of 1024 with a hop size of 20ms [BCP16]. Throughout the experiments, pre-trained encoders are frozen without updating their weights at any stage.

5.4.2 HuBERT Label Generation

Within the scope of generating CTC-labels using the HuBERT policy, two language-matched pre-trained HuBERT models are selected to extract HuBERT features for English and Mandarin. For the DAIC-WOZ dataset, the HuBERT-large model [HBT21] pre-trained on 60k hours Libir-light is used [KRZ20]. Regarding the CONVERGE dataset, a Chinese-HuBERT-large model, pre-trained on the 10k hours WenetSpeech training set [ZLG22] is applied. In the clustering phase, MinibatchKmeans is fitted using the Scikit-learn package [PVG11], aligning with the label generation method used in HuBERT pre-training [HBT21].

5.4.3 Model Configuration

The backbone Speechformer model consists of multiple Transformer encoders in each stage, with the MSA operation replaced by Speech-MSA. The number of encoder layers is 2, 2, 4, and 4 for stages F, P, W, and U, respectively. All Speech-MSA modules utilize 8 attention heads. The local span scale of Speech-MSA at each stage is manually determined to be 50ms, 400ms, 2000ms, and the input sequence length, respectively. The feature expansion factor r is set to $= \{1, 1, 1\}$. The final classifier, responsible for depression classification, consists of 3 linear layers activated by intermediate ReLU functions and a final softmax layer. The final output size is set to 2, representing scores for ND and D. Each linear layer reduces the feature dimension by half. These model configurations are selected to be the same as the original study [CXX22]. The classifier for the sequential CTC modeling task has an identical structure, with the only difference being the final output size. For the One-Hot policies, the output size is set to 3, and for the HuBERT-policy, it is set to $2k + 1$. The additional 1 output token is reserved for the $Null$ label in the HuBERT policy scenario.

5.4.4 Training and Evaluation Scheme

As indicated in Table 5.3, both datasets suffer from imbalance in terms of the number of segments from the D and ND classes. The scarcity of the D class in the DAIC-WOZ dataset is possibly caused by the lack of willingness of depressed individuals to engage in conversation. However, for the CONVERGE datasets, the datasets are collected through clinical interviews, where individuals with depression are more inclined to seek treatment and willingly describe their situation during conversations. This imbalanced sample scales from ND and D classes may result in overfitting problem on the majority class. Consequently, a downsampling strategy is applied for each majority-class speaker in a speaker-wise manner, with downsampling rates as 2, 3, and 2 for the DAIC-WOZ ND class, CONVERGE1 D class, and CONVERGE2 D class, respectively. It should be noted that the number of speakers is kept the same before and after the downsampling operation.

Models are trained at the segment level, with the maximum input sequence length set to the 80 percentile of all sequence lengths to mitigate the impact of extremely long segments. The specific determination of the maximum sequence length is done empirically for each case based on the applied features and datasets.

Experiments are conducted with Pytorch [PGM19]. Models are trained for 40 epochs with batch sizes of 16 for DAIC-WOZ and 64 for CONVERGE. The learning rates are selected empirically. A cosine annealing learning rate scheduler is applied, gradually decreasing the learning rate to 1/100 of the initial learning rate over the entire training process. An SGD optimizer is used with a momentum of 0.9 and a weight decay factor of $1e-3$ for DAIC-WOZ and 0 for CONVERGE. Regarding the factor α which controls the CTC-loss weight, we start the training with a default α as 1 and inspect the CTC-loss scale. The value of α is then chosen to maintain the CTC-loss and CE-loss to be at the same scale.

The evaluation is conducted at the speaker level using a majority voting approach. A speaker is classified as depressed if more segments are decoded as depressive than non-depressive, and vice versa. Classification performance is evaluated using the F1-score, in terms of the D class, the ND class, as well as the macro average of both (F1-avg) to avoid overoptimistic performance biased towards the majority class. Precision and recall scores of each class are also reported.

5.5 Results

Experimental results are presented in four parts. First, the proposed methods are applied to the DAIC-WOZ dataset. Two CTC-label generation policies and the corresponding results when applying CTC at various stages are compared to show the effectiveness of the proposed methods. The results obtained from the DAIC-WOZ dataset are analyzed to gain insights into non-uniform depression patterns within speech signals. We then show results for experiments with content features. The best-performing configurations obtained on the DAIC-WOZ dataset are then used to demonstrate the generalizability of the proposed methods on the CONVERGE datasets. Finally, a comparison is made between our results and other published research on depression detection.

Table 5.4: Results, in terms of F1-score, Precision, and Recall using Naive-One-Hot policy on the DAIC-WOZ dataset. s stands for the stage where the CTC objective is applied. F, P, W, and U are $frame$, *phoneme*, *word* and *utterance*.

Model	s	F1-score			Precision		Recall	
		Avg	ND	D	ND	D	ND	D
Speechformer	-	0.7142	0.8358	0.5926	0.8235	0.6154	0.8485	0.5714
Speechformer-CTC	F	0.6948	0.8182	0.5714	0.8182	0.5714	0.8182	0.5714
	P	0.7631	0.8387	0.6875	0.8966	0.6111	0.7879	0.7857
	W	0.7631	0.8387	0.6875	0.8966	0.6111	0.7879	0.7857
	U	0.7353	0.8254	0.6452	0.8667	0.5882	0.7879	0.7143

Table 5.5: F1-score, Precision, and Recall using E-one-hot policy on the DAIC-WOZ dataset. △ F1-avg column is relative improvement compared to corresponding F1-avg from Naive-One-Hot at each stage separately. D and ND stand for depression class and non-depression class, respectively. s stands for the stage where CTC objective is applied. F, P, W, and U are $frame$, *phoneme*, *word* and *utterance*. $*$ means the F1-avg change is not statistically significant. The best F1-avg score improvement is boldfaced.

s	F1-score				Precision		Recall	
	Avg	ND	D	△ F1-avg	ND	D	ND	D
F	0.8042	0.8750	0.7333	**15.75%**	0.9032	0.6875	0.8485	0.7857
P	0.8042	0.8750	0.7333	5.39%	0.9032	0.6875	0.8485	0.7857
W	0.7756	0.8615	0.6897	1.64%	0.8750	0.6667	0.8485	0.7143
U	0.6948*	0.8182	0.5714	-5.51%	0.8182	0.5714	0.8182	0.5714

5.5.1 CTC Label Generation Policies

5.5.1.1 E-One-Hot

Results for applying the E-One-Hot CTC label generation at different stages (*Frame*, *Phoneme*, *Word*, *Utterance*) of the Speechformer-CTC model are reported and compared against the Naive-One-Hot policy in Table 5.5. Naive-One-Hot results are reported in Table 5.4. Overall, the proposed E-One-Hot method achieves the best F1-avg score of 0.8042 on the *frame* and *phoneme* level, outperforming the best Naive-One-Hot performance of 0.7631 by 5.39% and the Speechformer baseline by 12.6% (which can be found in Table 5.4).

When the performance of the two One-Hot methods at individual stages are compared, it is observed that the relative improvement in F1-avg is more significant at fine-grained stages compared to coarse-grained when the proposed expectation-based CTC-label generation policy is used. The largest improvement of 15.75% is observed when the E-One-Hot policy is applied at the *F* stage. In contrast, when Naive-One-Hot is applied at the *F* stage, F1-avg performance degrades (from 0.7142 to 0.6948). A possible explanation is that the Naive-One-Hot policy might be less accurate at the fine-grained stages (when there are a larger number of samples) due to the sub-optimal assumptions about the depression density. However, as the features pass through the model, the merging operations aggregate the non-uniform depression characteristics across temporal regions which can result in a loss of fine-grained local depression characteristics. As a consequence, at coarser stages, it is expected that the benefits of the proposed method, specifically related to varying depression density, cannot be leveraged. This hypothesis is supported by a monotonic decrease in improvement at coarser stages observed for the E-One-Hot policy where the F1-avg improvement reduces to 5.39% at the *P* stage, 1.64% at the *W* stage, and a degradation of 5.51% at the *U* stage.

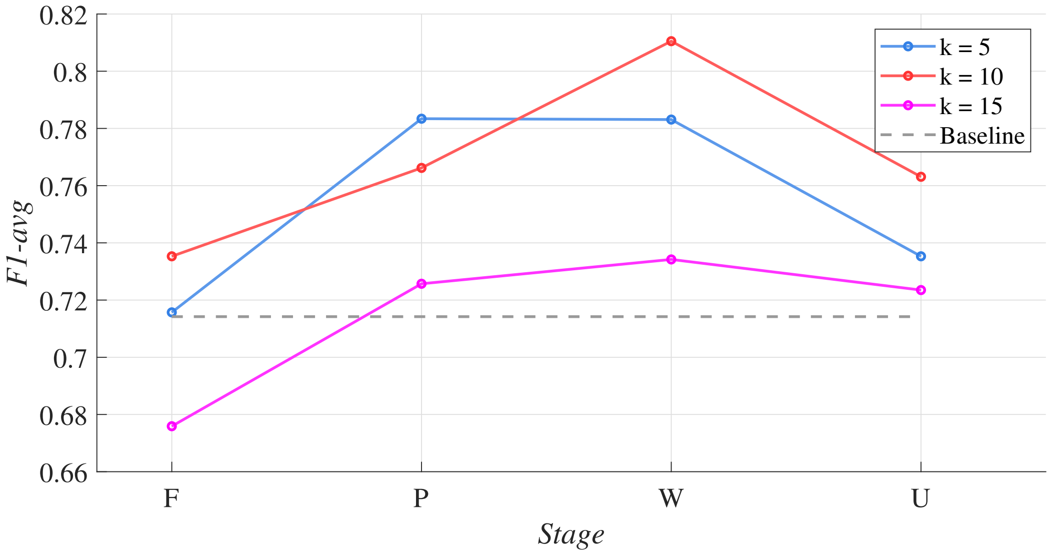


Figure 5.4: F1-avg using the HuBERT policy at different stages with different numbers of k centroids.

5.5.1.2 HuBERT Policy

Performance using CTC-label generation through HuBERT policy for various stages are shown in Figure 5.4. The performance of the Speechformer-CTC model is evaluated in different settings by changing the number of clusters k (5, 10, and 15) and the stage at which the CTC loss is applied. Overall, the best-performing model, using 10 cluster centroids on the W stage, yields a relative F1-avg improvement of 13.48% over the baseline (0.8105 vs. 0.7142, respectively).

For all experiments, applying HuBERT policy at the W stage performs the best, and the stage F performs the worst. Additionally, a consistent trend is observed where the performance improves from the F stage to the W stage and then degrades on the U stage (the only exception $k = 5$ which already achieves the best performance on the P level and the improvements saturate at later stages).

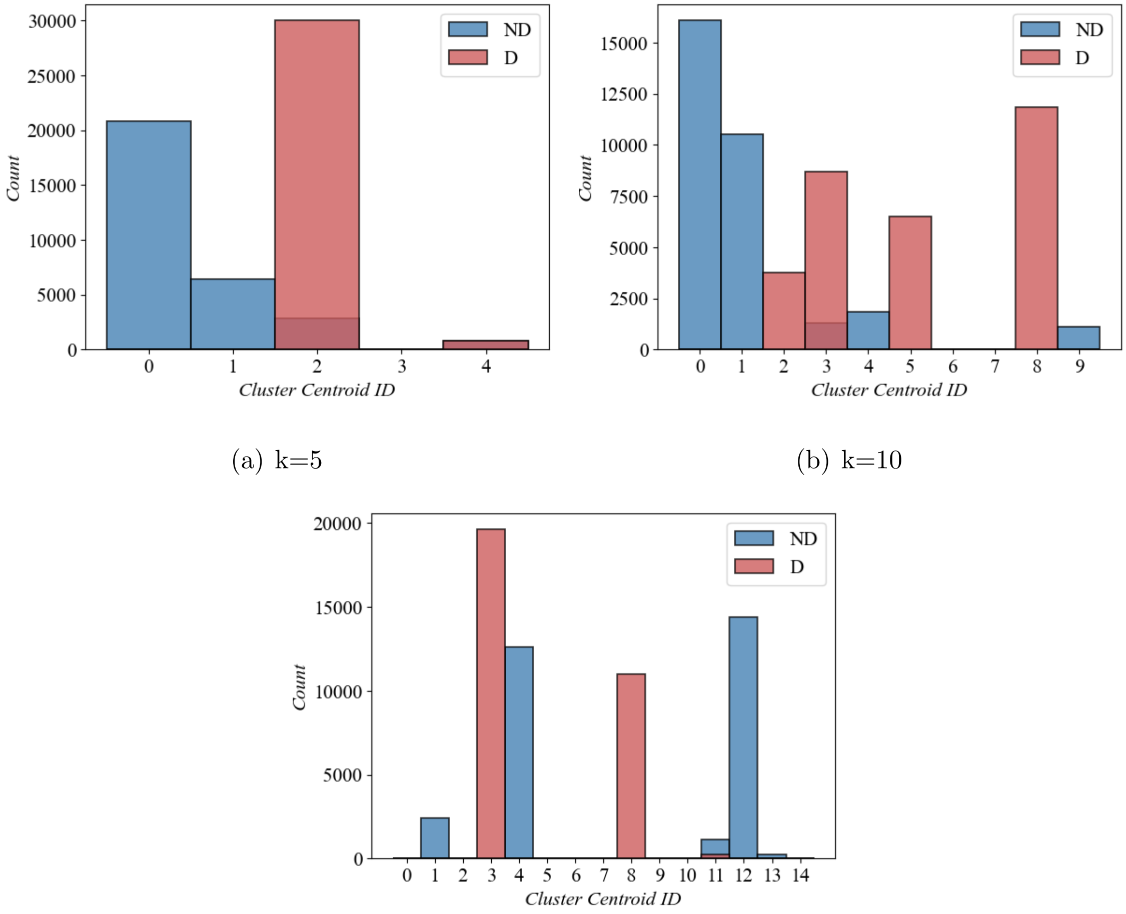
Comparing the performance when setting different K-means centroids (k), we observe the best overall performances when setting $k = 10$. In contrast, $k = 15$ results in the worst performance at all stages. The inferior performance of $k = 15$ vs $k = 5$ and 10 is further analyzed using the HuBERT label distribution plots shown in Figure 5.5.

The predicted CTC tokens obtained from the trained Speechformer-CTC models, excluding the $Null$ token, are plotted in Figure 5.5. k values are varied among 5, 10 or 15 and the CTC-loss is applied at the W stage. For each sample, the token with maximum probabilities, among ND or D HuBERT labels, is selected as the predicted token.[3]

It is observed for all values of k, some HuBERT labels are not predicted at all, for example, label 3 is not predicted for any sample when $k = 5$. This suggests that the model tends to shrink the predicted HuBERT label sets into a smaller group containing a candidate subset of original HuBERT clusters, which might be more correlated with depression. This is particularly significant given that the CTC-labels generated from the HuBERT policy are not obtained by explicitly incorporating any depression-related information across different labels. A possible explanation is that an unsupervised clustering operation on HuBERT representations cannot explicitly map depression-related characteristics to all clusters, and therefore, the Speechformer-CTC model discards some irrelevant clusters through model training.

From Figures 5.5(a) and (b), we observe that the number of possible HuBERT label predictions is approximately half of the number of clusters (for example, for $k = 10$, the predicted HuBERT label set has a size of 5 for ND class and 4 for D class). However, when $k = 15$, though each class is assigned 15 clusters, the model shrinks the prediction set size into 5 and 3 for ND and D classes, respectively. The majority of clusters is not contributing towards depression detection. Therefore, during training, those additional irrelevant clusters can result in more invalid alignment paths and degrade performance. These results therefore suggest that utilizing HuBERT labels is an effective way of introducing prior knowledge in

[3] Please note that, to make plots readable, HuBERT labels for the D class samples are shifted down by the bias factor k according to Eq. 5.5, such that all plots use the same cluster index (from 0 to $k - 1$).



(a) k=5

(b) k=10

(c) k=15

Figure 5.5: Predicted HuBERT centroids distribution with different k. The X-axis of each plot represents the Hubert centroids ID, and the Y-axis is the number of predicted tokens. (a) k = 5 (b) k = 10 (c) k = 15.

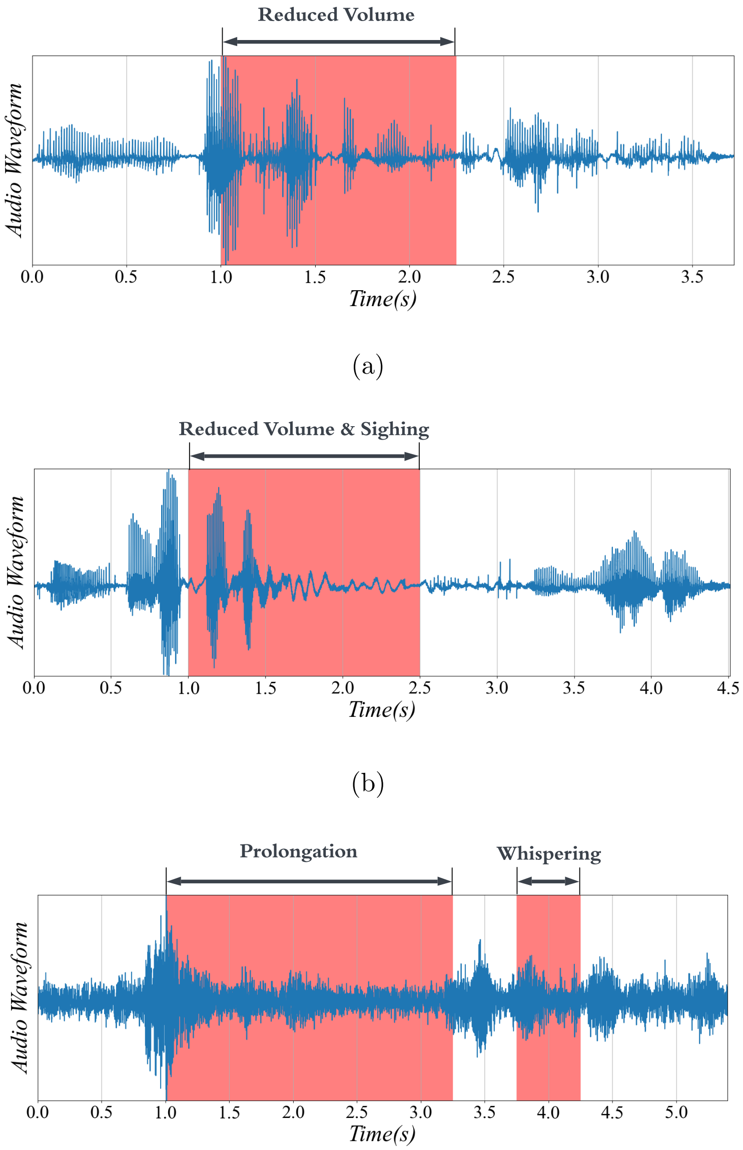
depression sequential modeling, but the number of clusters has to be carefully chosen to avoid the issue mentioned above.

Furthermore, we conduct a visualization analysis by mapping the regions with highly predicted HuBERT labels' probabilities back to the original audio clips to check whether the corresponding region has some depression-related patterns from a human perception perspective. We select a subset of mostly predicted HuBERT labels (2, 3, 5, 8) obtained from the best-performing model ($k = 10$ on the W stage) and highlight the corresponding regions as presented in Figure 5.6. The examples show that predicted HuBERT labels do correspond to some intuitive depression patterns, such as reduced volume [CSE14], sighing [VVB15], prolongation [FBC92], and whispering [JLZ20] for some cases. However, specific depression-related patterns do not correspond to HuBERT labels in a one-to-one manner, because labels are generated without cluster-wise supervision of depression patterns. Alignment between the orderless cluster centroid with specific speech activity could enhance the discrimination of different HuBERT labels and will be undertaken in future experiments.

5.5.2 Non-uniform Modeling of Content Features

In addition to conventional acoustic features, we demonstrate that non-uniform modeling of depression is also beneficial when content-related features are used. Therefore, we replace the input features from log-mel to content features (HuBERT-ft) extracted from the ASR fine-tuned HuBERT-large model for English. Results are as shown in Table 5.6.

Replacing the log-mel feature with the HuBERT-ft feature without incorporating the proposed CTC approach on the Speechformer network can yield an 8.6% improvement in terms of F1-avg. Even without non-uniform modeling, HuBERT-ft features are more representative and meaningful in capturing depression patterns compared to hand-crafted log-mel features. Second, combining HuBERT-ft features with the proposed method results in the highest performance in terms of F1-avg, F1-ND, and F1-D in this study, which are 0.8315, 0.8890, and 0.7742, respectively. The best F1-avg has a relative improvement of 16.42% compared



(a)

(b)

(c)

Figure 5.6: Visualization of audio waveforms from the DAIC-WOZ dataset, where highlighted regions represent clips with high probabilities of depression-related HuBERT centroids ID. (a) example1 (b) example2 (c) example3.

Table 5.6: F1-scores, Precision, and Recall using log-mel and HuBERT-ft features on the DAIC-WOZ dataset. The "*HuBERT CTC*" column marks whether CTC-label sequences generated by HuBERT policy are used. The best F1-avg is boldfaced.

Feature(dim)	HuBERT CTC	F1-score			Precision		Recall	
		Avg	ND	D	ND	D	ND	D
log-mel(128)	-	0.7142	0.8358	0.5926	0.8235	0.6154	0.8485	0.5714
	✓	0.8105	0.8710	0.7500	0.9310	0.6667	0.8182	0.8571
HuBERT-ft(1024)	-	0.7756	0.8615	0.6897	0.8750	0.6667	0.8485	0.7143
	✓	**0.8315**	0.8890	0.7742	0.9333	0.7059	0.8485	0.8571

to the baseline Speechformer model trained using log-mel features, and 7.21% compared to the Speechformer model trained on HuBERT-ft features without non-uniform sequential modeling.

5.5.3 Extension to the CONVERGE Datasets

We apply the proposed methods to the CONVERGE datasets to verify their generalizability to Mandarin. We use the best configurations obtained in DAIC-WOZ experiments and apply them to the CONVERGE datasets.

Overall, the performance on CONVERGE2 is lower compared to CONVERGE1. Because the system is trained with the CONVERGE1 training set and evaluated on CONVERGE2 with an additional challenge of domain mismatch. However, it still can be seen that by applying the proposed methods, F1-avg improves. With the Naive-One-Hot label generation policy, F1-avg improves by 1.77% and 2.10% on CONVERGE1 and CONVERGE2, respectively. However, applying E-One-Hot results in slight degradation. The degradation might suggest that CONVERGE, a dataset collected through clinical interviews of severely depressed patients and labeled by experts, may have a more constant depression density along the speech segment. Thus, the Naive-One-Hot generation policy gives a relatively good approximation

Table 5.7: F1-scores using the CONVERGE1 and CONVERGE2 datasets. CTC stands for CTC-label generation policy applied. * stands for the change is not statistically significant. The best F1-avg is boldfaced.

Input Features	CTC	CONVERGE1			CONVERGE2		
		F1-avg	F1-ND	F1-D	F1-avg	F1-ND	F1-D
log-mel	-	0.7463	0.7639	0.7290	0.6057	0.7139	0.4974
	Naive-One-Hot	0.7595*	0.7730	0.7460	0.6184	0.7129	0.5241
	E-One-Hot	0.7532*	0.7665	0.7400	0.6108*	0.7197	0.5026
	HuBERT	0.7651	0.7788	0.7515	0.6277	0.6776	0.5779
Whisper	HuBERT	**0.8196**	0.8435	0.7956	**0.7176**	0.8043	0.6309

without the necessity to introduce label length expectation.

By utilizing the HuBERT policy, F1-avg performances on CONVERGE1 and CONVERGE2 are improved by 2.52% and 3.63%, respectively, compared to the baseline system. However, unlike the observation of improving all F1-scores on CONVERGE1, using HuBERT policy results in a 5% F1-ND degradation on the CONVERGE2 dataset. It is suspected that generated HuBERT labels using a K-means model trained using the CONVERGE1 training set may have caused a domain mismatch problem on CONVERGE2, specifically on ND class samples. However, the significantly improved F1-D performance on the CONVERGE2 using the HuBERT policy, 16.18% compared to the baseline, shows that HuBERT labels still capture more discriminative depression-related information.

Finally, by replacing log-mel features with Whisper features, we achieve the best performances on CONVERGE1 and CONVERGE2 simultaneously, yielding relative improvements of 9.82% and 18.47%, respectively, compared to the baseline.

5.5.4 Comparison to SOTA Studies

In this section, we compare our proposed methods with existing SOTA studies for depression detection as shown in Table 5.8.

Table 5.8: Comparison with SOTA depression detection studies on the DAIC-WOZ and CONVERGE1 datasets, in terms of F1-scores. The best results are boldfaced. [MSH20] requires phoneme-level transcription, and therefore can not be applied to the CONVERGE datasets.

Dataset	Method	F1-score		
		F1-avg	F1-ND	F1-D
DAIC-WOZ	CAE ADD [SNR22]	0.7050	0.7100	0.7000
	Speechformer [CXX22]	0.7142	0.8358	0.5926
	AudVowConsNet [MSH20]	**0.8350**	**0.9000**	0.7700
	SFTN [HSS23]	0.7550	0.8400	0.6700
	Proposed Speechformer-CTC	0.8315	0.8890	**0.7742**
CONVERGE1	Fraug [RWF22a]	0.7390	-	-
	IDL [WRF22]	0.7435	0.7548	0.7323
	Speechformer [CXX22]	0.7463	0.7639	0.7290
	Proposed Speechformer-CTC	**0.8196**	**0.8435**	**0.7956**

On the DAIC-WOZ dataset, we achieve close-to-SOTA performance (Audvowconsnet [SNR22]), with a slightly improved F1-D but not F1-avg and F1-ND. Notably, Audvowconsnet relies on phoneme-level transcription alignment, to distinguish vowels and consonants, and involves pitch and noise augmentation techniques [KPP15]. In contrast, our approaches can achieve comparable performance without phoneme transcription, alignment, or augmentation, highlighting its effectiveness in depression detection. Regarding the CONVERGE datasets, because CONVERGE2 is a more recent database, we compare our results only against our previous work on CONVERGE1. The comparison shows that our method provides the best F1-scores on CONVERGE1.

5.6 Summary

In this chapter, we present a novel framework, Speechformer-CTC, to model non-uniform depression patterns within speech segments. This is achieved by introducing a CTC alignment task regularized by generated CTC-labels. Two novel CTC-label generation policies, namely the E-One-Hot and the HuBERT policies, are proposed and incorporated in objectives on various granularities. We show that: 1) Highly depressive regions can be observed in utterances of individuals without depression and vice versa. Utilizing the One-Hot policy, we show in a supervised way that depression and non-depression contours are dynamic and unevenly distributed. 2) HuBERT labels exhibit a high correlation with some depressive verbal cues within specific subsets of clustered centroids, highlighting the effectiveness of introducing prior knowledge in CTC-label generation. 3) The integration of ASR features with the proposed method further enhances the detection performance, demonstrating the compatibility of the proposed method with content-related information. Our results show that the performance of depression detection, in terms of Macro F1-score, is improved on both DAIC-WOZ (English) and CONVERGE (Mandarin) datasets. The best performances on DAIC-WOZ and CONVERGE achieve close-to-SOTA or SOTA performance but without the need for transcription.

In future work, we will apply this method to other paralinguistic speech processing tasks, including SER, and Alzheimer's disease detection. Additionally, depression-related prior knowledge, such as voiced activity detection, vowel regions, or emotional attributes, will be considered during the CTC label generation stage for better alignment.

CHAPTER 6

Conclusion

In this dissertation, we resolved three challenges in depression detection system development on speech signals: 1) mitigating data scarcity challenge with UPT, 2) disentangling speaker identity information from speech signal to preserve privacy with adversarial training, and 3) tackling non-uniform depression patterns through sequential modeling. In this chapter, main results of each topic are summarized and possible future work is discussed.

6.1 Summary

Chapter 2 described two depression corpora, namely DAIC-WOZ (English) and CONVERGE (Mandarin), along with feature sets used in this dissertation. The DAIC-WOZ dataset includes dialogues between a virtual interviewer and participants. It is the most widely used benchmark by the speech-based depression research community. The CONVERGE dataset is a large-scale database, characterized by more diverse phonetic and content variability. Within this dataset, CONVERGE1 is used for training and evaluation and CONVERGE2, a recent collected dataset, is specifically used for an out-of-domain replication study in Chapter 5. Feature sets used in this dissertation includes mel-spectrogram, raw audio and high-level features extracted from pre-trained models, namely HuBERT and Whisper.

In Chapter 3, an unsupervised pre-training framework is proposed to mitigate data scarcity by leveraging large-scale of out-of-domain data to improve speech-based depression detection system performance. We propose a novel Instance Discriminative Learning (IDL) framework to let the model learn to extract augment-invariant and instance-spread-out embeddings. We applied different augmentation techniques on the IDL pre-training stage to investigate the optimal augmentation strategies specifically for speech-based depression detection. Significant results are observed when using Time Masking (TM) and Volume Perturbation. Additionally, different sampling strategies are compared in terms of preserving speaker information or not. We showed that for the proposed IDL framework, preserving speaker information performed better than the baseline. Moreover, a novel sampling strategy, Pseudo Instance-based Sampling (PIS) was proposed to capture implicit discriminative characteristics through a clustering algorithm and showed superior performance. Comparing to the other two related pre-training techniques, Contrastive Predictive Coding and Speech SimCLR, IDL achieves better performance on two depression corpora, DAIC-WOZ and CONVERGE.

Chapter 4, a novel method, Non-uniform Speaker Disentanglement (NUSD), was proposed to more effectively disentangle speaker information by leveraging differential encoding behaviors of different components of the model. Built upon the previous USD study, we first introduced the combination of ECAPA-TDNN model with raw audio as input. Then we separated the model into Feature Extraction (FE) and Feature Processing (FP) components and scaled the auxiliary task, Speaker Identification (SID), loss towards these components differently. A Generalized Discrimination Value (GDV) analysis was conducted to gain an insight of layer-wise embedding discrimination characteristics. We showed that larger SID loss weight on FE components generally leads to greater performance on depression detection tasks. Experiments are conducted on the DAIC-WOZ dataset using two different model architectures, DepAudioNet and ECAPA-TDNN, to verify NUSD's generalizability. Results showed that NUSD can achieve better depression detection performance comparing to the system without disentanglement and with USD, while it simultaneously increases speaker disentanglement capability.

Finally, in Chapter 5, to address the non-uniformly distributed depression patterns within speech signals, we proposed a novel framework, Speechformre-CTC, to model the non-uniform depression pattern distribution through a Connectionist Temporal Classification (CTC) alignment task. Novel CTC-label generation policies were proposed, namely Expectation-One-Hot (E-One-Hot), and HuBERT policies. We investigated different speech granularities where the CTC loss is applied and found intermediate granularities (phone-level and word-level) regularization gives better performance than the finest frame-level and the most coarse utterance-level ones. Additionally, we conducted ablation experiments using content-related features extracted from pre-trained Automatic Speech Recognition (ASR) models and found the proposed method has great compatibility with those features. Finally we achieved close-to-state-of-the-art or state-of-the-art performance on DAIC-WOZ and CONVERGE datasets, respectively, without the requirement of phoneme-level transcription.

6.2 Future Work

In the UPT study, we have shown the effectiveness of speaker-based characteristics on depression detection. Since it has been shown that text modality, especially word lexical choices, can also yield good detection performance [She22, GP17, LDL19], it is worth investigating the effect of utilizing content-based characteristics discrimination at the pre-training stage. Instead of trying to distinguish different speakers, the model can be trained to distinguish different content spoken by the same speaker. We believe that this research will provide better insights into how depression is characterized/manifested in different spaces (speaker and content) and eventually leads to a robust depression detection system that can assist with clinical screening.

In the proposed NUSD study, DepAudioNet and ECAPA-TDNN models are roughly separated into FE and FP modules with prior knowledge. However, to generalize the technique to other model architectures, this separation needs to be defined empirically, such as in the Transformer-based architectures. Thus, future work will focus on exploring a more fine-grained, data-driven variant of NUSD such that the speaker-identification loss weights can be determined dynamically during model training.

In the IDL and NUSD studies, we have shown that preserving some speaker-identity information during pre-training can help with downstream depression detection tasks. Further in the NUSD study, attenuating unnecessary speaker-information bias can also improve detection performance and preserve patients' privacy simultaneously. We suspected that the depression space overlaps with speaker space. Therefore, it is worth investigating how to combine IDL and NUSD together, such that the goal is to maintain necessary speaker information for efficient depression diagnosis, but attenuate irrelevant speaker information under the consideration of privacy preservation.

Finally, for the Speechformer-CTC study, CTC-labels used for the alignment task were generated in an unsupervised way without prior knowledge. One valuable future direction would be incorporating specific speech attributes into CTC-label generation, such as Voice-Activity-Detection, emotional attributes or vowel/consonant regions. Additionally, the backbone Speechformer model conducts an even merging operation at each stage. These constant merging scales for all stages are defined through statistical analysis of speech units. A possible future work direction would be adaptively and dynamically merging acoustic units in a finer-grained manner depending on different vowels and consonants (fricatives, nasals, etc.). It also would be of interest to develop a system that addresses all three issues (scarcity, privacy, non-uniformity).cccccbngkrdiekncevvvlrcfgvuclgnngihulcbdfeec

REFERENCES

[AG18] Amber Afshan, Jinxi Guo, et al. "Effectiveness of voice quality features in detecting depression." *Interspeech*, 2018.

[AGG18] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. "Detecting Depression with Audio/Text Sequence Modeling of Interviews." In *Interspeech*, pp. 1716–1720, 2018.

[ASA15] 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.

[ASS14] Meysam Asgari, Izhak Shafran, and Lisa B Sheeber. "Inferring clinical depression from speech and spoken utterances." In *2014 IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–5. IEEE, 2014.

[Bae20] 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.

[Bai21] Andrew Bailey et al. "Gender Bias in Depression Detection Using Audio Features." In *2021 29th EUSIPCO*, pp. 596–600. IEEE, 2021.

[BCP16] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. "Openai gym." *arXiv preprint arXiv:1606.01540*, 2016.

[BNG21] SG Brederoo, FG Nadema, FG Goedhart, AE Voppel, JN De Boer, J Wouts, S Koops, and IEC Sommer. "Implementation of automatic speech analysis for early detection of psychiatric symptoms: what do patients want?" *Journal of psychiatric research*, **142**:299–301, 2021.

[BP20] 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.

[BRS04] R Michael Bagby, Andrew G Ryder, Deborah R Schuller, and Margarita B Marshall. "The Hamilton Depression Rating Scale: has the gold standard become a lead weight?" *American Journal of Psychiatry*, **161**(12):2163–2177, 2004.

[BSA19] Alexei Baevski, Steffen Schneider, and Michael Auli. "vq-wav2vec: Self-supervised learning of discrete speech representations." *arXiv preprint arXiv:1910.05453*, 2019.

[CGS23] Lisette Corbin, Emily Griner, Salman Seyedi, Zifan Jiang, Kailey Roberts, Mina Boazak, Ali Bahrami Rad, Gari D Clifford, and Robert O Cotes. "A comparison of linguistic patterns between individuals with current major depressive disorder, past major depressive disorder, and controls in a virtual, psychiatric research interview." *Journal of Affective Disorders Reports*, **14**:100645, 2023.

[Che22] S. Chen et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." *IEEE Journal of Selected Topics in Signal Processing*, **16**:1505–1518, 2022.

[CHR18] Xiaomin Chen, Wenjing Han, Huabin Ruan, Jiamu Liu, Haifeng Li, and Dongmei Jiang. "Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network." In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–6. IEEE, 2018.

[CKN20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[CP17] Vladimir Chernykh and Pavel Prikhodko. "Emotion recognition from speech with recurrent neural networks." *arXiv preprint arXiv:1701.08071*, 2017.

[CPJ11] Pamela Y Collins, Vikram Patel, Sarah S Joestl, Dana March, Thomas R Insel, Abdallah S Daar, Isabel A Bordin, E Jane Costello, Maureen Durkin, Christopher Fairburn, et al. "Grand challenges in global mental health." *Nature*, **475**(7354):27–30, 2011.

[CS15] Nicholas Cummins, Stefan Scherer, et al. "A review of depression and suicide risk assessment using speech analysis." *Speech Communication*, **71**:10–49, 2015.

[CSE14] Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, and Jarek Krajewski. "Probabilistic acoustic volume analysis for speech affected by depression." In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[CV95] Corinna Cortes and Vladimir Vapnik. "Support-vector networks." *Machine Learning*, **20**(3):273–297, 1995.

[CVS17] Nicholas Cummins, Bogdan Vlasenko, Hesam Sagha, and Björn Schuller. "Enhancing speech-based depression detection through gender dependent vowel-level formant features." In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, pp. 209–214. Springer, 2017.

[CXX22] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. "Speech-Former: A hierarchical efficient framework incorporating the characteristics of speech." *arXiv preprint arXiv:2203.03812*, 2022.

[CXX23] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. "Speechformer++: A hierarchical efficient framework for paralinguistic speech processing." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**:775–788, 2023.

[Des20] Brecht Desplanques et al. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification." In *Proc. Interspeech*, pp. 3830–3834, 2020.

[Dia94] American_Psychiatric_Association Diagnostic. "Statistical Manual of mental disorders.", 1994.

[Dub19] S Pavankumar Dubagunta et al. "Learning voice source related information for depression detection." In *ICASSP*, pp. 6525–6529, 2019.

[Dum21] Sri Harsha Dumpala et al. "Sine-Wave Speech and Privacy-Preserving Depression Detection." In *Proc. SMM21, Workshop on Speech, Music and Mind 2021*, pp. 11–15, 2021.

[Dum22] Sri Harsha Dumpala et al. "Detecting Depression with a Temporal Context Of Speaker Embeddings." *Proc. AAAI SAS*, 2022.

[DWL19] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng. "Learning discriminative features from spectrograms using center loss for speech emotion recognition." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7405–7409. IEEE, 2019.

[DWL21] Yazheng Di, Jingying Wang, Weidong Li, and Tingshao Zhu. "Using i-vectors from voice features to identify major depressive disorder." *Journal of Affective Disorders*, **288**:161–166, 2021.

[Ega22] J. V. Egas-López et al. "Automatic assessment of the degree of clinical depression from speech using X-vectors." In *ICASSP*, pp. 8502–8506. IEEE, 2022.

[FAA21] Ruchao Fan, Amber Afshan, and Abeer Alwan. "Bi-apc: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's asr." In *ICASSP*, pp. 7023–7027, 2021.

[FBC92] Alastair J Flint, Sandra E Black, Irene Campbell-Taylor, Gillian F Gailey, and Carey Levinton. "Acoustic analysis in the differentiation of Parkinson's disease and major depression." *Journal of Psycholinguistic Research*, **21**:383–399, 1992.

[FC23] Kexin Feng and Theodora Chaspari. "A knowledge-driven vowel-based approach of depression classification from speech using data augmentation." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[FCC21] Ruchao Fan, Wei Chu, Peng Chang, and Jing Xiao. "CASS-NAT: CTC alignment-based single step non-autoregressive transformer for speech recognition." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5889–5893. IEEE, 2021.

[FCC23] Ruchao Fan, Wei Chu, Peng Chang, and Abeer Alwan. "A ctc alignment-based non-autoregressive transformer for end-to-end automatic speech recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**:1436–1448, 2023.

[Fen22] K. Feng et al. "A knowledge-driven vowel-based approach of depression classification from speech using data augmentation." *arXiv preprint arXiv:2210.15261*, 2022.

[FK20] Yue Fan, JW Kang, et al. "Cn-celeb: a challenging chinese speaker recognition dataset." In *ICASSP*, pp. 7604–7608, 2020.

[FSA24] Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. "UniEnc-CASSNAT: An Encoder-only Non-autoregressive ASR for Speech SSL Models." *IEEE Signal Processing Letters*, 2024.

[FWG23] Ruchao Fan, Yiming Wang, Yashesh Gaur, and Jinyu Li. "CTCBERT: Advancing Hidden-Unit Bert with CTC Objectives." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[Gat22] Itai Gat et al. "Speaker normalization for self-supervised speech emotion recognition." In *ICASSP*, pp. 7342–7346. IEEE, 2022.

[GFG06] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

[GP17] Yuan Gong and Christian Poellabauer. "Topic modeling based multi-modal depression detection." In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 69–76, 2017.

[HBT21] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM TASLP*, **29**:3451–3460, 2021.

[HCM18] Sahar Harati, Andrea Crowell, Helen Mayberg, and Shamim Nemati. "Depression severity classification from speech emotion." In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5763–5766. IEEE, 2018.

[HEJ19] Zhaocheng Huang, Julien Epps, and Dale Joachim. "Investigation of speech landmark patterns for depression detection." *IEEE Transactions on Affective Computing*, 2019.

[HEJ20] Zhaocheng Huang, Julien Epps, and Dale Joachim. "Exploiting vocal tract coordination using dilated CNNs for depression detection in naturalistic environments." In *ICASSP*, pp. 6549–6553, 2020.

[HHK19] Jana M Havigerová, Jiří Haviger, Dalibor Kučera, and Petra Hoffmannová. "Text-based detection of the risk of depression." *Frontiers in psychology*, **10**:513, 2019.

[HRC18] Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, and Björn Schuller. "Towards temporal modelling of categorical speech emotion recognition." 2018.

[HSS22] 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.

[HSS23] Zhuojin Han, Yuanyuan Shang, Zhuhong Shao, Jingyi Liu, Guodong Guo, Tie Liu, Hui Ding, and Qiang Hu. "Spatial-Temporal Feature Network for Speech-Based Depression Recognition." *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[JH13] 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, p. 21, 2013.

[JLC20] Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning." *arXiv preprint arXiv:2010.13991*, 2020.

[JLZ20] Yuan Jia, Yuzhu Liang, and Tingshao Zhu. "An Analysis of Acoustic Features in Reading Speech from Chinese Patients with Depression." In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 128–133. IEEE, 2020.

[JZW18] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." *Advances in neural information processing systems*, **31**, 2018.

[KAM98] Ronald C Kessler, Gavin Andrews, Daniel Mroczek, Bedirhan Ustun, and Hans-Ulrich Wittchen. "The World Health Organization composite international diagnostic interview short-form (CIDI-SF)." *International journal of methods in psychiatric research*, **7**(4):171–185, 1998.

[KBB23] Sanne Koops, Sanne G Brederoo, Janna N de Boer, Femke G Nadema, Alban E Voppel, and Iris E Sommer. "Speech as a biomarker for depression." *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, **22**(2):152–160, 2023.

[KPP15] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. "Audio augmentation for speech recognition." In *Sixteenth annual conference of the international speech communication association*, 2015.

[KRZ20] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. "Libri-light: A benchmark for asr with limited or no supervision." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673. IEEE, 2020.

[KS09] 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.

[KYK21] Danish M Khan, Norashikin Yahya, Nidal Kamel, and Ibrahima Faye. "Automated diagnosis of major depressive disorder using brain effective connectivity and 3D convolutional neural network." *IEEE Access*, **9**:8835–8846, 2021.

[LB23] Wei-Cheng Lin and Carlos Busso. "Sequential Modeling by Leveraging Non-Uniform Distribution of Speech Emotion." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**:1087–1099, 2023.

[LDG14] Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo. "A study of acoustic features for the classification of depressed speech." In *MIPRO*, pp. 1331–1335. IEEE, 2014.

[LDL19] Genevieve Lam, Huang Dongyan, and Weisi Lin. "Context-aware deep learning for multi-modal depression detection." In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 3946–3950. IEEE, 2019.

[LGW23] Qifei Li, Yingming Gao, Cong Wang, Yayue Deng, Jinlong Xue, Yichen Han, and Ya Li. "Frame-level emotional state alignment method for speech emotion recognition." *arXiv preprint arXiv:2312.16383*, 2023.

[LH15] Zhenyu Liu, Bin Hu, et al. "Detection of depression in speech." In *2015 ACII*, pp. 743–747. IEEE, 2015.

[LHL16] Zhenyu Liu, Bin Hu, Fei Liu, Huanyu Kang, Xiaoyu Li, Lihua Yan, and Tianyang Wang. "Evaluation of depression severity in speech." In *Brain Informatics and Health: International Conference, BIH 2016, Omaha, NE, USA, October 13-16, 2016 Proceedings*, pp. 312–321. Springer, 2016.

[Li20] Haoqi Li et al. "Speaker-invariant affective representation learning via adversarial training." In *ICASSP*, pp. 7144–7148. IEEE, 2020.

[LJD14] Mariane Acosta Lopez Molina, Karen Jansen, Cláudio Drews, Ricardo Pinheiro, Ricardo Silva, and Luciano Souza. "Major depressive disorder symptoms in male and female young adults." *Psychology, Health & Medicine*, **19**(2):136–145, 2014.

[LLL22] Jiahao Lu, Bin Liu, Zheng Lian, Cong Cai, Jianhua Tao, and Ziping Zhao. "Prediction of Depression Severity Based on Transformer Encoder and CNN Model." In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 339–343. IEEE, 2022.

[LLZ23] Cheng Lu, Hailun Lian, Wenming Zheng, Yuan Zong, Yan Zhao, and Sunan Li. "Learning Local to Global Feature Aggregation for Speech Emotion Recognition." *arXiv preprint arXiv:2306.01491*, 2023.

[LML10] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. "Detection of clinical depression in adolescents' speech during family interactions." *IEEE Transactions on Biomedical Engineering*, **58**(3):574–586, 2010.

[LNZ22] Ya Li, Mingyue Niu, Ziping Zhao, and Jianhua Tao. "Automatic depression level assessment from speech by long-term global information embedding." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8507–8511. IEEE, 2022.

[Low11] Lu-Shih Low. *Detection of clinical depression in adolescents' using acoustic speech analysis*. PhD thesis, RMIT University, 2011.

[LS12] 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.

[LT15] Jinkyu Lee and Ivan Tashev. "High-level feature representation using recurrent neural network for speech emotion recognition." In *Interspeech 2015*, 2015.

[LYL21] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang. "Contrastive unsupervised learning for speech emotion recognition." In *ICASSP*, pp. 6329–6333, 2021.

[Ma19] Edward Ma. "NLP Augmentation." https://github.com/makcedward/nlpaug, 2019.

[MCL21] Shuiyang Mao, PC Ching, and Tan Lee. "Enhancing segment-based speech emotion recognition by iterative self-learning." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**:123–134, 2021.

[Mit10] Alex J Mitchell. "Why do clinicians have difficulty detecting depression." *Screening for depression in clinical practice: An evidence-based guide*, pp. 57–82, 2010.

[MRL15] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "Librosa: Audio and music signal analysis in python." In *Proceedings of the 14th Python in Science Conference*, volume 8, pp. 18–25. Citeseer, 2015.

[MSH20] Muhammad Muzammel, Hanan Salam, Yann Hoffmann, Mohamed Chetouani, and Alice Othmani. "AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis." *Machine Learning with Applications*, **2**:100005, 2020.

[MY16] 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*, pp. 35–42, 2016.

[NLT21] Mingyue Niu, Bin Liu, Jianhua Tao, and Qifei Li. "A time-frequency channel attention and vectorization network for automatic depression level prediction." *Neurocomputing*, **450**:208–218, 2021.

[ODZ16] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499*, 2016.

[OEB19] 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." *arXiv preprint arXiv:1904.01038*, 2019.

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748*, 2018.

[Org17] World Health Organization et al. "Depression and other common mental disorders: global health estimates." Technical report, World Health Organization, 2017.

[PC15] Dimitri Palaz, Ronan Collobert, et al. "Analysis of CNN-based speech recognition system using raw speech as input." 2015.

[PCP15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. "Librispeech: an asr corpus based on public domain audio books." In *ICASSP*, pp. 5206–5210, 2015.

[PCZ19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. "Specaugment: A simple data augmentation method for automatic speech recognition." *arXiv preprint arXiv:1904.08779*, 2019.

[PGM19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library." *Advances in NIPS*, **32**:8026–8037, 2019.

[PMT15] Anastasia Pampouchidou, Kostas Marias, Manolis Tsiknakis, P Simos, Fan Yang, and Fabrice Meriaudeau. "Designing a framework for assisting depression severity assessment from facial image analysis." In *ICSIPA*, pp. 578–583, 2015.

[PVG11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research*, **12**:2825–2830, 2011.

[Rav22] V. Ravi et al. "A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement." In *Proc. Interspeech*, pp. 3338–3342, 2022.

[RKM22] 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.

[RKX23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision." In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

[RWF22a] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. "FrAUG: A Frame Rate Based Data Augmentation Method for Depression Detection from Speech Signals." *arXiv preprint arXiv:2202.05912*, 2022.

[RWF22b] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. "A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement." In *Interspeech*, volume 2022, p. 3338. NIH Public Access, 2022.

[RWF24a] 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.

[RWF24b] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. "A Privacy-Preserving Unsupervised Speaker Disentanglement Method for Depression Detection from Speech." In *Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI*, volume 3649, pp. 57–63, 2024.

[SBC19] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. "wav2vec: Unsupervised pre-training for speech recognition." *arXiv preprint arXiv:1904.05862*, 2019.

[Sch21] A. Schilling et al. "Quantifying the separability of data classes in neural networks." *Neural Networks*, **139**:278–293, 2021.

[SCP17] Olympia Simantiraki, Paulos Charonyktakis, Anastasia Pampouchidou, Manolis Tsiknakis, and Martin Cooke. "Glottal Source Features for Automatic Speech-Based Depression Assessment." In *INTERSPEECH*, pp. 2700–2704, 2017.

[SEG18] Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. "A weakly supervised learning framework for detecting social anxiety and depression." *IMWUT*, **2**(2):1–26, 2018.

[Sha20] Uzzal Sharma. "Detection of Depression from Speech Signal through Linear SVM using MFCC Feature." In *ICICC*, 2020.

[She22] Ying Shen et al. "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model." In *ICASSP*, pp. 6247–6251. IEEE, 2022.

[SNR22] Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgoo, and Peter Eklund. "Audio based depression detection using Convolutional Autoencoder." *Expert Systems with Applications*, **189**:116076, 2022.

[Suh22] BN Suhas et al. "PRIVACY SENSITIVE SPEECH ANALYSIS USING FEDERATED LEARNING TO ASSESS DEPRESSION." *ICASSP*, 2022.

[TD23] Jenthe Thienpondt and Kris Demuynck. “ECAPA2: A Hybrid Neural Network Architecture and Training Strategy for Robust Speaker Embeddings.” In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

[TTN18] Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masayuki Suzuki, Toru Nagano, Ryuki Tachibana, Masafumi Nishimura, and Tetsuaki Arai. “Major depressive disorder discrimination using vocal acoustic features.” *Journal of affective disorders*, **225**:214–220, 2018.

[VG16] 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*, pp. 3–10, 2016.

[VLM14] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. “Deep neural networks for small footprint text-dependent speaker verification.” In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4052–4056. IEEE, 2014.

[VVB15] Elke Vlemincx, Ilse Van Diest, and Omer Van den Bergh. “Emotion, sighing, and respiratory variability.” *Psychophysiology*, **52**(5):657–666, 2015.

[Wan22a] D. Wang et al. “ECAPA-TDNN Based Depression Detection from Clinical Speech.” In *Proc. Interspeech*, pp. 3333–3337, 2022.

[Wan22b] J. Wang et al. “Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals.” In *Proc. Interspeech*, pp. 2018–2022, 2022.

[WLG15] Lingyun Wen, Xin Li, Guodong Guo, and Yu Zhu. “Automated depression diagnosis based on facial dynamic analysis and sparse coding.” *IEEE Transactions on Information Forensics and Security*, **10**(7):1432–1441, 2015.

[WLL24] Yong Wang, Cheng Lu, Hailun Lian, Yan Zhao, Björn Schuller, Yuan Zong, and Wenming Zheng. “Speech Swin-Transformer: Exploring a Hierarchical Transformer with Shifted Windows for Speech Emotion Recognition.” *arXiv preprint arXiv:2401.10536*, 2024.

[WLZ21] Hongbo Wang, Yu Liu, Xiaoxiao Zhen, and Xuyan Tu. “Depression speech recognition with a three-dimensional convolutional network.” *Frontiers in human neuroscience*, **15**:713823, 2021.

[WQH13] James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. “Vocal biomarkers of depression based on motor incoordination.” In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 41–48, 2013.

[WRA23] Jinhan Wang, Vijay Ravi, and Abeer Alwan. "Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals." In *Proc. INTERSPEECH 2023*, pp. 2343–2347, 2023.

[WRF22] Jinhan Wang, Vijay Ravi, Jonathan Flint, and Abeer Alwan. "Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals." In *Interspeech*, volume 2022, p. 2018. NIH Public Access, 2022.

[WRF24] Jinhan Wang, Vijay Ravi, Jonathan Flint, and Abeer Alwan. "Speechformer-CTC: Sequential Modeling of Depression Detection with Speech Temporal Classification." *Speech Communication, Manuscript submitted for publication*, 2024.

[WWL23] Pingping Wu, Ruihao Wang, Han Lin, Fanlong Zhang, Juan Tu, and Miao Sun. "Automatic depression recognition by intelligent speech signal processing: A systematic survey." *CAAI Transactions on Intelligence Technology*, **8**(3):701–711, 2023.

[WWY22] Wen Wu, Mengyue Wu, and Kai Yu. "Climate and weather: Inspecting depression detection via emotion recognition." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6262–6266. IEEE, 2022.

[WZF21] Jinhan Wang, Yunzheng Zhu, Ruchao Fan, Wei Chu, and Abeer Alwan. "Low Resource German ASR with Untranscribed Data Spoken by Non-native Children–INTERSPEECH 2021 Shared Task SPAPL System." *arXiv preprint arXiv:2106.09963*, 2021.

[Yan12] Ying Yang et al. "Detecting depression severity from vocal prosody." *IEEE transactions on affective computing*, **4**(2):142–150, 2012.

[YDX23] Faming Yin, Jing Du, Xinzhou Xu, and Li Zhao. "Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks." *Electronics*, **12**(2):328, 2023.

[Yin20] Yufeng Yin et al. "Speaker-invariant adversarial domain adaptation for emotion recognition." In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 481–490, 2020.

[YJS20] Le Yang, Dongmei Jiang, and Hichem Sahli. "Feature augmenting networks for improving depression severity estimation from speech signals." *IEEE Access*, **8**:24033–24045, 2020.

[YLC23] Wenju Yang, Jiankang Liu, Peng Cao, Rongxin Zhu, Yang Wang, Jian K Liu, Fei Wang, and Xizhe Zhang. "Attention guided learnable time-domain filterbanks for speech depression detection." *Neural Networks*, 2023.

[YZY19] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. "Unsupervised embedding learning via invariant and spreading instance feature." In *Proceedings of the IEEE/CVF CVPR*, pp. 6210–6219, 2019.

[ZBZ19] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Jun Deng, Nicholas Cummins, Haishuai Wang, Jianhua Tao, and Björn Schuller. "Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders." *IEEE Journal of Selected Topics in Signal Processing*, **14**(2):423–434, 2019.

[ZGH22] Zhiyuan Zhou, Yanrong Guo, Shijie Hao, and Richang Hong. "Hierarchical Multifeature Fusion via Audio-Response-Level Modeling for Depression Detection." *IEEE transactions on computational social systems*, 2022.

[ZLC20] Ziping Zhao, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, and Björn Schuller. "Hybrid network feature extraction for depression assessment from speech." 2020.

[ZLD21] Yan Zhao, Zhenlin Liang, Jing Du, Li Zhang, Chengyu Liu, and Li Zhao. "Multi-head attention-based long short-term memory for depression detection from speech." *Frontiers in Neurorobotics*, **15**:684037, 2021.

[ZLG22] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186. IEEE, 2022.

[ZWD21] Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. "Depa: Self-supervised audio embedding for depression detection." In *Proceedings of the 29th ACM-MM*, pp. 135–143, 2021.