Title

Speechformer-CTC: Sequential Modeling of Depression Detection with Speech Temporal Classification.

Authors

Wang, Jinhan
Ravi, Vijay
Flint, Jonathan Frederic Rest
et al.

Publication Date

2024-09-01

DOI

10.1016/j.specom.2024.103106

Peer reviewed

HHS Public Access

Author manuscript

Speech Commun. Author manuscript; available in PMC 2024 October 03.

Published in final edited form as:

Speech Commun. 2024 September ; 163: . doi:10.1016/j.specom.2024.103106.

Speechformer-CTC: Sequential Modeling of Depression Detection with Speech Temporal Classification

Jinhan Wang[a,*], **Vijay Ravi**[a], **Jonathan Flint**[b], **Abeer Alwan**[a]

[a]Department of Electrical and Computer Engineering, University of California, Los Angeles, California, 90095, USA

[b]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, California, 90095, USA

Abstract

Speech-based automatic depression detection systems have been extensively explored over the past few years. Typically, each speaker is assigned a single label (Depressive or Non-depressive), and most approaches formulate depression detection as a speech classification task without explicitly considering the non-uniformly distributed depression pattern within segments, leading to low generalizability and robustness across different scenarios. However, depression corpora do not provide fine-grained labels (at the phoneme or word level) which makes the dynamic depression pattern in speech segments harder to track using conventional frameworks. To address this, we propose a novel framework, Speechformer-CTC, to model non-uniformly distributed depression characteristics within 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 on 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 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. On the DAIC-WOZ dataset, the system with HuBERT ASR features and a CTC objective optimized using HuBERT policy for label generation achieves 83.15% F1-score, which is close to state-of-the-art without the need for phoneme-level transcription or data augmentation. On the CONVERGE dataset, using Whisper features with the HuBERT policy improves the F1-score by 9.82% on CONVERGE1 (in-domain test set) and 18.47% on CONVERGE2 (out-of-domain test set). These findings show that depression detection can benefit from modeling non-uniformly distributed depression patterns and the proposed framework can be potentially used to determine significant depressive regions in speech utterances.

Keywords

Depression-detection; CTC; Non-uniform distribution

[*]Corresponding author: wang7875@g.ucla.edu.

1. Introduction

Recently, depression detection through verbal cues has gained substantial traction, largely due to the ease of collecting speech data and the rapid advancements in neural network-based modeling methods [1, 2, 3, 4]. Previous research has extensively examined speech-based depression detection from various perspectives, encompassing diverse acoustic features (vocal source [5, 6], voice quality [7], vocal tract articulators [8]), model types (spatial [9], Recurrent Neural Networks (RNNs) [10], self-attention [11]), data augmentation (Generative Adversarial Network (GAN) [12], FrAUG [13]) and backend modeling techniques (transfer learning [14], self-supervised pre-training [15, 16]).

Despite notable performance improvements over the years [17, 18, 19], several challenges persist. One of such challenges is that 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 [20, 21, 22]. Previous research has unveiled that depression severity affects different regions of speech segments disparately, considering both time and frequency intervals [21]. This influence is evident in word choices [23] and variations in vowels between individuals with depression and those without [24].

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 [25]. 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.

To overcome the challenges mentioned earlier, we present a Speechformer-CTC (Connectionist Temporal Classification) framework. Speechformer is a hierarchical architecture proposed to model cognitive speech-processing tasks, including Speech Emotion Recognition (SER) and depression detection [26]. This setup utilizes the CTC loss objective function [27] and works as follows - first, for each speech segment, a CTC-label sequence is generated. Next, in parallel with the depression classification task, the depression detection task is reframed into a sequential modeling problem that involves utilizing the generated CTC-labels, regularizing intermediate representations at different granularities, and optimizing the model using the CTC loss. Additionally, for the sequential modeling task, alignment between the input and the generated CTC-label sequence is unnecessary because the CTC framework allows for label collapsing, and hence, it accommodates input-output sequence length mismatches. This proposed approach, therefore, allows us to capture salient depression regions in an alignment-free manner.

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. Our contributions can be summarized as follows:

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 states 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 Automatic Speech Recognition (ASR) features with the proposed method further enhances the detection performance, demonstrating the compatibility of the proposed method with content-related information.

The remainder of the paper is organized as follows. Section 2 reviews related work in the base model architecture, Speechformer [26], and in modeling non-uniform paralinguistic speech patterns. Our proposed methods are introduced in Section 3. Experimental details, including model, datasets, and features, are introduced in Section 4. The results are presented in Section 5 along with a detailed discussion and analysis of model performance. We summarize the article in Section 6 along with suggestions for future work.

2. Related Work

In this section, we begin by examining previous studies on model architectures for detecting depression with a focus on Speechformer, the neural network at the core of our proposed method. We then review studies exploring the temporal non-uniformity in related domains of speech-based emotion recognition and depression detection.

2.1. Model Architectures

Automatic depression detection systems were previously built based on architectures like Gaussian Mixture Models [28], logistic regression [7], Naive Bayes and Random Forest [29]. 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 [10, 30, 31]. Notably, Convolutional Neural Networks (CNNs), RNNs, and their variants have gained prominence in the research community. DepAudioNet [10], 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 [13, 32, 33, 34, 3]. In [35], a dilated CNN is proposed and coupled with full vocal tract coordination features to leverage its potential in determining depression states. [36] 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 [37], the Transformer encoder is cascaded with a CNN to capture long-range dependencies. In [9], a parallel CNN-Transformer architecture to simultaneously capture local knowledge and

temporal sequential information was proposed. More recently, a new framework named "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 [26]. Additionally, the Speechformer model achieved superior performance compared to its Transformer counter-part.

Given the variability in the characteristics of depression across different granularities of the speech signal [6, 38, 39, 40], 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 for our study.

2.1.1. Speechformer—In [26], two key ideas were introduced based on the Transformer model to facilitate 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 [26]. 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.

2.2. Non-uniform Temporal Variation in Depression

Several studies have attempted to leverage the non-uniformly distributed speech patterns in depression detection tasks. In [22], 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 [41], a long-term global information embedding (GIE) is proposed to reweight each frame output from the LSTM module, allowing frames with more significant depression cues to be emphasized through the attention function. Additionally, in [21], the authors introduce the Time-Frequency Attention (TFA) and merge it with the Squeeze-and-Excitation component to emphasize timestamps, frequency bands, and 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 "un-supervised" 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 classification tasks has been investigated in SER, another domain that can benefit from modeling the non-uniformity in a given speech utterance [42, 43, 44]. Prior studies include [45, 46, 47, 48], where a CTC [27] objective is used to reformulate the SER task into a sequence-to-sequence task. Here, we briefly introduce the CTC method.

Denote an input sequence $\mathbf{X} = [x_1, \ldots, x_t, \ldots x_T]$ of length T, and the corresponding target sequence $\mathbf{Y} = [y_1, \ldots, y_m, \ldots y_M]$ of length M. Let $\mathbf{Z} = [z_1, \ldots, z_t, \ldots z_T]$ 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 Z's are defined as those Z's where $Z = \beta^{-1}(Y)$. Since the collapse function β is a many-to-one mapping, multiple Z's can be mapped to the groundtruth target sequence Y. As a result, CTC loss is defined as the summation of all valid Z's negative log-probabilities, which is formulated as:

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

Building upon the achievements in SER, notably in pseudo-label generalization and optimizing models via CTC-based methods, our paper presents a novel framework for depression detection. This framework aims to capture the non-uniformly distributed patterns of depression through sequential modeling methodology.

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 content features.

3.1. Speechformer-CTC

Before introducing the proposed Speechformer-CTC framework (shown in Figure 1), the backbone Speechformer model [26] 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 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 [26] (detailed description in Appendix B). 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 [49, 50]. However, for such a non-aligned sequence-to-sequence task, CTC objective optimization is an ideal candidate [27, 51]. 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[1].

Before describing the proposed CTC-label generation policy, let us define the variables in the framework. Based on definitions of Speechformer variables in previous paragraphs, 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:

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

$$
\begin{aligned}
L_{Dep} &= -y \cdot \log(\tilde{y}) - (1-y) \cdot \log(1-\tilde{y}) \\
L_{CTC} &= -\log P\left(Y_{ctc} \mid \hat{X}_s\right) \\
&= -\log \sum_{Z \in \beta^{-1}(Y_{ctc})} P\left(Z \mid \hat{X}_s\right) \\
&= -\log \sum_{Z \in \beta^{-1}(Y_{ctc})} \prod_{t=1}^{T_s} P\left(z_t \mid \hat{X}_s\right) \quad s \in \left\{F, P, W, U\right\} \\
L_{total} &= L_{Dep} + \alpha L_{CTC}
\end{aligned}
\tag{2}
$$

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 choose to apply the CTC objective on each stage separately to investigate the effect of different granularities for depression state alignment.

Compared to previous SER works [45, 46, 48], 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 [14], voiced/unvoiced regions [39], or different vowels [2]. Therefore, preserving L_{Dep} can avoid overfitting the model to only some local speech patterns.

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.

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 [45, 47, 48], 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 [45, 46, 47], 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:

$$Y_{ctc} = \begin{cases} 0^M & y = 0 \\ 1^M & y = 1 \end{cases}$$
$$\text{where } M = \frac{\text{len}(\hat{X}_s)}{k} \quad s \in \{F, P, W, U\} \tag{3}$$

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 [26] is selected as the baseline and for the Naive-One-Hot experiment, the proposed Speechformer-CTC framework (Section 3.1) is used. Experiments are conducted on the DAIC-WOZ dataset [52]. DAIC-WOZ is an English dataset with 133 ND and 56 D speakers. 128-dimensional log-melspectrograms are used as input features. Evaluation is conducted using F1-scores. Detailed configurations are described in Section 4.

Table 1 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.

To visualize the non-uniform distribution of depression patterns, the CTC scores for ND and D classes are plotted for four speakers in Figures 2 and 3, respectively.

As shown in Figures 2(a) and 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 2(b), we observe that, for this non-depressed individual, certain regions exhibit higher D class scores compared to ND class scores. Similarly, Figure 3(b) displays a comparable pattern, where only half of the session demonstrates higher D class scores than ND class scores for a depressed 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 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 [53] 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 [54] features for Mandrain).

3.3. Expectation-based One-Hot Policy (E-One-Hot)

As observed in the preliminary experiments described in Section 3.2, even for individuals with depression, some cases have majority speech regions classified as ND. 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, with filler *Null* states before and after this region. Repeatedly choosing different values for M randomly 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:

$$
\begin{aligned}
&M_i \sim \text{uniform}\left(1, \frac{\text{len}\left(\hat{X}_{s,i}\right)}{2}\right)\\
&Y_{ctc,i} = \begin{cases} 0^{M_i} & y = 0 \\ 1^{M_i} & y = 1 \end{cases}\\
&L_{CTC,i} = -\log P\left(Y_{ctc,i} \mid \hat{X}_{s,i}\right)\\
&L_{CTC} \approx -\mathbb{E}_{Y_{ctc}}\left[\log P\left(Y_{ctc} \mid \hat{X}_s\right)\right]\\
&\text{where } s \in \left\{F, P, W, U\right\}
\end{aligned}
\tag{4}
$$

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

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 [53], pseudo-labels generated by clustering algorithms can reveal implicit correlations between hidden representations and underlying acoustic units. For example, [55] demonstrates that fine-tuning a HuBERT model can provide frame-level pseudo-emotion labels for SER, aiding in distinguishing emotional/non-emotional frames. Additionally, [16] 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 $\mathbf{A} = [a_1, \ldots, a_n, \ldots a_{N_F}]$, where N_F is frame-level feature length. Depending on the stage at which the HuBERT label will be used, the dimension of $\mathbf{A}$ can be reduced through average pooling operations following merging scales presented in Appendix B. The resulting feature sequence lengths are denoted by N_P, N_W, and N_U, for phoneme, word, and utterance, 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 $\mathbf{C} = [c_1, \ldots, c_n, \ldots c_{N_s}]$, 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.

$$
\begin{aligned}
A &= \text{HuBERT}\left(Raw\ Audio\right) \\
K_e &\sim f(A_e, k) \ \text{ where } e \in \left\{D, ND\right\} \\
C_i &= \begin{cases} K_{ND}(A_i) & y_i = 0 \\ K_D(A_i) & y_i = 1 \end{cases} \\
Y_{ctc,i} &= \begin{cases} \gamma(C_i) & y_i = 0 \\ \gamma(C_i) + k & y_i = 1 \end{cases}
\end{aligned}
\tag{5}
$$

where f is 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.

3.5. Content Features

Previous research has shown that text modality is effective in depression detection [56, 57]. 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 [23]. 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.

4. Experimental Details

The effectiveness of the proposed methods is shown on two datasets, one in English and the other one in Mandarin. Detailed experimental setups of dataset configurations, acoustic features, model, and training/evaluation scheme are presented in this section.

4.1. Datasets

4.1.1. DAIC-WOZ—The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [52] is an English dataset containing audio, text, and video of interviews collected by a virtual interviewer from 189 male and female participants. The audio files have durations ranging from 7 ~ 33 min (16 min on average), with a non-depression vs depression ratio of 3:1 and a total duration ~50 hours, sampled at 16kHz. Segments are extracted from the responses of speakers using the provided time stamps , where each segment behaves as one sample in the dataset. The dataset is split into train, validation, and test sets with a ratio of 107: 35: 47, following the official split in [52]. Sessions with errors are properly handled manually by the authors based on the provided documentation.

4.1.2. CONVERGE—The Mandarin dataset used is part of China, Oxford, and Virginia Commonwealth University Experimental Research Genetic Epidemiology (CONVERGE) [58]. The dataset is divided 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 speaker are segmented into multiple audio clips by annotators ,where each clip is a data sample. 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 newly collected dataset 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.

A summary of the statistics of the two datasets is shown in Table 2.

4.2. Acoustic Features

Log-melspectrogram (log-mel) features are selected as the acoustic features for a fair comparison with the Speechformer study [26]. 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.

In experiments where content features are explored, representations extracted from pre-trained ASR models are used as input features. For the DAIC-WOZ dataset, the HuBERT-large model [53], pre-trained on 60k hours of Libr-light [59] and fine-tuned on 960 hours of Librispeech [60], is used to extract the 1024-dimensional representation with hop size of 20ms. Extraction is performed using the fairseq toolkit [61]. Regarding the CONVERGE datasets, Whisper [54] is selected as the feature 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-medium model demonstrates good performance in terms of Word Error Rate (WER) on Mandarin ASR tasks. Extracted Whisper feature vectors have a dimension of 1024 with a hop size of 20ms [62]. Throughout the experiments, pre-trained feature encoders are frozen without updating their weights at any stage.

4.3. 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 [53] pre-trained on Libirspeech-light is used [59]. Regarding the CONVERGE dataset, a Chinese-HuBERT-large model, pre-trained on the 10k hours WenetSpeech training set [63] is applied. In the clustering phase, MinibatchKmeans is fitted using the Scikit-learn package [64], aligning with the label generation method used in HuBERT pre-training [53].

4.4. 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 [26]. 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 $2k + 1$, where k is the number of HuBERT clusters. The additional 1 output token is reserved for the *Null* label in the HuBERT policy scenario.

4.5. Training and Evaluation Scheme

As indicated in Table 2, 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 to engage in conversation for depressed individuals. 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 [65]. 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 $1e \square 3$ for DAIC-WOZ and 0 for CONVERGE. Regarding the factor α which controls the CTC-loss weight, we start the training and inspect the CTC-loss scale. The value of α is then chosen to maintain CTC-loss and CE-loss to be at the same scale. The best-performing model on the validation set is selected for evaluation on the test set.

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. Detection 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. Results and Discussion

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.

5.1. CTC Label Generation Policies

5.1.1. E-One-Hot—Results for applying the E-One-Hot CTC label generation on different stages (*Frame*, *Phoneme*, *Word*, *Utterance*) of the SpeechFormer-CTC model are reported and compared against the Naive-One-Hot policy in Table 3. Naive-One-Hot results are reported in Appendix A. 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 Appendix A).

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 on 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.

5.1.2. HuBERT Policy—Performance using CTC-label generation through HuBERT policy for various stages are shown in Figure 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 numbers of K-means clusters *k*, we observe the best overall performances when setting $k = 10$. In contrast, $k = 15$ results in the worst performance on 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.

The predicted CTC tokens obtained from the trained Speechformer-CTC models, excluding the *Null* token, are plotted in Figure 5. *k* values are varied among 5, 10 or 15 and the CTC-loss is applied on the *W* stage. For each sample, the token with maximum probabilities, among ND or D HuBERT labels, is selected as the predicted token.[2]

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

[2]To make plots readable, HuBERT labels for the D class samples are shifted down by the bias factor *k* according to Eq. 5, such that all plots use the same cluster index (from 0 to $k - 1$).

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(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 are not contributing towards depression detection. Therefore, during training, those additional irrelevant clusters can result in more invalid alignment paths and degrade the performance. These results therefore suggest that utilizing HuBERT labels is an effective way of introducing prior knowledge in 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 6. The examples show that predicted HuBERT labels do correspond to some intuitive depression patterns, such as reduced volume [66], sighing [67], prolongation [68], and whispering [69] 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.

5.2. 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 4.

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 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.

5.3. Extension to the CONVERGE Datasets

We apply the proposed methods to the CONVERGE datasets to verify their generalizability to another language, which is 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. This is as expected 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 without the necessity to introduce label length expectation.

By utilizing the HuBERT policy, F1-avg performance 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.

5.4. Comparison to SOTA Studies

In this section, we compare our proposed methods with existing SOTA studies for depression detection as shown in Table 6.

On the DAIC-WOZ dataset, we achieve close-to-SOTA performance (Audvowconsnet [70]), 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 [72]. 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 outperforms three previous works in terms of F1-scores on CONVERGE1.

6. Conclusion and Future Work

In this paper, 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. Additionally, experiments using ASR features are conducted to demonstrate the compatibility of the proposed method with content-based features. 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.

Acknowledgement

This work was funded by the National Institutes of Health under the award number R01MH122569- Combining Voice and Genetic Information to Detect Heterogeneity in Major Depressive Disorder.

Appendix

Appendix

Appendix A. Naive-One-Hot Results

Results of the baseline and using Naive-One-Hot policy label generation on different stages are presented in Table A.7.

Table A.7:

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.

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

Appendix B. Speechformer Merge Scales

Merging scales of the Speechformer model [26] across different stages are manually determined through statistical speech characteristics as shown in Table B.8.

Table B.8:

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

References

[1]. Othmani A, Kadoch D, Bentounes K, Rejaibi E, Alfred R, Hadid A, Towards robust deep neural networks for affect and depression recognition from speech, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II, Springer, 2021, pp. 5–19.

[2]. Feng K, Chaspari T, 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), IEEE, 2023, pp. 1–5.

[3]. Ravi V, Wang J, Flint J, Alwan A, Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement, Computer Speech & Language 86 (2024) 101605. [PubMed: 38313320]

[4]. Huang Z, Epps J, Joachim D, Investigation of speech landmark patterns for depression detection, IEEE Transactions on Affective Computing (2019).

[5]. Dubagunta SP, Vlasenko B, Doss MM, Learning voice source related information for depression detection, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6525–6529.

[6]. Koops S, Brederoo SG, de Boer JN, Nadema FG, Voppel AE, Sommer IE, Speech as a biomarker for depression, CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders) 22 (2023) 152–160.

[7]. Afshan A, Guo J, Park SJ, Ravi V, Flint J, Alwan A, Effectiveness of Voice Quality Features in Detecting Depression, in: Proc. Interspeech 2018, 2018, pp. 1676–1680. doi:10.21437/Interspeech.2018-1399.

[8]. Cummins N, Vlasenko B, Sagha H, Schuller B, 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, Springer, 2017, pp. 209–214.

[9]. Yin F, Du J, Xu X, Zhao L, Depression detection in speech using transformer and parallel convolutional neural networks, Electronics 12 (2023) 328.

[10]. Ma X, Yang H, Chen Q, Huang D, Wang Y, Depaudionet: An efficient deep model for audio based depression classification, in: Proceedings of the 6th international workshop on audio/visual emotion challenge, 2016, pp. 35–42.

[11]. Zhao Z, Li Q, Cummins N, Liu B, Wang H, Tao J, Schuller BW, Hybrid Network Feature Extraction for Depression Assessment from Speech, in: Proc. Interspeech 2020, 2020, pp. 4956–4960. doi:10.21437/Interspeech.2020-2396.

[12]. Yang L, Jiang D, Sahli H, Feature augmenting networks for improving depression severity estimation from speech signals, IEEE Access 8 (2020) 24033–24045.

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

[14]. Wu W, Wu M, Yu K, Climate and weather: Inspecting depression detection via emotion recognition, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6262–6266.

[15]. Zhang P, Wu M, Dinkel H, Yu K, Depa: Self-supervised audio embedding for depression detection, in: Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 135–143.

[16]. Wang J, Ravi V, Flint J, Alwan A, Unsupervised instance discriminative learning for depression detection from speech signals, in: Interspeech, volume 2022, NIH Public Access, 2022, p. 2018. [PubMed: 36341466]

[17]. Chen W, Xing X, Xu X, Pang J, Du L, Speechformer++: A hierarchical efficient framework for paralinguistic speech processing, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023) 775–788.

[18]. Feng K, Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels, in: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, 2022, pp. 01–07.

[19]. Tasnim M, Ehghaghi M, Diep B, Novikova J, Depac: a corpus for depression and anxiety detection from speech, arXiv preprint arXiv:2306.12443 (2023).

[20]. Simantiraki O, Charonyktakis P, Pampouchidou A, Tsiknakis M, Cooke M, Glottal source features for automatic speech-based depression assessment., in: INTERSPEECH, 2017, pp. 2700–2704.

[21]. Niu M, Liu B, Tao J, Li Q, A time-frequency channel attention and vectorization network for automatic depression level prediction, Neurocomputing 450 (2021) 208–218.

[22]. Wang H, Liu Y, Zhen X, Tu X, Depression speech recognition with a three-dimensional convolutional network, Frontiers in human neuroscience 15 (2021) 713823. [PubMed: 34658815]

[23]. Corbin L, Griner E, Seyedi S, Jiang Z, Roberts K, Boazak M, Rad AB, Clifford GD, Cotes RO, 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 (2023) 100645.

[24]. Yang W, Liu J, Cao P, Zhu R, Wang Y, Liu JK, Wang F, Zhang X, Attention guided learnable time-domain filterbanks for speech depression detection, Neural Networks 165 (2023) 135–149. [PubMed: 37285730]

[25]. Khan DM, Yahya N, Kamel N, Faye I, Automated diagnosis of major depressive disorder using brain effective connectivity and 3d convolutional neural network, Ieee Access 9 (2021) 8835–8846.

[26]. Chen W, Xing X, Xu X, Pang J, Du L, Speechformer: A hierarchical efficient framework incorporating the characteristics of speech, arXiv preprint arXiv:2203.03812 (2022).

[27]. Graves A, Fernández S, Gomez F, Schmidhuber J, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[28]. Williamson JR, Quatieri TF, Helfer BS, Horwitz R, Yu B, Mehta DD, Vocal biomarkers of depression based on motor incoordination, in: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013, pp. 41–48.

[29]. Liu Z, Hu B, Liu F, Kang H, Li X, Yan L, Wang T, Evaluation of depression severity in speech, in: Brain Informatics and Health: International Conference, BIH 2016, Omaha, NE, USA, October 13-16, 2016 Proceedings, Springer, 2016, pp. 312–321.

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

[31]. Lam G, Dongyan H, Lin W, Context-aware deep learning for multi-modal depression detection, in: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2019, pp. 3946–3950.

[32]. Ravi V, Wang J, Flint J, Alwan A, A step towards preserving speakers' identity while detecting depression via speaker disentanglement, in: Interspeech, volume 2022, NIH Public Access, 2022, p. 3338. [PubMed: 36341467]

[33]. Wang J, Ravi V, Alwan A, Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals, in: Proc. INTERSPEECH 2023, 2023, pp. 2343–2347. doi:10.21437/Interspeech.2023-2101.

[34]. Ravi V, Wang J, Flint J, Alwan A, 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, 2024, pp. 57–63.

[35]. Huang Z, Epps J, Joachim D, Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6549–6553.

[36]. Zhao Y, Liang Z, Du J, Zhang L, Liu C, Zhao L, Multi-head attention-based long short-term memory for depression detection from speech, Frontiers in Neurorobotics 15 (2021) 684037. [PubMed: 34512301]

[37]. Lu J, Liu B, Lian Z, Cai C, Tao J, Zhao Z, Prediction of depression severity based on transformer encoder and cnn model, in: 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2022, pp. 339–343.

[38]. Zhao Z, Bao Z, Zhang Z, Deng J, Cummins N, Wang H, Tao J, Schuller B, Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders, IEEE Journal of Selected Topics in Signal Processing 14 (2019) 423–434.

[39]. Muzammel M, Salam H, Hoffmann Y, Chetouani M, Othmani A, Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis, Machine Learning with Applications 2 (2020) 100005.

[40]. Zhou Z, Guo Y, Hao S, Hong R, Hierarchical multifeature fusion via audio-response-level modeling for depression detection, IEEE transactions on computational social systems (2022).

[41]. Li Y, Niu M, Zhao Z, Tao J, 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), IEEE, 2022, pp. 8507–8511.

[42]. Wang Y, Lu C, Lian H, Zhao Y, Schuller B, Zong Y, Zheng W, Speech swin-transformer: Exploring a hierarchical transformer with shifted windows for speech emotion recognition, arXiv preprint arXiv:2401.10536 (2024).

[43]. Lu C, Lian H, Zheng W, Zong Y, Zhao Y, Li S, Learning local to global feature aggregation for speech emotion recognition, arXiv preprint arXiv:2306.01491 (2023).

[44]. Lin W-C, Busso C, Sequential modeling by leveraging non-uniform distribution of speech emotion, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023) 1087–1099.

[45]. Lee J, Tashev I, High-level feature representation using recurrent neural network for speech emotion recognition, in: Proc. Interspeech 2015, 2015, pp. 1537–1540. doi:10.21437/Interspeech.2015-336.

[46]. Chen X, Han W, Ruan H, Liu J, Li H, Jiang D, 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), IEEE, 2018, pp. 1–6.

[47]. Han W, Ruan H, Chen X, Wang Z, Li H, Schuller B, Towards Temporal Modelling of Categorical Speech Emotion Recognition, in: Proc. Interspeech 2018, 2018, pp. 932–936. doi:10.21437/Interspeech.2018-1858.

[48]. Chernykh V, Prikhodko P, Emotion recognition from speech with recurrent neural networks, arXiv preprint arXiv:1701.08071 (2017).

[49]. Wang J, Zhu Y, Fan R, Chu W, Alwan A, Low Resource German ASR with Untranscribed Data Spoken by Non-Native Children — INTERSPEECH 2021 Shared Task SPAPL System, in: Proc. Interspeech 2021, 2021, pp. 1279–1283. doi:10.21437/Interspeech.2021-1974.

[50]. Fan R, Afshan A, Alwan A, Bi-apc: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's asr, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 7023–7027.

[51]. Fan R, Wang Y, Gaur Y, Li J, Ctcbert: Advancing hidden-unit bert with ctc objectives, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[52]. Valstar M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres Torres M, Scherer S, Stratou G, Cowie R, Pantic M, Avec 2016: Depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th international workshop on audio/visual emotion challenge, 2016, pp. 3–10.

[53]. Hsu W-N, Bolte B, Tsai Y-HH, Lakhotia K, Salakhutdinov R, Mohamed A, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460.

[54]. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[55]. Li Q, Gao Y, Wang C, Deng Y, Xue J, Han Y, Li Y, Frame-level emotional state alignment method for speech emotion recognition, arXiv preprint arXiv:2312.16383 (2023).

[56]. Havigerová JM, Haviger J, Kučera D, Hoffmannová P, Text-based detection of the risk of depression, Frontiers in psychology 10 (2019) 513. [PubMed: 30936845]

[57]. Al Hanai T, Ghassemi MM, Glass JR, Detecting depression with audio/text sequence modeling of interviews., in: Interspeech, 2018, pp. 1716–1720.

[58]. Li Y, Shi S, Yang F, Gao J, Li Y, Tao M, Wang G, Zhang K, Gao C, Liu L, Li K, Liu Y, Wang X, Zhang J, Lv L, Wang X, Chen Q, Hu J, Sun L, Shi J, Chen Y, Xie D, Flint J, Kendler K, Zhang Z, Patterns of co-morbidity with anxiety disorders in chinese women with recurrent major depression, Psychological medicine 42 (2012) 1239–1248. [PubMed: 22126712]

[59]. Kahn J, Rivière M, Zheng W, Kharitonov E, Xu Q, MazaryÉ P, Karadayi J, Liptchinsky V, Collobert R, Fuegen C, Likhomanenko T, Synnaeve G, Joulin A, Mohamed A, Dupoux E, Librilight: A benchmark for asr with limited or no supervision, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7669–7673. doi:10.1109/ICASSP40776.2020.9052942.

[60]. Panayotov V, Chen G, Povey D, Khudanpur S, Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 5206–5210.

[61]. Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, Grangier D, Auli M, fairseq: A fast, extensible toolkit for sequence modeling, in: Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

[62]. Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W, Openai gym, arXiv preprint arXiv:1606.01540 (2016).

[63]. Zhang B, Lv H, Guo P, Shao Q, Yang C, Xie L, Xu X, Bu H, Chen X, Zeng C, Wu D, Peng Z, 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), 2022, pp. 6182–6186. doi:10.1109/ICASSP43922.2022.9746682.

[64]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[65]. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[66]. Cummins N, Sethu V, Epps J, Krajewski J, Probabilistic acoustic volume analysis for speech affected by depression, in: Proc. Interspeech 2014, 2014, pp. 1238–1242. doi:10.21437/Interspeech.2014-311.

[67]. Vlemincx E, Van Diest I, Van den Bergh O, Emotion, sighing, and respiratory variability, Psychophysiology 52 (2015) 657–666. [PubMed: 25524012]

[68]. Flint AJ, Black SE, Campbell-Taylor I, Gailey GF, Levinton C, Acoustic analysis in the differentiation of parkinson's disease and major depression, Journal of Psycholinguistic Research 21 (1992) 383–399. [PubMed: 1447729]

[69]. Jia Y, Liang Y, Zhu T, 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 Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, 2020, pp. 128–133.

[70]. Sardari S, Nakisa B, Rastgoo MN, Eklund P, Audio based depression detection using convolutional autoencoder, Expert Systems with Applications 189 (2022) 116076.

[71]. Han Z, Shang Y, Shao Z, Liu J, Guo G, Liu T, Ding H, Hu Q, Spatial-temporal feature network for speech-based depression recognition, IEEE Transactions on Cognitive and Developmental Systems 16 (2024) 308–318. doi:10.1109/TCDS.2023.3273614.

[72]. Ko T, Peddinti V, Povey D, Khudanpur S, Audio augmentation for speech recognition, in: Proc. Interspeech 2015, 2015, pp. 3586–3589. doi:10.21437/Interspeech.2015-711.

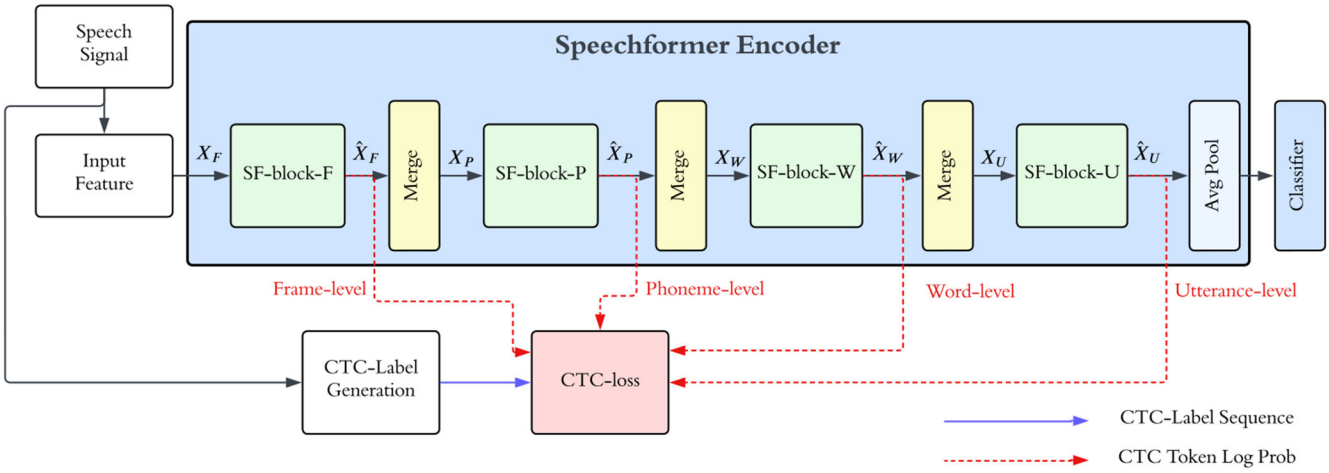

Figure 1:
A block diagram of the proposed Speechformer-CTC model.

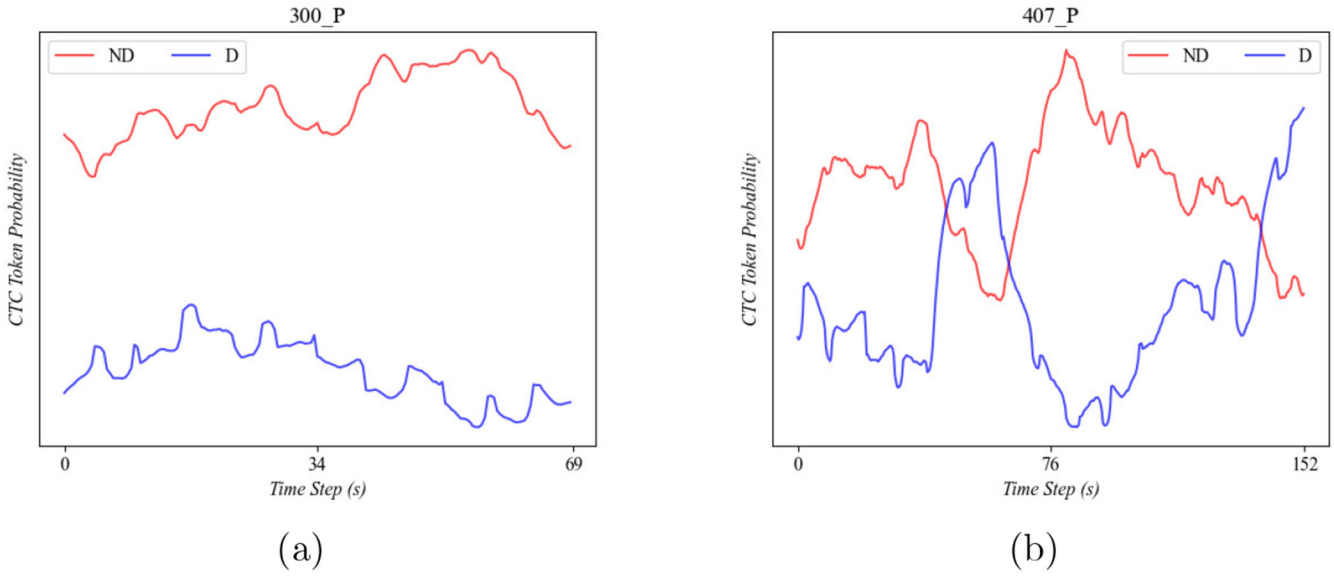

Figure 2:
Visualization of the CTC token probabilities on two ND individuals. (a) 300 (b) 407.

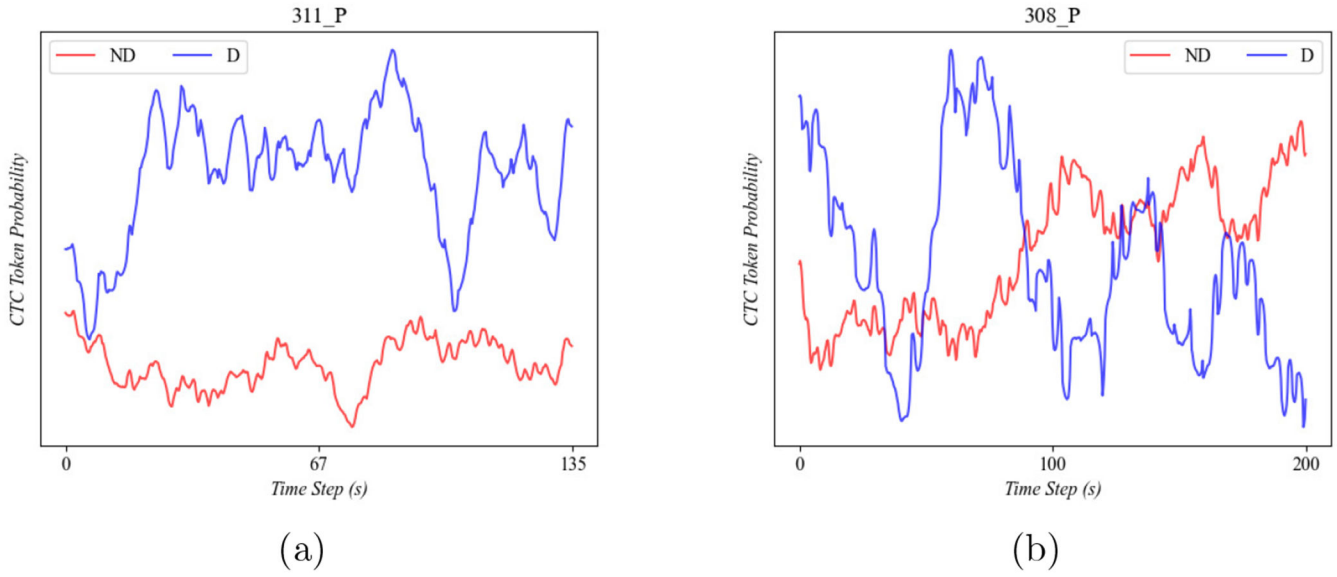

Figure 3:
Visualization of the CTC token probabilities on two D individuals. (a) 311 (b) 308.

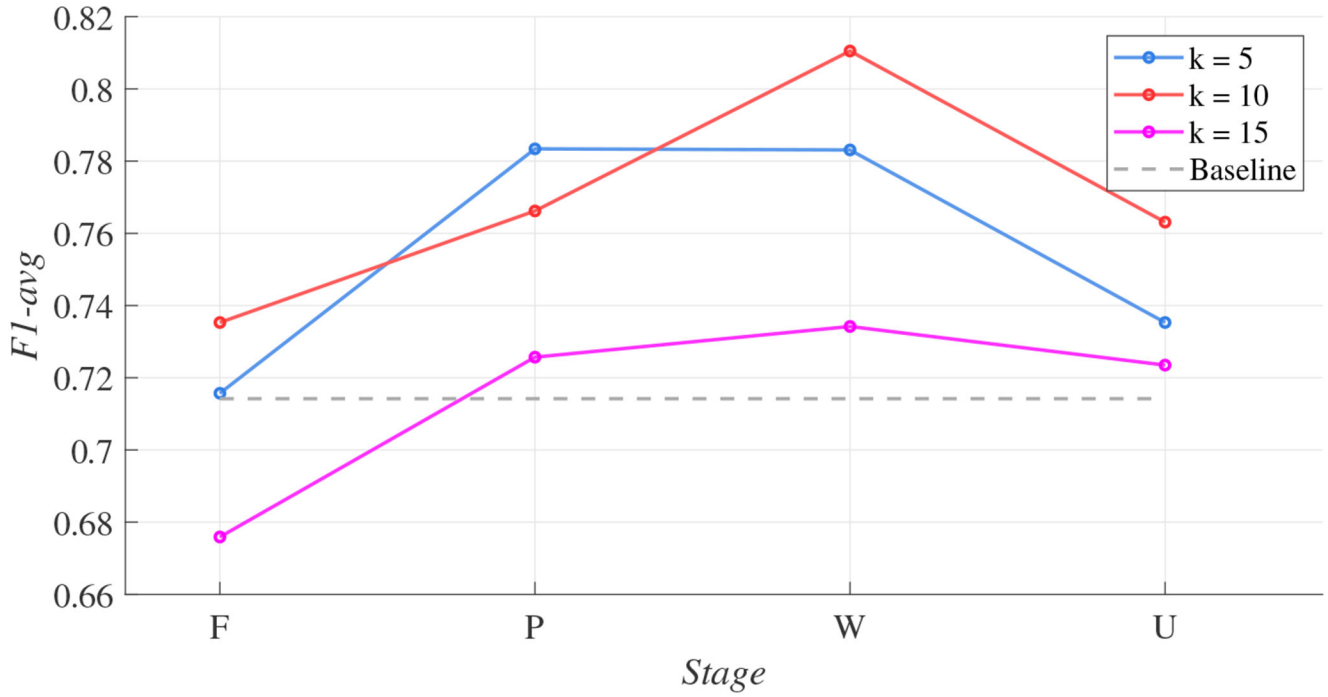

Figure 4:
F1-avg using the HuBERT policy on different stages with different number of centroids k values.

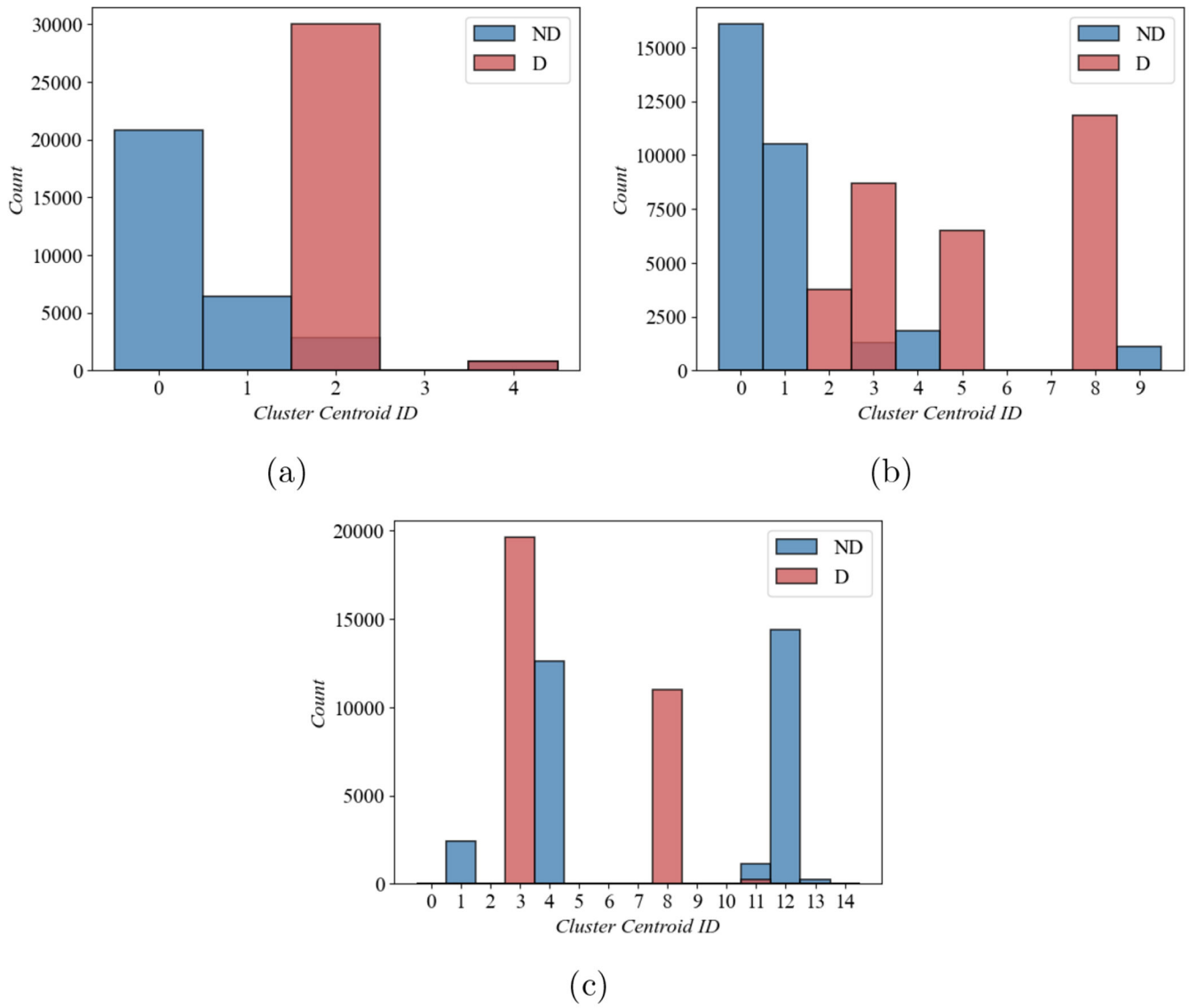

Figure 5:
Predicted HuBERT centroids distribution with different k. The X-axis of each plot represents the Hubert centroids ID, and the Y-axis is the number of predicted tokens. (a) k = 5 (b) k = 10 (c) k = 15.

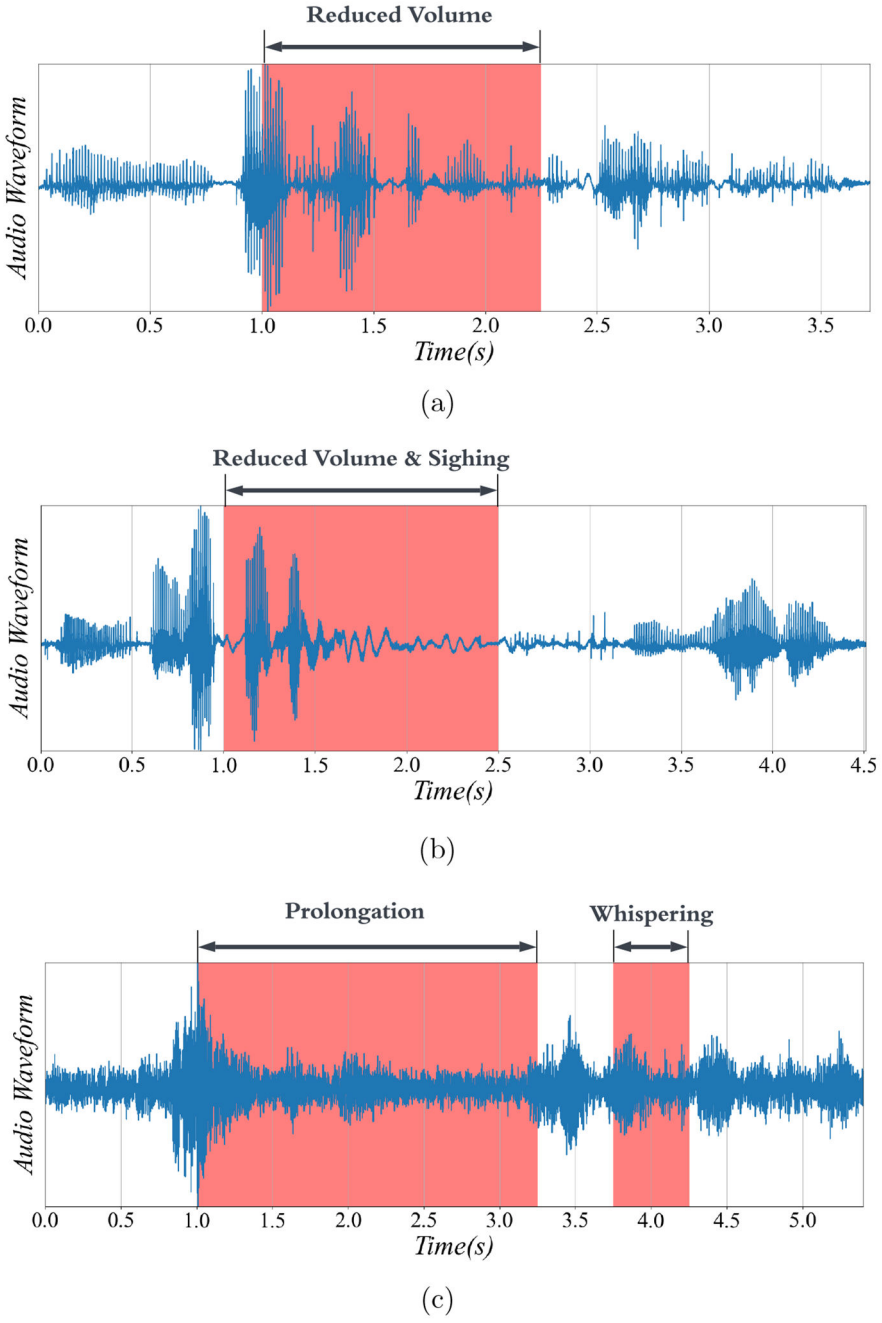

Figure 6:
Visualization of audio waveforms where highlighted regions represent clips with high probabilities of depression-related HuBERT centroids ID. (a) example1 (b) example2 (c) example3.

Table 1:

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 2:

Summary of datasets used in this paper.

	DAIC-WOZ	CONVERGE1	CONVERGE2
Language	English	Mandarin	Mandarin
Num 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
Num of Segments	32k	300k	65k
D/ND	10k/22k	219k/82k	45k/20k

Table 3:

F1-score, Precision, and Recall using E-one-hot policy on the DAIC-WOZ dataset. Δ 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*.

s	F1-score				Precision		Recall	
	Avg	ND	D	Δ 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

*means the F1-avg change is not statistically significant.

The best F1-avg score improvement is boldfaced.

Table 4:

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

Table 5:

F1-scores using the CONVERGE1 and CONVERGE2 datasets. CTC stands for CTC-label generation policy applied.

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

* stands for the change is not statistically significant.

The best F1-avg is boldfaced.

Table 6:

Comparison with SOTA depression detection studies on the DAIC-WOZ and CONVERGE1 datasets, in terms of F1-scores. The best results are bold-faced. [39] 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 [70]	0.7050	0.7100	0.7000
	Speechformer [26]	0.7142	0.8358	0.5926
	AudVowConsNet [39]	**0.8350**	**0.9000**	0.7700
	SFTN [71]	0.7550	0.8400	0.6700
	Our Work	0.8315	0.8890	**0.7742**
CONVERGE1	Fraug [13]	0.7390	-	-
	IDL [16]	0.7435	0.7548	0.7323
	Speechformer [26]	0.7463	0.7639	0.7290
	Our Work	**0.8196**	**0.8435**	**0.7956**