



ECAPA-TDNN Based Depression Detection from Clinical Speech

Dong Wang^{1,2}, Yanhui Ding¹, Qing Zhao³, Peilin Yang², Shuping Tan⁴, Ya Li^{2,*}

¹School of Information Science and Engineering, Shandong Normal University, China

²School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China

³Peking University Huilongguan Clinical Medical School, Beijing Huilongguan Hospital, China

⁴Beijing Huilongguan Hospital, Peking University Huilongguan Clinical Medical School, China

201611010218@stu.sdu.edu.cn, yanhuiding@126.com, tiancaqpl@126.com,
peilin-yang@bupt.edu.cn, shupingt@126.com, yli01@bupt.edu.cn

Abstract

Depression is a serious mood disorder that has become one of the major diseases that endanger human mental health. The automatic detection of depression using speech signals has become a promising approach for the early diagnosis of depression currently. However, there is still a performance gap between clinical practice and research, considering the lab-recorded corpus was used in most of the current studies. Therefore, we collected a Chinese clinical depression corpus, of which 131 participants with their speech during the Hamilton Rating Scale for Depression (HAMD) interview were included in this study. Furthermore, we developed a depression speech detection system based on a Time-Delay Neural Network (TDNN) model to distinguish depression¹. Our approach achieves a mean F1 score of 90.8% and an accuracy of 90.4% by five-fold cross-validation. The result suggests that the developed TDNN-based model has a potential clinical meaning in the diagnosis of depression.

Index Terms: depression, HAMD, audio classification, deep learning, TDNN

1. Introduction

Depression is a prevalent mental disease that mainly manifests as low mood, slow thought process and decreased will activity, which has become one of the major health problems today [1]. On the one hand, depression was associated with significant impairment in mental health. On the other hand, comorbidity is common with depression and other neuropsychiatric disorders, such as anxiety disorders [2] and Alzheimer's disease [3]. Therefore, accurate and automatic detection of depression at the earliest disease stage is crucial for the health and treatment of depression patients.

The use of speech signals for detecting depression has become a focus of research since massive information about the mental health of speakers is carried by speech [4]. Alpert *et al.* [5] found that depressed patients showed less prosody as early as 2003, then Cannizzaro *et al.* [6] found that the speaking rate and percent pause time correlated with HAMD score. Recently, deep learning techniques have been successfully applied to depression detection and significantly improve the performance. Ma *et al.* [7] proposed DepAudioNet to classify the speech of patients with depression. The model includes 1-dimensional Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM), where CNN models the spatial representation of the original speech signal and LSTM obtained

short-term and long-term feature representation from Fbank feature. Cai *et al.* [8] designed a time-domain channel attention network, using dilated convolution to extract features from the raw waveform and an efficient channel attention module to capture information more relevant to depression cues. Huang *et al.* [9] applied CNN and domain adaptation technology to enhance cross-corpus generalizability, and Lopez *et al.* [10] proposed Deep neural network based structure with 23 Mel-Frequency Cepstral Coefficients (MFCC) and 40 filter-banks for discriminating the degrees of depression. Previous studies have demonstrated that it is feasible to detect depression by speech [11, 12].

At present, one challenge facing depression detection by speech is to generalize from lab-based research to clinical practice. Although there are several audio depression datasets publicly available for research, such as AVEC2013 [13], DAIC [14] and MOMDA [15], and most of the research work is carried out in those datasets, the process of data collection is completely different from the real clinical diagnosis. The participants in those corpora were asked to perform various tasks, such as reading speeches, telling stories and communicating with a virtual interviewer. In fact, HAMD is most commonly used for rating depression severity [16], it requires a 20-30 min face-to-face interview and clinician completes questionnaire according to the symptoms. It is considered that recording of HAMD interviews is more widely used in clinical diagnosis, and the automatic diagnosis method based on this recording can be easily extended to clinical application.

In this paper, we collected the recordings of clinical HAMD interviews from Beijing Huilongguan Hospital. To the best of our knowledge, this is the largest Chinese speech depression corpus. Along with the corpus, we proposed a TDNN-based depression detection with MFCC as feature input to the model. Experimental results on clinical recordings demonstrate the effectiveness of our proposed system. In addition, we found that there were significant differences in the expression of depressed patients through the analysis of transcriptions.

2. Depression corpus

Data collection is both an essential and challenging constituent of depression research. Many studies are conducted on the AVEC2013 as well as AVEC2014 datasets, and both datasets are partially selected from the audiovisual depressive corpus and applied to the Audio/Visual Emotion Challenge. The AVEC2013 dataset includes 340 videos generated by 292 German-speaking participants performing human-computer interaction tasks, including reading speech, counting down, telling stories, etc. The participants performed human-computer interaction tasks and wore headsets connected to lap-

*Corresponding author

¹Code is available at <https://github.com/dong-8080/ETDNN>

tops to collect speech data. They had to complete complicated tasks according to the prompts of the computer screen. Only 150 videos are available for public use. And the AVEC2014 corpus is just a partial interception of AVEC2013. DAIC contains four types of interviews, including face-to-face, teleconference, semi-automated and fully automated virtual interviewer. Besides, the vast majority of the audio depression dataset was collected in the European and US regions, with the exception of MODMA database [17]. MODMA is a multi-modal open dataset for mental disorder analysis in China, containing EEG and audio data, with the audio data recorded with microphones while 52 participants were interviewed, reading stories and viewing pictures. This dataset is more commonly used in EEG studies.

In our corpus, physicians organizes face-to-face interviews with participants in a quiet room, asking questions based on HAMD and subjects' own situations, and frequently asked questions are listed in Table 1. It is considered that face-to-face interviews could be a more natural way to stimulate the depression cues of patients.

Table 1: *Questions often asked by doctors in interviews.*

| Questions |
|-------------------------------------------------------------|
| Do you have insomnia recently? |
| Have you ever thought about suicide? |
| Do you often have feelings of nervousness or scared? |
| Have you ever suspected that you have an incurable disease? |
| How about your appetite in the last two weeks? |
| Do you have high mood swings from morning to night? |
| Do you have any plans for the future? |

The participants are recruited from Huilongguan Hospital from December 2020 to December 2021, including 36 male and 95 female subjects. The age of subjects ranges from 18 to 55 years with an average of 30.4 years. Every participant is a fluent Mandarin speaker without significant accent and the interviews were conducted in Mandarin. Furthermore, HAMD is conducted to identify the depression status, and those with HAMD scores higher than eight are defined as having depressive symptoms. In the patient group, the average HAMD score is 27, and the number of patients with mild, moderate and severe depression are 8, 47 and 11, respectively. A summary of our corpus details can be found in Table 2.

Table 2: *Summary of our corpus. N , m and f refer to the number of subjects, male subjects and female subjects, respectively.*

| Corpus | N (m/f) | Age | Duration (min) |
|----------------|------------|----------------|----------------|
| Depression | 66 (18/48) | 30.0 \pm 8.4 | 17.0 \pm 5.4 |
| Not Depression | 65 (18/47) | 31.0 \pm 8.6 | 11.0 \pm 1.5 |

Collecting clean and high-quality data is one of the most challenging parts of depression research. The original audio is recorded using a dual channel recording device manufactured by asoundgen. Voices of subjects and physicians are separated into two channels at the sample rate of 16kHz and saved in WAV format. To protect the privacy of the participants, sensitive personal information is cut before further analysis. In total, we obtained 66 voices from depressed patients and 65 voices from normal controls with an average duration of 13.9 minutes.

3. Methods

3.1. Preprocessing

Firstly, the mono audios of the subject's speech were separated from the stereo audios. In order to differentiate the speech from non-speech sections, we used Voice Activity Detection (VAD) based on energy and zero-crossing rate [18]. In our experiment, we divided each audio into several segments with a duration of three seconds and overlap is 50%, which is the same as in the previous study [8].

3.2. Feature extraction

MFCC is applied as a classification feature because it provides stable and reliable partitioning of acoustic space and concentrates with the levels of speaker depression, and is utilized in most studies on emotion and depression detection. We extract 80-dimensional MFCC with a window size of 25 ms and a frame shift of 10 ms as neural network inputs. The feature extraction process is implemented by torchaudioⁱⁱ.

3.3. Data augmentation

Data augmentation techniques are widely utilized to introduce new data samples and improve the robustness of the deep learning models. In our experiment, we use SpecAugment [19] for data augmentation that operates on MFCC of the input audio rather than the raw audio itself. This method is simple to apply and does not require additional data, consisting of frequency masking and time masking. SpecAugment is used so that f consecutive mel frequency channels and t consecutive time steps are masked, where f and t are randomly chosen from [0, 8] and [0, 10], respectively. Data Augmentation is applied on-the-fly to every speech segment.

3.4. Baseline

Our baseline system is Support Vector Machine (SVM) with radial basis function kernel combined with AVEC2013 feature set [13], which is commonly used in previous depression studies. AVEC2013 feature set contains several low-level descriptors and functions that represent parameters of the audio signal, such as energy, pitch and spectral. We extract those features using openSMILE toolkit [20]. A one-way ANOVA was applied to select the distinguishing feature subset and the hyperparameters of SVM were optimized by GridSearch using the scikit-learn [21].

3.5. Proposed classification system

At present, the most widely used neural network topology in speech classification is based on Time Delay Neural Network (TDNN) architectures, also as known as x-vector [22, 23]. We further propose an Emphasized Channel Attention, Propagation and Aggregation (ECAPA) TDNN based system for depression detection, as is shown in Figure 1.

The basic x-vector is composed of TDNN layer, statistical pooling layer and fully connected linear layer. The first five TDNN layers of the network operate on speech frame level with different temporal contexts, and statistic pooling layer aggregates frame-level outputs from the last TDNN layer and derives its mean and standard deviation, then these segment-level statistics are concatenated and fed into two fully connected layers [24].

ⁱⁱ<https://pytorch.org/audio/stable/index.html>

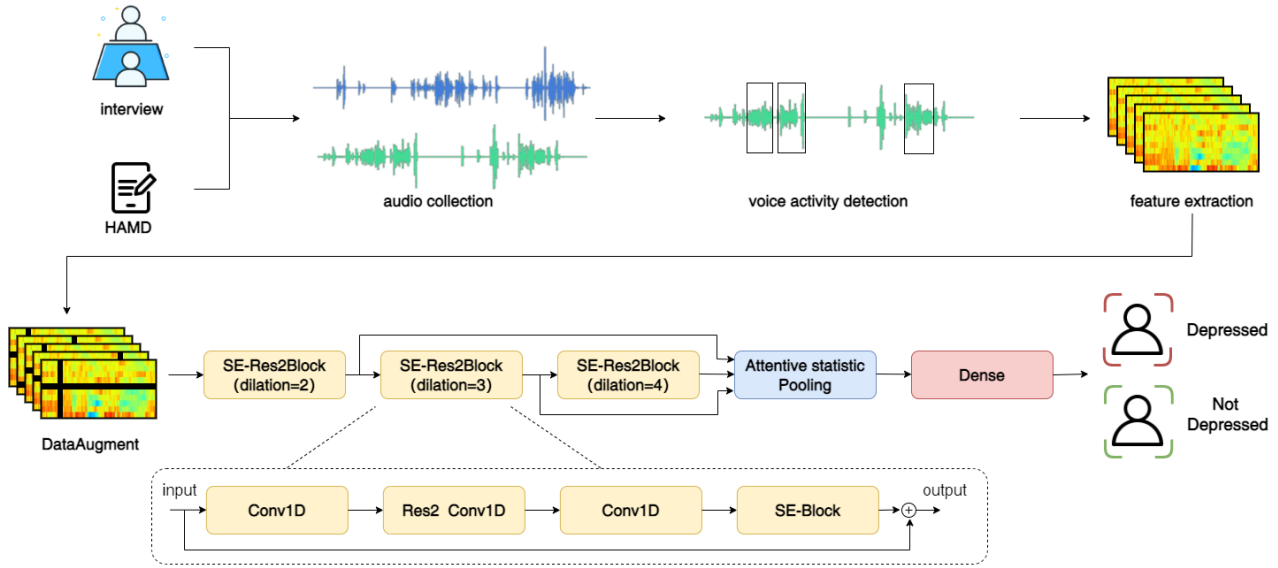


Figure 1: The framework for ECAPA-TDNN based depression detection system.

The ECAPA-TDNN [25] is a variant that has made multiple enhancements on x-vector architecture. It replaces the original five TDNN layers with three 1-Dimensional SE-Res2Blocks, in which Res2Net module implements convolution operations and (Squeeze and Excitation) SE block performing dynamic channel-wise feature recalibration to improve the representation ability of the network. The Res2Net model replaces a group of filters with smaller groups of filters and then connects different filter groups with a hierarchical residual-like way, which can process multi-scale features and reduce the number of parameters significantly. SE block has the greatest influence on the experimental results of the model, it first produces channel descriptors z by the mean of features across the time domain, then calculating a weight for every channel:

$$s = \sigma(W_2 f(W_1 z + b_1) + b_2) \quad (1)$$

with $\sigma(\cdot)$ and $f(\cdot)$ representing sigmoid and tanh functions, respectively. Finally, the origin input features are reweighted by multiplying s and fed into the next layer [20]. The dilation of these blocks is 2, 3, and 4, respectively. The output of each block is concatenated and fed into the pooling layer so that both shallow and deep features which express different pieces of information are effectively utilized.

Attention statistics pooling layer produces weighted means as well as standard deviations of frame-level features by attention mechanism [26]. The statistics pooling can calculate the mean and standard deviation of the input frame-level feature, which converts features of the side length into a fixed-dimensional vector. Generally speaking, frame-level features of some frames are more important than others, so it makes sense to introduce attention mechanism for extracting more important frames. At first, a scalar score e_t is calculated for each frame-level feature:

$$e_t = v^T f(Wh_t + b) + k \quad (2)$$

where h_t represents frame-level features from the last layer.

Then e_t is normalized over all frames by softmax:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{\tau} \exp(e_{\tau})} \quad (3)$$

And the weighted mean vector and standard deviation are defined as follows:

$$\mu = \sum_t \alpha_t h_t \quad (4)$$

$$\sigma = \sqrt{\sum_t \alpha_t h_t^2 - \mu^2} \quad (5)$$

This enables speakers embedding to more accurately and efficiently capture speaker factors with respect to long-term variations.

Finally, the utterance-level statistical features are calculated at two fully connected layers and then passed to the softmax layer to get the final classification results.

3.6. Speech transcription

In order to explore the differences in verbal expression between depressed and healthy subjects, we further analyze the word frequency in the corpus. The recordings of speech were transcribed using WeNet [27]. This end-to-end speech recognition model adopts the hybrid CTC/attention architecture with conformer blocks that achieves state-of-the-art performance on Chinese speech corpus Aishell-1 [28].

Word segmentation and Part-of-speech (POS) were performed by Jiebaⁱⁱⁱ, a simple and effective tool for Chinese texts. Six types of POS were labeled for future analysis, i.e., nouns, verbs, adjectives, adverbs, pronouns, and prepositions. We calculated the proportion of POS in depressed and healthy groups and compared their differences.

ⁱⁱⁱ<https://github.com/tachikomahub/jieba>

4. Experiment Result

4.1. Data preparation

After preprocessing, there were a total of 14433 audio segments with a duration of three seconds, including 8681 depressive speech and 5752 healthy speech. Then we extracted MFCC from the speech segments, applying online frequency masking and time masking before feeding into the neural network. For dividing training set and testing set, group k-fold cross-validation is used to ensure that the voice generated by the same speaker does not appear in both the training and evaluation process at the same time. We set $k=5$ in our experiment.

4.2. Model training

The ECAPA-TDNN is implemented under the Deep Learning framework of Pytorch [29]. The Adam optimizer is used as an optimization algorithm, and the loss function is cross-entropy loss function. The cosine decay with warm-up strategy is utilized for adjusting the learning rate. That is, the learning rate is initially set to 0.00001, increases to 0.0001 after 20 epochs, and gradually declines to 0.00001 in a cosine-like schedule in the next 180 epochs. The iteration of training stop at epoch 200. The experiments are repeated five times. For SVM, different combinations of cost of 0.001, 0.01, 0.1 and 10 and gamma of 0.00001, 0.0001, 0.001 were evaluated. The 2286 features extracted by openSMILE were standardized and then selected from 100 to 2286 with steps of 100 via ANOVA as feature selection methods.

4.3. Classification result

To observe the overall performance of the proposed, Accuracy, Precision, Recall and F1 score are taken to evaluate the utterance and speaker level result. The utterance-level performances of the proposed depression detection system and baseline are listed in Table 3. The proposed approach achieves an average accuracy of 83.4%, a precision of 86.8%, a recall of 85.1%, and an F1 score of 86.5%, which is obviously better than the performance of baseline. It has to be noticed that we used the accuracy of utterance-level results for model training and evaluation, and the experiments were repeated five times.

Table 3: Classification result in utterance level (mean and standard deviation).

| | Accuracy | Precision | Recall | F1 score |
|-----------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Baseline | 0.712 (± 0.043) | 0.750 (± 0.022) | 0.798 (± 0.077) | 0.772 (± 0.048) |
| Proposed | 0.834 (± 0.001) | 0.868 (± 0.004) | 0.851 (± 0.008) | 0.865 (± 0.006) |

Since the purpose of classifying depression is to determine whether an individual has depressive symptoms, we aggregated the prediction results at the utterance level to speaker level. After majority voting, the prediction results of 14433 short speech were converged into the prediction results of 131 speakers, and better classification results were achieved, as shown in Table 4. The accuracy, precision, recall and F1 score are 90.4%, 88.8%, 92.8% and 90.8%, respectively.

4.4. Transcriptions analysis

By analyzing the transcriptions of each utterance, we found there was significant difference between the normal and de-

Table 4: Classification result in speaker level (mean and standard deviation).

| | Accuracy | Precision | Recall | F1 score |
|-----------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Baseline | 0.758 (± 0.056) | 0.719 (± 0.052) | 0.868 (± 0.056) | 0.785 (± 0.041) |
| Proposed | 0.904 (± 0.019) | 0.888 (± 0.018) | 0.928 (± 0.019) | 0.908 (± 0.018) |

pressed groups. Significance testing was performed using Student t-test and the significance level was set to 0.001. The average frequency of the six lexical categories is represented in Figure 2. P values of the word frequency in nouns, pronouns and adjectives are less than significance level, which indicates that patients with depression have the tendency to use more nouns and adjectives during interviews, while the use of prepositions and pronouns is less.

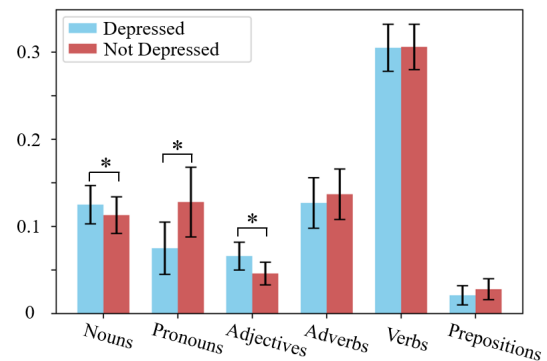


Figure 2: Word frequency statistics between depressed and not depressed groups (* represents $p < 0.001$).

5. Conclusions

Although progress has been made to automatic detection of depression from speech, experimental verification on the clinical data still needs to be carried out before practical applications. In this paper, we collected the clinical recording of HAMD interview and established a relatively large Chinese depression corpus. Besides, a TDNN-based depression detection model was proposed and achieved high performance in diagnosing depression, in which the clinical speeches were cut into three seconds after VAD, and MFCC features were extracted from them as the input of the model. This study performed a solid foundation for identifying depression via clinical speech and shined a light on reducing the gap between research and clinical practice. In the future work, we will continue to expand our corpus as well as track the depression changes overtime through longitudinal data, and deploy our system to assist clinicians in the diagnosis of depression.

6. Acknowledgements

This work is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (202200042), New Talent Project of Beijing University of Posts and Telecommunications (2021RC37) and the Natural Science Foundation of Shandong Province (ZR2020MF051).

7. References

- [1] M. Wang, S. Yan, Y. Zhou, and P. Xie, “trans-cinnamaldehyde reverses depressive-like behaviors in chronic unpredictable mild stress rats by inhibiting nf- κ b/nlrp3 inflammasome pathway,” *Evidence-Based Complementary and Alternative Medicine*, vol. 2020, 2020.
- [2] S. M. Meier, L. Petersen, M. Mattheisen, O. Mors, P. B. Mortensen, and T. M. Laursen, “Secondary depression in severe anxiety disorders: a population-based cohort study in denmark,” *The Lancet Psychiatry*, vol. 2, no. 6, pp. 515–523, 2015.
- [3] K. R. R. Krishnan, M. DeLong, H. Kraemer, R. Carney, D. Spiegel, C. Gordon, W. McDonald, M. A. Dew, G. Alexopoulos, K. Buckwalter *et al.*, “Comorbidity of depression with other medical diseases in the elderly,” *Biological psychiatry*, vol. 52, no. 6, pp. 559–588, 2002.
- [4] C. Demiroglu, A. Beşirli, Y. Ozkanca, and S. Çelik, “Depression-level assessment from multi-lingual conversational speech data using acoustic and text features,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–17, 2020.
- [5] M. Alpert, E. R. Pouget, and R. R. Silva, “Reflections of depression in acoustic measures of the patient’s speech,” *Journal of affective disorders*, vol. 66, no. 1, pp. 59–69, 2001.
- [6] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, “Voice acoustical measurement of the severity of major depression,” *Brain and cognition*, vol. 56, no. 1, pp. 30–35, 2004.
- [7] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, “Depaudionet: An efficient deep model for audio based depression classification,” in *Proceedings of the 6th international workshop on audiovisual emotion challenge*, 2016, pp. 35–42.
- [8] C. Cai, M. Niu, B. Liu, J. Tao, and X. Liu, “Tdca-net: time-domain channel attention network for depression detection,” in *Interspeech*, 2021, pp. 2511–2515.
- [9] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, “Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns,” in *INTERSPEECH*, 2020, pp. 4561–4565.
- [10] J. V. Egas-López, G. Kiss, D. Sztahó, and G. Gosztolya, “Automatic assessment of the degree of clinical depression from speech using x-vectors,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8502–8506.
- [11] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, R. Cowie, and M. Pantic, “Summary for avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1483–1484.
- [12] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, and M. Pantic, “Summary for avec 2017: Real-life depression and affect challenge and workshop,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1963–1964.
- [13] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 3–10.
- [14] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, “The distress analysis interview corpus of human and computer interviews,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, 2014, pp. 3123–3128.
- [15] H. Cai, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao *et al.*, “Modma dataset: a multi-modal open dataset for mental-disorder analysis,” *arXiv preprint arXiv:2002.09283*, 2020.
- [16] T. A. Furukawa, M. Reijnders, S. Kishimoto, M. Sakata, R. J. DeRubeis, S. Dimidjian, D. J. Dozois, U. Hegerl, S. D. Hollon, R. B. Jarrett *et al.*, “Translating the bdi and bdi-ii into the hamd and vice versa with equipercetile linking,” *Epidemiology and psychiatric sciences*, vol. 29, 2020.
- [17] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang *et al.*, “Deep learning for depression recognition with audiovisual cues: A review,” *Information Fusion*, vol. 80, pp. 56–86, 2022.
- [18] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.
- [19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [20] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
- [21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [22] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *Interspeech*, vol. 2017, 2017, pp. 999–1003.
- [23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [24] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [25] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” in *Interspeech2020*, 2020, pp. 3830–3834.
- [26] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” in *Conference on Computer Vision*. Springer, 2006, pp. 531–542.
- [27] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, “Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” *arXiv preprint arXiv:2102.01547*, 2021.
- [28] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, 2017, pp. 1–5.
- [29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.