Computer Life
ISSN: 1819-4818 | Vol. 12, No. 2, 2024

Depression detection based on speaker information disentanglement using pre-trained features

Shugang Liu, Yuhang Zhao

School Of Computer Science, NORTH CHINA ELECTRIC POWER University, Baoding 071003, China

Abstract: This article proposes using pre trained features to address the challenge of detecting depression through speech. Traditional raw audio has shown low accuracy and insufficient generalization performance in depression detection. We use pre trained models that have been developed to extract features, which can be used to extract general feature representations from speech data. During the pre training process, we further decouple the speakers, introducing prior information and providing a better starting point for training downstream models. The results indicate that we achieved the best performance when using the extracted features from the CONTANTVEC pre- trained model with speaker decoupling improvement.

Keywords: Pre-trained, depression detection, speech classification, speaker information.

1. Introduction

As one of the global public health issues in the 21st century, depression has become the biggest killer of mental illness due to its high incidence, disability, recurrence, missed diagnosis, and suicide rates. There is a huge gap in the prevalence of diagnosis and treatment for depression. There is an urgent need to develop a low-cost, efficient, and universal automatic detection method for depression.

The latest advances in speech recognition technology have shown enormous potential in addressing the challenges posed by this severe disease. Multiple feature and model architectures for MDD diagnosis have been proposed in the past [3,4,5], each with its unique advantages and limitations. This includes various acoustic features [6,7,8,9,10], as well as complex backend modeling techniques [11,12,13]. Among them, features related to speaker identity have also been used for depression detection [14,15], and these studies mainly focus on x vectors or speaker embeddings.

In recent studies, it has been pointed out that in small datasets, models are prone to overfitting to speaker classification models during the process of learning depression classification. If this situation can be avoided, the performance of the model can be improved to a certain extent.

[16] Adversarial learning is used to remove speaker related information from speech signals for speech emotion recognition and depression detection [17]. An unsupervised speaker disentanglement method is used for speech depression detection. Although they achieved good results, their features all achieved good results on the original audio. In the current stage of vigorous development of large models, in many speech processing tasks, including Automatic Speech Recognition (ASR) and Automatic Emotion Recognition (AER) [18,19], state-of-the-art (SOTA) results have been achieved through the use of fine-tuning speech universal pre trained models. Can we use some pre trained models that have been used for speaker entanglement resolution to extract pre trained features and achieve better results? This is the research we are going to conduct in this paper.

This article investigates the challenge of using pre trained speech based models to address (speech depression detection) in SDD. By using pre trained models, universal feature representations can be extracted from speech data, introducing prior information and providing a better starting point for downstream model training. We first compared the advantages and disadvantages of pre trained features and original audio features through experiments. We used three pre trained models: Hubert base [20], CONTANTVEC [21], and FACodec [22]. It was found that the use of an improved self supervised pre training model, CONTANTVEC, based on speaker separation, performed the best. We also investigated the impact of using different downstream models on the experimental results and found that ECAPA-TDNN [23] achieved better results compared to linear layers. Then we analyze each block by block to explore the representations of different layers of CONTANTVEC and understand which type of information is more effective in SDD.

Our contributions include:

1. Explore the advantages and disadvantages of using pre trained features compared to raw audio for depression detection.

2. Explore the performance of using pre trained features for speaker entanglement resolution in depression detection.

3. Explore the superiority and inferiority of using different output layers of CONTANTVEC in depression detection results.

2. Method

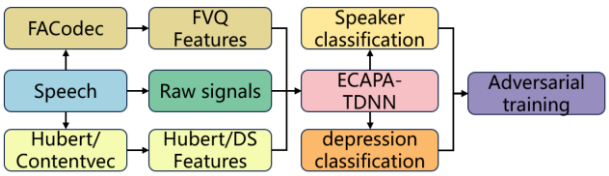

Fig. 1 The general flow diagram of proposed methods.（FVQ：Factorial Vector Quantization, characterized by content, prosody and acoustic details;DS: Disentangling Speakers）

2.1. Raw signals based approach

As shown in Figure 1, we used Raw signals as the baseline to investigate whether the depression detection performance of using pre trained features can surpass that of Raw signals features.

2.2. Features of self supervised pre-trained models based approach

Here we explore two pre trained models, namely Hubert base and Contentvec.

Among them, HuBERT base is the basic version of HuBERT and one of the most commonly used voice SSL models. It adopts a multi-layer Transformer structure. This includes self attention mechanisms and feedforward neural networks. This Transformer structure is used to extract speech features from the original audio signal and model contextual information. The Hubert Base model learns universal audio representations by pre training on large-scale speech data. Introducing prior information to give features a higher starting point.

CONTANTVEC is built on the basis of the HuBERT framework and introduces three key disentanglement components: teacher's disentanglement, student's disentanglement, and teacher's adjustment of the predictor. By focusing on the representation of audio content, speaker decoupling can be achieved, enabling speaker disengagement without significant loss of content.

2.3. FACodec vector decomposition features based approach

FACodec is the latest development by Microsoft for the Text to Speech (TTS) model, which can decouple audio into Prosody, Content, Acoustic Detail, and Timber. Among them, Timber is an important information for speaker recognition, but it does not play a significant role in depression detection. Therefore, we can decouple Timbre and leave Prosody, Content, and Acoustic Detail information for depression detection.

3. Experiments

3.1. Datasets

We have chosen the most widely used speech depression detection dataset, DAIC-WOZ [24]. It collected English audiovisual interviews with 189 males and females who underwent psychological distress assessment. This dataset contains 107 speakers for training and 35 speakers for evaluation, consistent with the database description. Extract audio data from patients only using the provided timestamps.

3.2. Data Pre-processing

We have taken preprocessing measures to address the issue of data imbalance. Firstly, we perform random cropping and sampling preprocessing on the training data. This means that we randomly crop each utterance to the shortest utterance length and divide it into multiple segments with a duration of 3.84 seconds. Each segment contains 61440 original audio samples. To ensure the balance of the training set, we adopted a random sampling method to extract the same number of depressive and non depressive fragments without replacement. This can ensure that the sample size for each category is the same. In each experiment, we trained five models using randomly generated training subsets. Finally, we averaged the prediction results of these five models to obtain the final prediction result. The batch size we use is 20.

3.3. Input Features

We evaluated four features, namely raw audio features, HuBERT features, CONTANTVEC features, and Prosody, Content, and Acoustic Detail fusion features decoupled by FACodec. The original audio features are 1-dimensional, HuBERT and CONTANTVEC are 768-dimensional, and FACodec features are 256 dimensional.

3.4. Models

For downstream models, we use ECAPA-TDNN, which receives an audio input and outputs speaker classification results and depression detection results, respectively. In order to ensure operation, we have made changes to the input dimension of ECAPA-TDNN to adapt to the dimensions of different features.

4. Result and analysis

In the experimental results, we will demonstrate the performance comparison between the original audio and pre trained features with and without speaker entanglement. And compare the performance of different output layers of CONTANTVEC.

4.1. Result

The results are shown in Table 1, where we use the original audio results as the baseline and the results of pre trained features as our proposed method. The speaker level F1 scores of depression (F1-D), non depression (F1-ND) categories and their unweighted (macro) mean (F1-AVG) were compared.

Table 1 The experimental results of using pre trained models and original audio with and without speaker decoupling are presented, respectively, on the validation and test sets(ADV: Whether to conduct speaker adversarial training)

Input Features	ADV	Valid F1-ND	Valid F1-D	Valid F1-AVG	Test F1-ND	Test F1-D	Test F1-AVG
Raw	No	0.727	0.549	0.638	0.6546	0.2495	0.4521
	Yes	0.757	0.551	0.654	0.6767	0.2533	0.4650
HuBERT	NO	0.708	0.442	0.575	0.6642	0.2982	0.4812
	Yes	0.721	0.435	0.578	0.6878	0.2825	0.4852
CONTENTVEC	NO	0.743	0.575	0.659	0.6861	0.3266	0.5064
	Yes	0.780	0.574	0.667	0.7530	0.3158	0.5433
FAcodec	No	0.766	0.522	0.644	0.7430	0.2500	0.4965
	Yes	0.750	0.455	0.602	0.7160	0.2960	0.5060

Then we conducted experiments on different output layers of CONTANTVEC, all of which incorporated speaker adversarial training. Similarly, speaker level F1 scores were compared for categories of depression (F1-D), non depression (F1-ND), and their unweighted (macro) mean (F1-AVG). However, since our goal was to find the optimal output layer and did not consider the performance of model generalization, we only compared the results on Valid. The variation curve of the result with the output layer is shown in Figure 2.

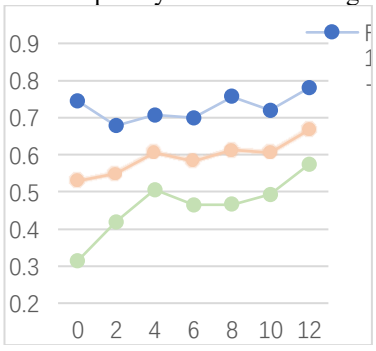

Fig. 2 Change the result curve of different output layers of CONTANTVEC

4.2. Analysis

In Table 1, we can observe that after introducing speaker adversarial training, except for FACodec, all other results show an improvement compared to not introducing speaker adversarial training. And using the CONTANTVEC feature with speaker decoupling improvement based on HuBERT, there is also a significant improvement compared to HuBERT feature, which verifies the conclusion that it is easy to overfit the previous DAIC-WOZ dataset for speaker classification. And we found that using raw audio features performs well on the validation set, but differs significantly on the test set, resulting in poor generalization performance. After using pre trained features, whether it is HuBERT features, CONTANTVEC features, or FACodec features, our performance on the test set has improved compared to raw audio features. Among them, CONTANTVEC achieved the best performance on the validation set, with F1-AVG of 0.667, and also achieved the best results on the test set with F1-AVG of 0.5433 and F1-ND of 0.7530.

From this, we can conclude that compared to using the original audio features, our use of speaker decoupled CONTANTVEC features outperforms the original audio features in terms of performance on the validation set and model generalization. And after comparing different output layers, the data achieved the best performance in the last layer of the model after passing through 12 layers of transformers.

5. Conclusion

Our proposed method of using pre trained features for depression detection has shown satisfactory results. The CONTANTVEC feature achieved the best results, with a validation set of F1-AVG of 0.667 and a training set of F1-AVG of 0.5433 and F1-ND of 0.7530. Compared to the original audio, there is not only an improvement in accuracy on the validation set, but also a significant improvement in generalization performance. However, during the experiment, we found that due to the uneven distribution of the DAIC-WOZ dataset, the F1-D score was relatively low. Future work can focus on addressing dataset imbalance and exploring more fine-grained variants of pre trained models.

References

[1] C. D. Mathers, D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030,PLoS Med. 3 (2006) e442.

[2] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder,J. Epps, T. F. Quatieri, A review of depression and suicide risk assessment using speech analysis,Speech Commun. 71 (2015) 10–49.

[3] E. Rejaibi et al., “Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech,” Biomedical Signal Processing and Control, vol. 71, p.103107, 2022.

[4] Y . Shen et al., “Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model,” in ICASSP.IEEE, 2022, pp. 6247–6251.

[5] K. Chlasta et al., “Automated speech-based screening of depression using deep convolutional neural networks,” Procedia Computer Science, vol. 164, pp. 618–628, 2019.

[6] M. H. Sanchez et al., “Using prosodic and spectral features in detecting depression in elderly males,” in Interspeech, 2011, pp.3001–3004.

[7] S. P . Dubagunta et al., “Learning voice source related information for depression detection,” in ICASSP. IEEE, 2019, pp. 6525–6529.

[8] Y . Yang et al., “Detecting depression severity from vocal prosody,” IEEE transactions on affective computing, vol. 4, no. 2,pp. 142–150, 2012.

[9] A. Afshan et al., “Effectiveness of voice quality features in detecting depression,” Interspeech, 2018.

[10] N. Seneviratne and C. Espy-Wilson, “Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings,” in ICASSP. IEEE, 2022, pp.6252–6256.

[11] L. Yang et al., “Feature augmenting networks for improving depression severity estimation from speech signals,” IEEE Access,vol. 8, pp. 24 033–24 045, 2020.

[12] A. Vázquez-Romero et al., “Automatic detection of depression in speech using ensemble convolutional neural networks,” Entropy,vol. 22, no. 6, p. 688, 2020.

[13] A. Harati et al., “Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus,” in ICASSP. IEEE, 2021, pp. 7273–7277.

[14] J. V . Egas-López et al., “Automatic assessment of the degree of clinical depression from speech using x-vectors,” in ICASSP.IEEE, 2022, pp. 8502–8506.

[15] V . Ravi et al., “Fraug: A frame rate based data augmentation method for depression detection from speech signals,” in ICASSP.IEEE, 2022, pp. 6267–6271.

[16] Wang J, Ravi V, Alwan A. Non-uniform speaker disentanglement for depression detection from raw speech signals[C]//Interspeech. NIH Public Access, 2023, 2023: 2343.

[17] Ravi V, Wang J, Flint J, et al. A Privacy-Preserving Unsupervised Speaker Disentanglement Method for Depression Detection from Speech[C]//CEUR workshop proceedings. NIH Public Access, 2024, 3649: 57.

[18] Y u Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Y uanzhong Xu, Yanping Huang,Shibo Wang, et al., “BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition,” IEEE Journal of Selected Topics in Signal Processing,vol. 16, no. 6, pp. 1519–1532, 2022.

[19] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat,Matheus Damasceno, and Hagai Aronowitz, “Speech emotion recognition using self-supervised features,” in Proc. ICASSP,Singapore, 2022.

[20] Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 3451-3460.

[21] Qian K, Zhang Y, Gao H, et al. Contentvec: An improved self-supervised speech representation by disentangling speakers[C]//International Conference on Machine Learning. PMLR, 2022: 18003-18017.

[22] Ju Z, Wang Y, Shen K, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models[J]. arXiv preprint arXiv:2403.03100, 2024.

[23] Desplanques B, Thienpondt J, Demuynck K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification[J]. arXiv preprint arXiv:2005.07143, 2020.

[24] Burdisso S, Reyes-Ramírez E, Villatoro-Tello E, et al. DAIC-WOZ: On the Validity of Using the Therapist's prompts in Automatic Depression Detection from Clinical Interviews[J]. arXiv preprint arXiv:2404.14463, 2024.