Journal of Affective Disorders 323 (2023) 299–308

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Depression recognition using a proposed speech chain model fusing speech production and perception features

Minghao Du [a,1], Shuang Liu [a,*,1], Tao Wang [a], Wenquan Zhang [a], Yufeng Ke [a], Long Chen [a], Dong Ming [a,b,**]

[a] *Tianjin International Joint Research Center for Neural Engineering, Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China*
[b] *Lab of Neural Engineering & Rehabilitation, Department of Biomedical Engineering, College of Precision Instruments and Optoelectronics Engineering, Tianjin University, Tianjin, China*

A R T I C L E I N F O

Keywords:
Depression
Deep learning
Audio
Feature fusion
Auxiliary diagnosis

A B S T R A C T

Background: Increasing depression patients puts great pressure on clinical diagnosis. Audio-based diagnosis is a helpful auxiliary tool for early mass screening. However, current methods consider only speech perception features, ignoring patients' vocal tract changes, which may partly result in the poor recognition.
Methods: This work proposes a novel machine speech chain model for depression recognition (MSCDR) that can capture text-independent depressive speech representation from the speaker's mouth to the listener's ear to improve recognition performance. In the proposed MSCDR, linear predictive coding (LPC) and Mel-frequency cepstral coefficients (MFCC) features are extracted to describe the processes of speech generation and of speech perception, respectively. Then, a one-dimensional convolutional neural network and a long short-term memory network sequentially capture intra- and inter-segment dynamic depressive features for classification.
Results: We tested the MSCDR on two public datasets with different languages and paradigms, namely, the Distress Analysis Interview Corpus-Wizard of Oz and the Multi-modal Open Dataset for Mental-disorder Analysis. The accuracy of the MSCDR on the two datasets was 0.77 and 0.86, and the average F1 score was 0.75 and 0.86, which were better than the other existing methods. This improvement reveals the complementarity of speech production and perception features in carrying depressive information.
Limitations: The sample size was relatively small, which may limit the application in clinical translation to some extent.
Conclusion: This experiment proves the good generalization ability and superiority of the proposed MSCDR and suggests that the vocal tract changes in patients with depression deserve attention for audio-based depression diagnosis.

1. Introduction

Depression is a common but serious psychological disorder characterized by persistent pessimism, cognitive decline, and social dysfunction (Hammar et al., 2022). To prevent depression scientifically, timely diagnosis is necessary to ensure appropriate treatment (Costantini et al., 2021). The World Health Organization estimates that 322 million people currently suffer from depression (Organization, 2017), which severely increases the burden of diagnosis. Therefore, automatic methods are needed to improve diagnostic capabilities. Although electroencephalogram-based (Saeedi et al., 2021), heart rate-based (Hartmann et al., 2019), and blood-based (Sealock et al., 2021) methods have shown good performance in depression diagnosis due to the objectivity of physiological signals, the high cost of the equipment and cumbersome collection processes make it difficult to popularize them. In contrast, audio-based depression diagnosis is more suitable for early mass screening. This method captures paralinguistic differences to diagnose depression, such as prosody and speech quality. Not focusing

* Corresponding author.
** Correspondence to: D. Ming, Tianjin International Joint Research Center for Neural Engineering, Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China.
E-mail addresses: shuangliu@tju.edu.cn (S. Liu), richardming@tju.edu.cn (D. Ming).
[1] These authors contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.jad.2022.11.060
Received 15 July 2022; Received in revised form 22 October 2022; Accepted 20 November 2022
Available online 30 November 2022
0165-0327/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

on the conscious and subjective semantic information, paralinguistics is an unconscious human communication phenomenon, containing rich content of attitude, themes, and emotions (Madhavi et al., 2020). Associated with neuromotor systems and physiological states, paralinguistic information in speech is highly sensitive to the effects of neurodegenerative illnesses (Gómez-Rodellar et al., 2020), which could serve as the objective depression-related marker. In addition, audio collection only requires a microphone, and the process could be contactless, both of which increase accessibility. Importantly, subject privacy could be protected due to the openness of audio. Therefore, audio-based depression diagnosis could become a complementary method to improve current diagnostic capabilities.

Phonetic differences in patients with depression have been confirmed by previous research. Patients usually present the clinical phonetic representation of speaking less, in low volume, and with hesitation (Sahu and Espy-Wilson, 2014). After further quantification, a significant difference in amplitude-frequency characteristics was found between depressed and non-depressed groups with respect to the glottis, fundamental frequency, jitter, and shimmer (Jia et al., 2019; Silva et al., 2021; Simantiraki et al., 2017). For explaining this phenomenon, some studies hypothesized that the neuromuscular coordination of patients is impaired by cognitive decline, and they further speculated that the vocal tract is affected by depression (Espy-Wilson et al., 2019; Seneviratne et al., 2020). Based on this research, the feasibility of implementing phonetic features as diagnostic clues of depression can be considered as confirmed. The goal of audio-based depression diagnosis is to identify depression by pronunciation features, regardless of language, content, or habits of speech. To achieve this, recent effort has mainly involved two aspects: phonetic feature extraction and optimization modeling.

For phonetic features, considering the perception differences in depressive speech, handcrafted descriptors, such as speed, prosodic features, and spectral features have been widely used. With respect to emotional perception, short time energy, intensity, loudness, and zero-crossing rate were extracted as handcrafted descriptors, and they showed robustness in the depression classification task (Long et al., 2017). With respect to tonal perception, Lam-Cassettari and Kohlhoff (2020) and Patil and Wadhai (2021) analyzed the difference of pitch between depressed and non-depressed groups and demonstrated the feasibility of pitch serving as a classification marker. With respect to auditory perception, the second dimension of the Mel-frequency cepstral coefficients (MFCC-2) of depressed patients was significantly higher than that of non-depressed subjects, which reflected an energy difference of frequencies around 2000–3000 Hz (Taguchi et al., 2018). Based on these differences, the MFCC, Mel-spectrogram, and spectrogram, which reflect time-frequency information, have been used in depression diagnosis and have shown positive performance (He et al., 2022; Rejaibi et al., 2022; Vázquez-Romero and Gallardo-Antolín, 2020; Yadav and Sharma, 2021). However, these features are extracted only from the speech perception process based on the sensory differences in how the depressive speech sounded rather than how it is produced. As speech difference maybe originate from changes in the vocal tract, extracting features only from the process of speech perception will lead to information loss according to the speech chain (Denes et al., 1993; Tjandra et al., 2020). Recently, a study regarding speaker identity recognition built a speech chain model that could capture phonetic identity features from the processes of speech production and speech perception (Chowdhury and Ross, 2020). This work used linear predictive coding (LPC) to model the vocal tract of the speaker and MFCC to describe the perceptual law of the human ear. The efficacy of the proposed model over existing methods was demonstrated, which may represent an advancement in depression diagnosis. Hence, we suppose that extracting phonetic features from both the processes of speech production and of speech perception can further improve depression recognition.

Models for automatic depression diagnosis have broadly employed two key approaches: traditional machine learning and neural networks. Representatives of traditional machine learning such as support vector machine (Dai et al., 2021; Valstar et al., 2016), linear regression (Jiang et al., 2018; Pan et al., 2018), and decision tree (Liu et al., 2020; Pampouchidou et al., 2016) are often selected for classification, but they have some limitations. As required for the input dimension of these models, statistical functions (mean, median, variance) of the phonetic features extracted from the whole speech are often used as inputs, which ignores the dynamic changes of the speech that are strongly associated with depression (Wichers, 2014). In contrast, neural networks are not limited by input dimensions and can extract dynamic information in the time or frequency domain. Srimadhur and Lalitha (2020) proposed an end-to-end convolutional neural network (CNN) framework to identify depression based on processed audio and achieved better classification results than the traditional machine learning models. Muzammel et al. (2020) divided the whole speech into segments and extracted spectral features, then established a phoneme-level CNN architecture to capture vowel and consonant acoustic features. This method provided excellent results on speech segments, but the whole speech was not tested. Zhao et al. (2021) focused on emotionally salient regions and proposed an attention-based long short-term memory network (LSTM) network to obtain key depression information in time information for classification. These studies have shown that neural networks are sensitive to dynamic information.

Another challenge of classification models is class imbalance, such as inconsistency in quantity and speech duration. Previous studies used random sampling (He and Cao, 2018; Zhao et al., 2021), resizing (Dong and Yang, 2021; Othmani et al., 2021), and cropping (Ma et al., 2016; Negi et al., 2018) of the whole speech to ensure non-bias of the models, but the depression-related information could have been lost. In other words, it is unreasonable to diagnose only by a few seconds in several minutes of speech, and the meaning of spectrograms would be changed after compression. Fortunately, (Rejaibi et al., 2022) proposed an ensemble system that divided speech into segments for detection with unit length and performed final classification by a hard voting classifier. However, depression is reflected not just in the classification proportion of segments, for example, non-depressed subjects also say negative things, and depressed subjects also show fewer expressions of positivity. Therefore, more complex relationships between segments need to be explored.

In this work, we propose a novel machine speech chain model for depression recognition (MSCDR). It has three main steps. First, raw speech is preprocessed to segments and then 40-dimensional LPC and 39-dimensional MFCC features are extracted to describe the processes of speech generation and of speech perception, respectively. Second, a one-dimensional convolutional neural network (1D-CNN) is proposed to extract intra-segment depressive features, which is composed of two networks processing in parallel with LPC and MFCC features as the inputs. Finally, a feature-level fusion algorithm is used to conduct the fusion of temporal features, and an LSTM is proposed to capture inter-segment depressive correlation features for classification. We employed the proposed MSCDR on the English dataset Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) and Chinese dataset Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) and compared the classification results with existing methods to demonstrate the superiority and generalization of the MSCDR.

The contributions of this work can be summarized as follows:

- Based on the machine speech chain, LPC and MFCC features are extracted from the speaker's mouth to the listener's ear to represent the pronunciation representation complementarily.
- A segmentation and fusion method is proposed to extract intra- and inter-segment features from variable-length speech without cropping and redundancy.
- A framework is constructed to capture text-independent depressive features for recognition, which suggests that the vocal tract changes in patients also deserve attention for audio-based depression diagnosis.

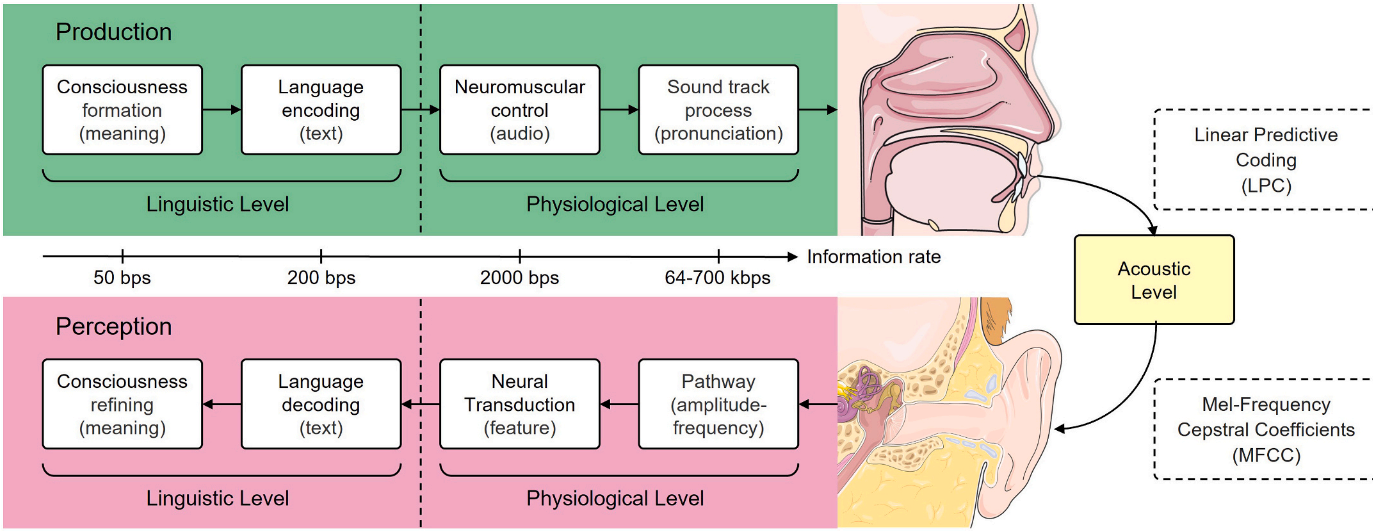

Fig. 1. The speech chain according to Denes et al. (1993); Tjandra et al. (2020).

- The rest of this paper is organized as follows. The theoretical foundations and the details of proposed MSCDR are introduced in Section 2. Section 3 reports the experimental setup and recognition results, which are discussed in Section 4. Finally, the conclusion and future work directions are in Section 5.

2. Materials and methods

2.1. Speech production and perception features

The speech chain concept was first introduced by Denes et al. (1993), and it explains the physics and biology involved during a closed-loop process of a message's production, propagation, and perception from the speaker to the listener. Based on this theory, Chowdhury and Ross (2020) took a step further and first developed a closed-loop speech chain model based on deep learning that integrated human speech perception and production behaviors for identity recognition. This work improved the performance compared to that of separate systems. We reproduced the visualization of the speech chain based on the descriptions in previous studies (Denes et al., 1993; Tjandra et al., 2020) (Fig. 1). During the production process, the message is encoded to text at the linguistic level, and the vocal tract generates the sound from the articulation, thus imparting the spoken language its acoustic properties at the physiological level. During the perception process, the acoustic meatus extracts speech features essential at the physiological level, and the text is decoded to meaning at the linguistic level. It is worth noting that the information rate contained in the transmitted spoken message is significantly higher than the base information rate of the text message itself. Therefore, the phonetic difference of depressed patients could be reflected objectively in the process of speech production and generation rather than the separation process used in previous studies. We first implemented the speech chain model for depression diagnosis that integrated human speech perception and production phonetic behaviors. Referring to the previous study for speaker identity recognition

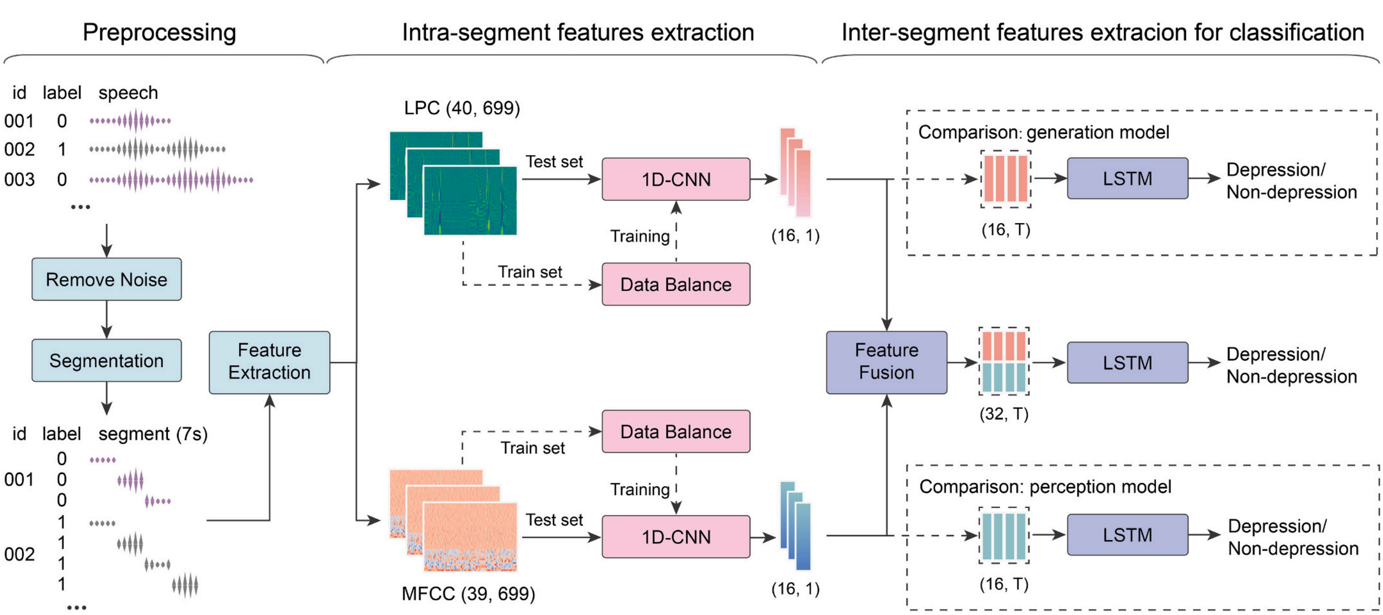

Fig. 2. Visualization of the overall process of the proposed MSCDR, including three parts: preprocessing, intra-segment features extraction and inter-segment features extraction for classification. The black dotted line represents the training stage. There is a balancing process for positive and negative sample numbers during the training stage of the 1D-CNN. In the dotted box are two single models for comparison.

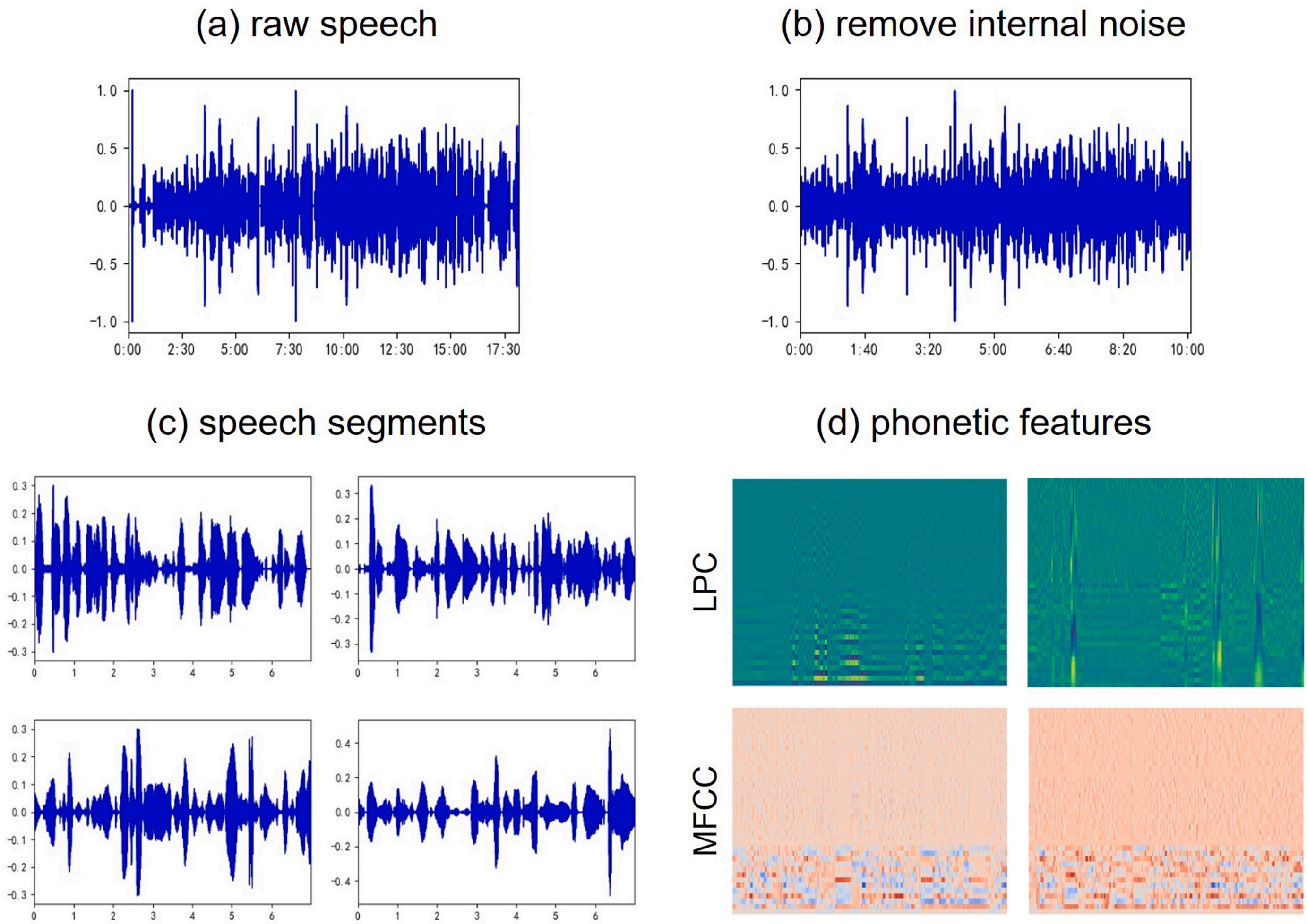

Fig. 3. Partial preprocessing results of one subject with (a) raw speech, (b) speech without internal noise, (c) speech segments with 7 s, and (d) phonetic features.

(Chowdhury and Ross, 2020), we used LPC to model the speech production process and MFCC to model the speech perception process. The combination covers the closed-loop speech chain complementarily and extracts depression-related information effectively.

2.1.1. Linear predictive coding

According to the source-filter model of speech, the human voice is excited from the lung as energy source and processed through the vocal tract as filter (Guzman et al., 2020; Mittal and Sharma, 2021). The information contained in the speech signal is formed by the modulation of the vocal tract as a time-varying filter, rather than the energy source. Linear predictive coding (LPC) is the digital filter parameter for simulating the vocal tract to reflect the characteristics of the speaker. Because human voice is a highly correlated sequence, a linear combination of p past speech samples could predict the next speech sample $\widehat{x}(n)$, as given by:

$$\widehat{x}(n) = \sum_{k=1}^{p} a_k x(n-k).$$

a_k are the vocal tract filter coefficients; $x(k)(k = 1, 2, 3, \ldots, p)$ is the k past speech sample. The real n speech sample is $x(n)$, and the prediction error e_n could be given as:

$$e_n = x(n) - \widehat{x}(x) = x(n) - \sum_{k=1}^{p} a_k x(n-k).$$

By minimizing the mean square error of e_n, the filter coefficients $a_k(k = 1, 2, 3, \ldots, p)$ can be obtained as the p-order LPC, which provides an estimate of the human vocal tract filter coefficients.

2.1.2. Mel-frequency Cepstral Coefficients

As shown in Fig. 1, the auditory pathway separates sounds based on their frequency content and converts sound waves into neural signals for brain. Mel-frequency Cepstral Coefficients (MFCC) model the human peripheral auditory system and are widely used in speech recognition (Rejaibi et al., 2022). MFCC describes the energies of the cepstrum in a nonlinear scale, the Mel scale. This scale reflects the characteristics of the human ear, which is more sensitive to low-frequency sounds than to high-frequency sounds. The relationship between the Mel scale and frequency can be approximated by:

$$Mel(f) = 2595 \times lg\left(1 + \frac{f}{700}\right).$$

MFCC is extracted as follows: 1) Calculate Fast Fourier Transform spectrum from the frequency, 2) Extract the filter bank output allocated on the Mel scale, and 3) Obtain the cepstrum coefficient through Discrete Cosine Transform (Guzman et al., 2020).

2.2. The proposed MSCDR

The overall process of the proposed MSCDR consists of three parts: preprocessing, intra-segment features extraction, and inter-segment features extraction for classification, as shown in Fig. 2. The raw speech of each subject is divided into segments sequentially and then

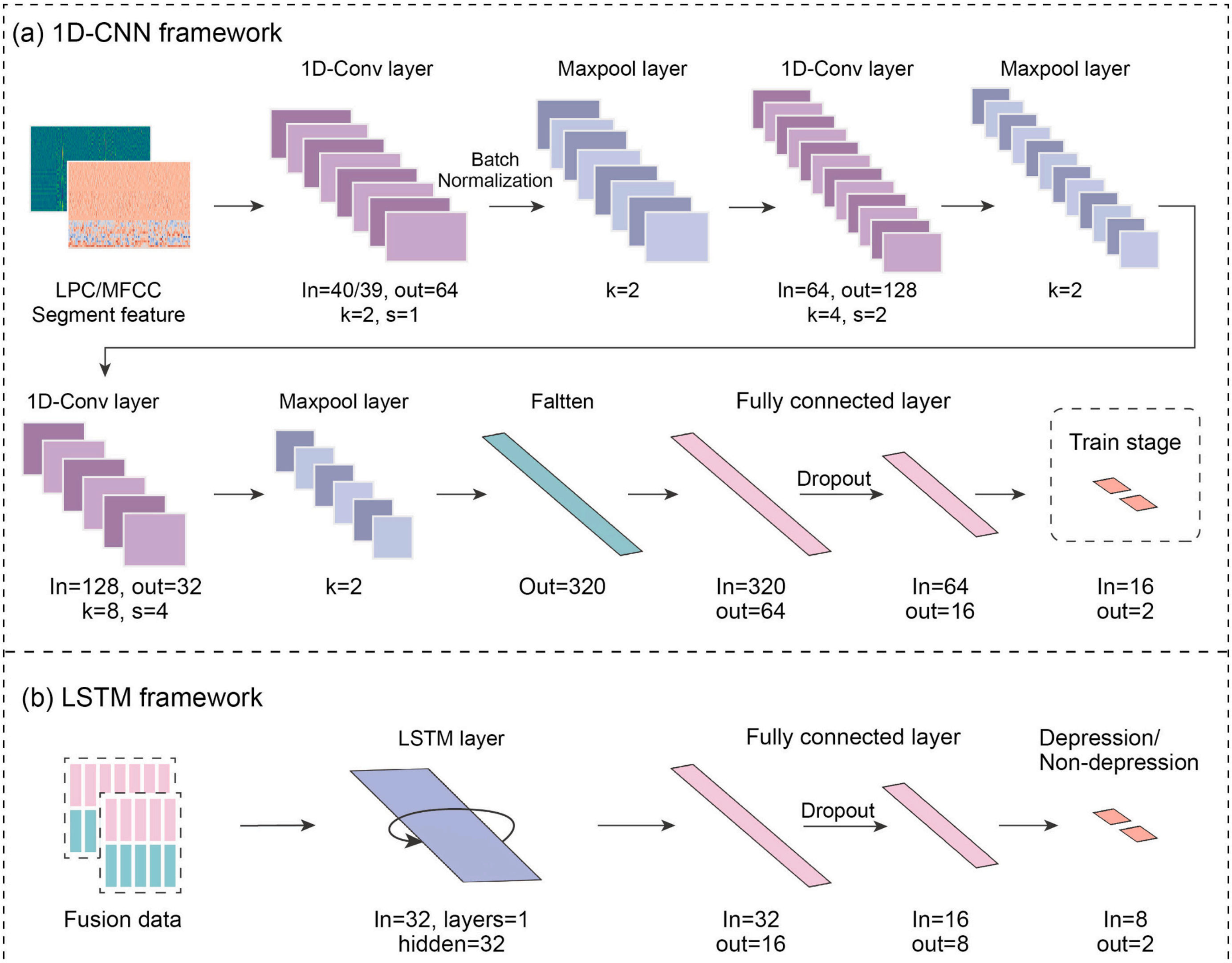

Fig. 4. Proposed deep learning architecture with (a) a 1D-CNN framework for intra-segment features and (b) an LSTM framework for inter-segment features and classification. The dense layer in the dotted line is used only during training to calculate the network weights. For LSTM, the input dimension of the MSCDR is 32 and the input dimension of the single models is 16. K stands for the kernel size and s stands for the stride.

LPC and MFCC features are extracted after preprocessing. A 1D-CNN is established to extract intra-segment high-level depressive features from the LPC and MFCC features. After that, all of the segment features of each subject are fused in the time domain, and the depressive correlation information between segments is extracted through the LSTM for classification. To further verify the improvement of the machine speech chain, two single models are constructed: the generation model extracts phonetic features only from the speech generation process, and the perception model extracts only from the speech perception process.

2.2.1. Preprocessing

Raw speech contains internal noise captured during collection, such as the interviewer's voice and mute clips, which are unassociated with depression and therefore affect the recognition performance. We removed the noise part, then divided the whole speech into segments of 7 s without overlap, and recorded their sequence, as shown in Fig. 3. As mentioned earlier, such segmentation is used to unify the speech with different lengths and it increases the number of samples for training. Semantic destruction during segmentation does not affect the text-independent classification model. 7 s as the segment length is based on the results of enumeration experiments and is consistent with previous research (Alghifari et al., 2019). After that, LPC and MFCC features are extracted using the same sliding window of length [0.025 × fs] with [0.01 × fs] stride (fs is the sampling frequency) and a Hamming window. The LPC feature comprises 20 filter coefficients and 20 first-order delta coefficients, and the MFCC feature comprises 13 Mel-cepstral coefficients, 13 first-order, and 13 second-order delta coefficients.

2.2.2. Intra-segment features extraction

After preprocessing, all of the segments of each subject are mixed to eliminate the influence of subject identity, and a 1D-CNN is established to extract high-level depression-related features from the segments. Since the axes of the LPC and MFCC represent different magnitudes and correspond to time and frequency dimensions with completely different meanings, the convolution of all frequencies by the 1D-CNN increases the model's sensitivity to the frequency domain (Vázquez-Romero and Gallardo-Antolín, 2020). Fig. 4(a) shows the structure of the 1D-CNN framework. The combination of 1D-CNN and 1D maximum pooling enables the model to capture short-term temporal dynamic information and frequency correlations effectively. The batch normalization and dropout layers improve training speed and prevent overfitting. Dense

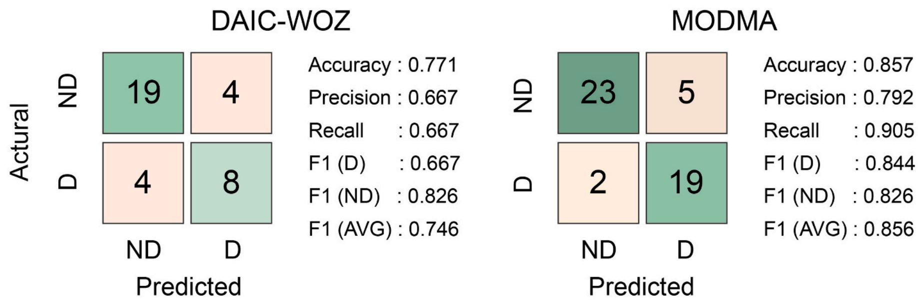

Fig. 5. Confusion matrixes of the proposed MSCDR were generated on the test sets of DAIC-WOZ and MODMA. ND stands for non-depression and D stands for depression.

layers further extract features and reduce the output dimension. To avoid class imbalance, the quantity of depressed and non-depressed segments in the training set is balanced at the ratio of 1:1 during the training stage. During the test stage, the 16-dimensional output of the penultimate dense layer for each segment is only reserved for the next session.

2.2.3. Inter-segment features extraction for classification

After intra-segment features extraction, 16-dimensional outputs that include depression-related information of the short segments are obtained. The MSCDR concatenates the outputs of each LPC and MFCC feature from each segment to the 32-dimensional segment feature for integrating the process of speech generation and perception. Then, all of the features from each subject are spliced from the segment level to the personal level at the original time domain order. For single models, all of the segment features are directly spliced to the personal layer without dimension concatenation. Finally, a one-layer LSTM is built to capture short- and long-term temporal correlation features between the segments at the personal level. Two dense layers are included to reduce the dimensionality and conclude the classification, as shown in Fig. 4(b). The recurrent layers of the LTSM consume variable-length inputs and ultimately produce only the layer's output at the final sequential step, which effectively deals with the inconsistent length speech of different subjects.

2.2.4. Performance metrics

Classification performance is determined using the confusion matrix, accuracy, and F1 score, similar to previous studies (Valstar et al., 2016; Zhao et al., 2021). F1 score is the harmonic mean of precision and recall, and it is a helpful evaluation criterion for unbalanced classification problems.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

TP/*FP* indicates true/false positives samples, and *TN*/*FN* indicates true/false negatives samples. A larger F1 score implies better discrimination.

3. Results

The proposed MSCDR can identify depression by pronunciation representation, regardless of the language, content, or habits of speech. To verify the text-independence of this method, we tested it on two datasets with different paradigms and languages and compared the classification results with previous studies.

Table 1
Comparison of the proposed MSCDR with existing methods. The second and third lines of MODMA are the reproduction results based on a RNN method (Rejaibi et al., 2022) and DepAudioNet (Ma et al., 2016), respectively. ND stands for non-depression, and D stands for depression.

Dataset	Method	Feature	Model	F1 score (D/ND)	F1 score (AVG)
DAIC-WOZ	Valstar et al., 2016	F0, VUV, and MFCC	SVM	0.41/0.58	0.50
	Ma et al., 2016	Mel-Spectrogram	CNN-LSTM	0.52/0.70	0.61
	Huang et al., 2020	MFCC	FVTC-CNN	0.40/0.84	0.62
	Rejaibi et al., 2022	MFCC	RNN	0.46/0.85	0.64
	Othmani et al., 2021	MFCC, Spectrogram	CNN	0.49/0.82	0.66
	Dumpala et al., 2021	OpenSMILE	LSTM	0.50/0.82	0.66
	MSCDR (ours)	LPC-MFCC	CNN-LSTM	0.67/0.83	0.75
MODMA	Chen and Pan, 2021	eGeMAPS	Decision Tree	0.80/–	–
	Rejaibi et al., 2022	MFCC	RNN	0.66/0.73	0.70
	Ma et al., 2016	Mel-Spectrogram	CNN-LSTM	0.82/0.75	0.79
	MSCDR (ours)	LPC-MFCC	CNN-LSTM	0.84/0.87	0.86

3.1. Two public datasets

DAIC-WOZ (Gratch et al., 2014) is supported by the AVEC2017 challenge (Ringeval et al., 2017) and is useful to understand several typical mental disorders: anxiety, depression, and post-traumatic stress disorder, for example. It recorded clinical interview audio of 189 subjects and discriminated between depressed and non-depressed patients by professionals combined with PHQ-8 binary. The paradigm hosted by a human-controlled virtual interviewer called Ellie. Ellie has a fixed set of utterances, and it provides feedback based on subjects' responses in real time. English is the language for questions and answers. The average length of audio recordings is 15 min, with a sampling frequency of 16 kHz. Consistent with the public split of DAIC-WOZ, the training set (30 depression vs. 77 non-depression) and development set (12 depression vs. 23 non-depression) are used to train and test.

MODMA (Cai et al., 2022), supported by Lanzhou University, China, is applicable for mental disorder analysis. It contains 52 subjects (23 depressed outpatients, 29 non-depressed subjects, and 1 depression data defective). Depressed patients were recruited among inpatients and outpatients that met the major depression diagnostic criteria of the

Table 2
Comparison of the proposed MSCDR with two single methods: the generation model is only from the speech generation process and the perception model is only from the speech perception process. ND stands for non-depression and D stands for depression.

Dataset	Model	Accuracy	F1 score		
			D	ND	AVG
DAIC-Woz	Generation Model	0.771	0.500	0.852	0.676
	Perception Model	0.714	0.583	0.783	0.683
	MSCDR	0.771	0.667	0.826	0.746
MODMA	Generation Model	0.837	0.789	0.867	0.828
	Perception Model	0.816	0.790	0.836	0.814
	MSCDR	0.857	0.844	0.868	0.856

Diagnostic and Statistical Manual of Mental Disorders (DSM). Healthy controls were recruited by posters and were excluded for other diseases. Each subject is asked to complete 29 recording tasks in different speaking patterns: interview, words and passage reading, and picture description under three kinds of emotional valences: positive, neutral, and negative. The passage reading recordings were excluded due to poor induction, which is explained in the discussion part. The spoken language is Chinese, and the recording length ranges from seconds to minutes. To avoid too short recordings, we combined several recordings from the same subject and randomly divided all into a training set (89 depression vs. 117 non-depression) and a test set (21 depression vs. 28 non-depression) in the 8:2 ratio.

3.2. Recognition performance

During the training stage of the MSCDR, PyTorch was used to implement the system with 0.01 as the learning rate, 16 as the batch size, cross-entropy as the loss function, and an early stop mechanism to prevent overfitting. Fig. 5 shows the confusion matrix generated on the test sets of DAIC-WOZ and MODMA. We calculated evaluation indicators and compared them with previous models (Table 1). The accuracy of the proposed MSCDR on DAIC-WOZ and MODMA was 0.77 and 0.86, respectively, and the average F1 score was 0.75 and 0.86, respectively, which was 0.09 and 0.07 higher than the existing methods. In particular, the F1 score for depression significantly improved by 0.17 on DAIC-WOZ, indicating that the MSCDR significantly enhanced the ability of capturing depression information.

The baseline (Valstar et al., 2016) of DAIC-WOZ and the method in (Chen and Pan, 2021) on MODMA only used statistical functions (e.g., mean, median) on whole phonetic features, which were not sensitive to temporal changes. In contrast, the studies (Ma et al., 2016; Huang et al., 2020; Rejaibi et al., 2022; Othmani et al., 2021) extracted time-frequency characteristics of speech (Mel-Spectrogram, MFCC) that contained dynamic information and obtained better results. However, the features used by these methods were only from the speech perception process and did not take into account the changes in the patient's vocal tract. Compared to these, the proposed MSCDR extracts from both the processes of speech generation and of speech perception and further improves classification performance effectively. The improvement of F1 score for depression reflects the effectiveness of the 1D-CNN and LSTM in extracting depression-related information. Considering that not the whole speech of depressed people is depressed and non-depressed

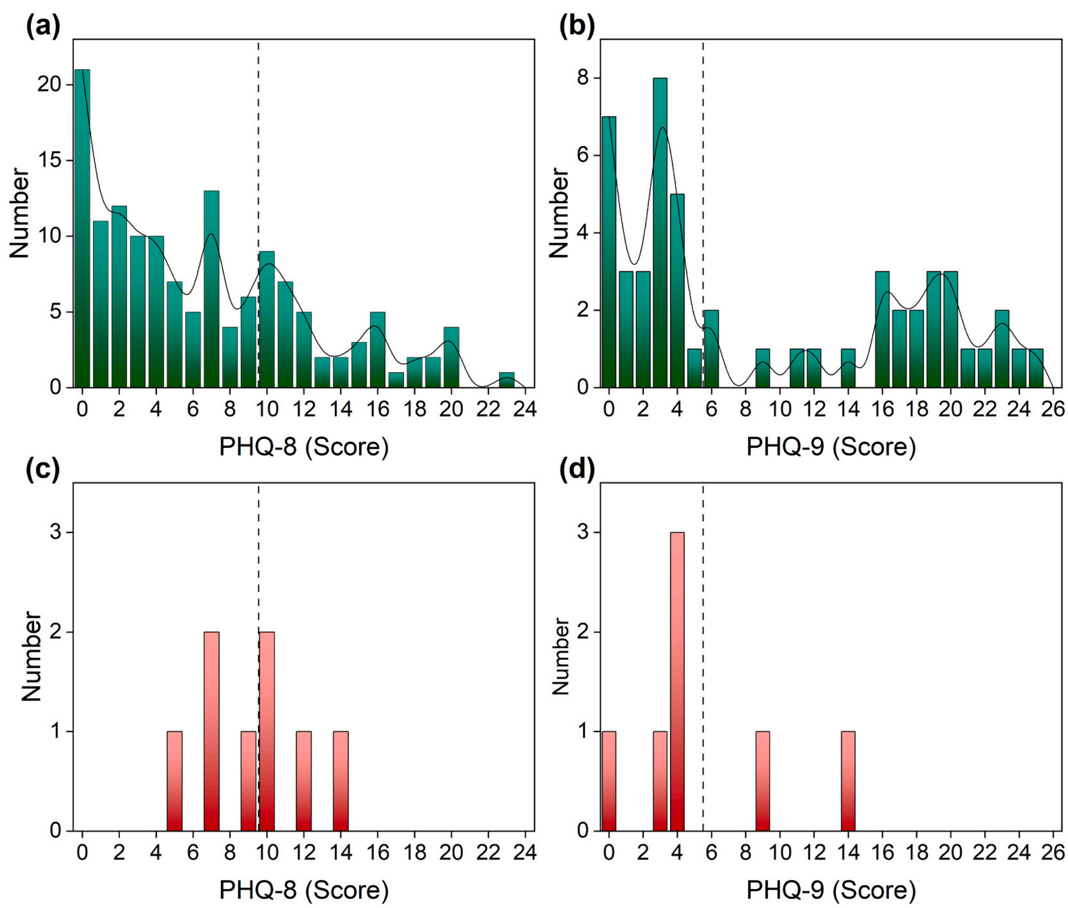

Fig. 6. (a) PHQ-8 score distribution of DAIC-WOZ subjects. (b) PHQ-9 score distribution of MODMA subjects. (c) PHQ-8 score distribution of the misclassified subjects of DAIC-WOZ. (d) PHQ-9 score distribution of the misclassified subjects of MODMA. The vertical dashed lines indicate the threshold for depression and non-depression.

people may also have depressed segments, the 1D-CNN first extracts depression-related information at the segment level rather than the whole speech by virtue of its sensitivity to the frequency domain. Then, the LSTM network classifies by capturing inter-segment correlation. Compared with whole speech-based classification, this method provides a new idea for depression diagnosis. In addition, the excellent results for the English dataset DAIC-WOZ and Chinese dataset MODMA demonstrate the text-independence of the MSCDR, which meets the requirements of audio-based depression diagnosis, namely, to classify pronunciation features, regardless of the language, content, or speech habits.

4. Discussion

4.1. Comparison to single models

To further verify the improvement of the machine speech chain, we compared the MSCDR with two single models under the same conditions, as shown in Fig. 2. The generation model extracted the depression-related features only from the speech generation process, and the perception model extracted them only from the speech perception process. The results in Table 2 indicate that there was a significant improvement from using the MSCDR compared to the single models. The proposed MSCDR extracts LPC from the generation process and MFCC from the perception process. As the LPC describes the vocal tract of the speaker and MFCC describes the perceptual law of the human ear, their combination represents the pronunciation representation complementarily. This improvement also proves to some extent that the vocal tract of depressed patients has changed, which is consistent with the hypothesis in (Espy-Wilson et al., 2019; Seneviratne et al., 2020). Therefore, the change of the physiological structure of depressed patients also deserves attention for audio-based automatic diagnosis of depression, which will lead to an improvement compared with manual diagnosis.

4.2. Misclassification analysis

Although the classification results of the MSCDR on both datasets were excellent, the average F1 score on DAIC-WOZ was significantly lower than that on MODMA, with the obvious gap of 0.11. We analyzed three factors that potentially affected the performance on DAIC-WOZ. The first factor is the possible label errors in DAIC-WOZ. The labels of MODMA are scientifically co-labeled by multiple scales and physician diagnosis, including PHQ-9 (Kroenke and Spitzer, 2002), (Gerdner and Allgulander, 2009), and GAD-7 (Spitzer et al., 2006). However, DAIC-WOZ binarizes the labels of subjects using only the PHQ-8 scale (Kroenke et al., 2009), which is not rigorous in clinical diagnosis, with a great likelihood of mislabeling. The second factor is that depressive symptoms are not prominent in DAIC-WOZ. As shown in Fig. 6(a) and (b), the score distribution of DAIC-WOZ is concentrated in non-depressed or mild patients whereas that of MODMA is more scattered. Such concentrated distribution of DAIC-WOZ might affect the training of the model. In addition, Fig. 6(c) shows that the misclassification subjects in DAIC-WOZ are concentrated at the threshold for depression and non-depression. The same phenomenon also happens on the MODMA dataset, as seen in Fig. 6(d). Thus, the third factor is that the audio features of depression may be subject to aliasing, that is, some non-depressed subjects with high scale scores may also show depressive phonetic features, and some mild patients may have no depressive phonetic symptoms. This phenomenon is also noteworthy and has not been mentioned before.

4.3. Speech tasks selection

Unlike DAIC-WOZ that has only interview tasks, MODMA consists of five speech tasks: interview, passage reading, words reading, picture description, and the Thematic Apperception Test (TAT), which have different inducing effects. Similar to (Taguchi et al., 2018), we performed statistical analysis on each dimension of the MFCC features between depressed and non-depressed groups to analyze the inducing effect of the different tasks. First, we used the Kolmogorov–Smirnov test to verify that the samples conform to the normal distribution. Then, Levene's test was used to test the homogeneity of variance. If satisfied, the Student's *t*-test was used; otherwise, Welch's t-test was used. And we corrected our findings for the hypothesis testing with the use of a false discovery rate (FDR) calculation. Table 3 shows the 39-dimensional statistical analysis results of each task. As we can see, the number of features with a significant difference in the passage reading task was far less than that in other tasks, indicating the low induction effect of this task, which is consistent with the previous results (Long et al., 2017; Rejaibi et al., 2022). We believe that this was due to its fixed content and the inconsistent familiarity of the subjects and therefore excluded passage reading recordings in MODMA from this study.

Table 3
p values of the 39-dimensional MFCC features for five speech tasks in MODMA. $p < 0.05$ indicates significant difference.

Feature	Interview	Passage reading	Words reading	Picture description	TAT
MFCC-0	0.01	0.02	0.07	0.01	0.01
MFCC-1	0.01	0.08	0.04	0.00	0.00
MFCC-2	0.84	0.59	0.79	0.92	0.79
MFCC-3	0.84	0.90	0.53	0.34	0.49
MFCC-4	0.85	0.98	0.85	0.58	0.94
MFCC-5	0.28	0.97	0.44	0.24	0.00
MFCC-6	0.77	0.62	0.48	0.75	0.68
MFCC-7	0.00	0.09	0.00	0.00	0.45
MFCC-8	0.22	0.61	0.41	0.25	0.14
MFCC-9	0.04	0.04	0.07	0.01	0.01
MFCC-10	0.70	0.62	0.48	0.73	0.92
MFCC-11	0.39	0.27	0.02	0.19	0.52
MFCC-12	0.14	0.58	0.59	0.04	0.21
MFCC-13	0.35	0.63	0.43	0.85	0.02
MFCC-14	0.49	0.49	0.87	0.66	0.19
MFCC-15	0.12	0.37	0.70	0.20	0.32
MFCC-16	0.12	0.37	0.70	0.20	0.32
MFCC-17	0.81	0.88	0.60	0.31	0.35
MFCC-18	0.83	0.71	0.43	0.46	0.36
MFCC-19	0.54	0.13	0.87	0.86	0.46
MFCC-20	0.72	0.70	0.61	0.79	0.36
MFCC-21	0.49	0.10	0.86	0.05	0.56
MFCC-22	0.52	0.35	0.42	0.84	0.65
MFCC-23	1.00	0.90	0.66	0.68	0.86
MFCC-24	0.12	0.53	0.02	0.05	0.77
MFCC-25	0.83	0.76	0.03	0.47	0.19
MFCC-26	0.45	0.62	0.39	0.85	0.02
MFCC-27	0.44	0.51	0.85	0.60	0.20
MFCC-28	0.12	0.42	0.91	0.21	0.34
MFCC-29	0.17	0.05	0.53	0.10	0.62
MFCC-30	0.81	0.94	0.60	0.42	0.51
MFCC-31	0.90	0.61	0.54	0.31	0.28
MFCC-32	0.66	0.11	0.91	0.85	0.44
MFCC-33	0.52	0.60	0.53	0.87	0.28
MFCC-34	0.59	0.07	0.89	0.03	0.78
MFCC-35	0.41	0.33	0.27	0.76	0.51
MFCC-36	1.00	0.98	0.69	0.73	0.94
MFCC-37	0.03	0.41	0.03	0.05	0.83
MFCC-38	0.63	0.66	0.04	0.38	0.12
The number of $p < 0.05$	5	2	7	6	6

4.4. Brief summary

Our results demonstrate the potential association between paralinguistic representation and depression, and further suggest that speech could be used as a powerful tool for early detection of mental disorders. Currently, there are many physiological programs on psychiatric disorders to explore their cognitive and pathological mechanisms. For example, the studies of metabolism (Bocchio-Chiavetto et al., 2018), genes (Li et al., 2021), electroencephalogram (Saeedi et al., 2021), magnetic resonance imaging (Squarcina et al., 2017) have made some important progress in exploring their physiological mechanism. We believe speech could be an important supplement to the understanding of mental disorders with its low acquisition cost and strong popularity. Furthermore, the audio-based diagnosis technology could be applied to smart devices such as mobile phones and bracelet to actively detect people's mental health, which can deal with potential mental health risks in society and has broad application prospects.

5. Conclusion

In this study, we proposed a MSCDR that extracts phonetic features from the speech perception and production processes complementarily for automatic depression recognition. The excellent classification results on two datasets with different paradigms and languages prove the good generalization ability and superiority of the proposed MSCDR. Due to the limitation of small sample size, we cannot apply MSCDR to the diagnosis of depression levels. Next, we will expand the sample size and make further verification before clinical translation. We believe this study suggests that the changes of the vocal tract in patients with depression deserve attention, and also provides theoretical basis and inspiration for the research of audio-based depression diagnosis.

CRediT authorship contribution statement

M. Du, S. Liu and D. Ming designed the study. T. Wang and L. Chen did data curation and formal analysis. M. Du, W. Zhang and Y. Ke built the model, tested and visualized. M. Du and S. Liu drafted the manuscript. All the authors contributed to the interpretation of the results, manuscript revision, and approved the final version of the manuscript.

Funding

Research supported by the National Natural Science Foundation of China under Grant 81925020 and 81801786, and the General Program of Tianjin, China under Grant 19JCYBJC29200.

Conflict of interest

No potential conflict of interest was reported by the authors.

Acknowledgments

The authors sincerely thank to the collectors and participants of DAIC-WOZ and MODMA for providing the audio data for this study. The Fig. 1 contains modified images from Servier Medical Art licensed by a Creative Commons Attribution 3.0 Unported License.

References

Alghifari, M.F., Gunawan, T.S., Nordin, M.A.W., Kartiwi, M., Borhan, L., 2019. On the optimum speech segment length for depression detection. In: 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA). IEEE, pp. 1–5. https://doi.org/10.1109/ICSIMA47653.2019.9057319.

Bocchio-Chiavetto, L., Zanardini, R., Tosato, S., Ventriglia, M., Ferrari, C., Bonetto, C., Lasalvia, A., Giubilini, F., Fioritti, A., Pileggi, F., Pratelli, M., Pavanati, M., Favaro, A., De Girolamo, G., Frisoni, G.B., Ruggeri, M., Gennarelli, M., 2018. Immune and metabolic alterations in first episode psychosis (FEP) patients. Brain Behav. Immun. 70, 315–324. https://doi.org/10.1016/j.bbi.2018.03.013.

Cai, H., Yuan, Z., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., Li, J., Yang, Z., Li, X., Zhao, Q., Liu, Z., Yao, Z., Yang, M., Peng, H., Zhu, J., Zhang, X., Gao, G., Zheng, F., Li, R., Guo, Z., Ma, R., Yang, J., Zhang, L., Hu, X., Li, Y., Hu, B., 2022. A multi-modal open dataset for mental-disorder analysis. Sci. Data 9, 178. https://doi.org/10.1038/s41597-022-01211-x.

Chen, X., Pan, Z., 2021. A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health. Int. J. Environ. Res. Public Health 18, 6441. https://doi.org/10.3390/ijerph18126441.

Chowdhury, A., Ross, A., 2020. Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals. IEEE Trans. Inf. Forensics Secur. 15, 1616–1629. https://doi.org/10.1109/TIFS.2019.2941773.

Costantini, L., Costanza, A., Odone, A., Aguglia, A., Escelsior, A., Serafini, G., Amore, M., Amerio, A., 2021. A breakthrough in research on depression screening: from validation to efficacy studies. Acta Biomed. Ateneo Parmense 92. https://doi.org/10.23750/abm.v92i3.11574 e2021215–e2021215.

Dai, Z., Zhou, H., Ba, Q., Zhou, Y., Wang, L., Li, G., 2021. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. J. Affect. Disord. 295, 1040–1048. https://doi.org/10.1016/j.jad.2021.09.001.

Denes, P.B., Denes, P., Pinson, E., 1993. The Speech Chain. Macmillan.

Dong, Y., Yang, X., 2021. A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing 441, 279–290. https://doi.org/10.1016/j.neucom.2021.02.019.

Dumpala, S.H., Rodriguez, S., Rempel, S., Uher, R., Oore, S., 2021. Significance of Speaker Embeddings and Temporal Context for Depression Detection. https://doi.org/10.48550/arXiv.2107.13969 arXiv.

Espy-Wilson, C., Lammert, A.C., Seneviratne, N., Quatieri, T.F., 2019. Assessing neuromotor coordination in depression using inverted vocal tract variables. In: Interspeech 2019. Presented at the Interspeech 2019. ISCA, pp. 1448–1452. https://doi.org/10.21437/Interspeech.2019-1815.

Gerdner, A., Allgulander, C., 2009. Psychometric properties of the swedish version of the childhood trauma Questionnaire—Short form (CTQ-SF). Nord. J. Psychiatry 63, 160–170. https://doi.org/10.1080/08039480802514366.

Gómez-Rodellar, A., Palacios-Alonso, D., Ferrández Vicente, J.M., Mekyska, J., Álvarez-Marquina, A., Gómez-Vilda, P., 2020. A methodology to differentiate Parkinson's disease and aging speech based on glottal flow acoustic analysis. Int. J. Neural Syst. 30, 2050058. https://doi.org/10.1142/S0129065720500586.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., 2014. The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 3123–3128.

Guzman, M., Bertucci, T., Pacheco, C., Leiva, F., Quintana, F., Ansaldi, R., Quezada, C., Munoz, D., 2020. Effectiveness of a physiologic voice therapy program based on different semioccluded vocal tract exercises in subjects with behavioral dysphonia: a randomized controlled trial. J. Commun. Disord. 87, 106023 https://doi.org/10.1016/j.jcomdis.2020.106023.

Hammar, Å., Ronold, E.H., Rekkedal, G.Å., 2022. Cognitive impairment and neurocognitive profiles in major Depression—A clinical perspective. Front. Psychiatry 13. https://doi.org/10.3389/fpsyt.2022.764374.

Hartmann, R., Schmidt, F.M., Sander, C., Hegerl, U., 2019. Heart rate variability as indicator of clinical state in depression. Front. Psychiatry 9, 735. https://doi.org/10.3389/fpsyt.2018.00735.

He, L., Cao, C., 2018. Automated depression analysis using convolutional neural networks from speech. J. Biomed. Inform. 83, 103–111. https://doi.org/10.1016/j.jbi.2018.05.007.

He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., 2022. Deep learning for depression recognition with audiovisual cues: a review. Inf. Fusion 80, 56–86. https://doi.org/10.1016/j.inffus.2021.10.012.

Huang, Z., Epps, J., Joachim, D., 2020. 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). Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6549–6553. https://doi.org/10.1109/ICASSP40776.2020.9054323.

Jia, Y., Liang, Y., Zhu, T., 2019. An analysis of voice quality of Chinese patients with depression. In: 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA). IEEE, pp. 1–6. https://doi.org/10.1109/O-COCOSDA46868.2019.9060848.

Jiang, H., Hu, B., Liu, Z., Wang, G., Zhang, L., Li, X., Kang, H., 2018. Detecting depression using an ensemble logistic regression model based on multiple speech features. Comput. Math. Methods Med. 2018, 6508319. https://doi.org/10.1155/2018/6508319.

Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr. Ann. 32, 509–515. https://doi.org/10.3928/0048-5713-20020901-06.

Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T., Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. J. Affect. Disord. 114, 163–173. https://doi.org/10.1016/j.jad.2008.06.026.

Lam-Cassettari, C., Kohlhoff, J., 2020. Effect of maternal depression on infant-directed speech to prelinguistic infants: implications for language development. PLOS ONE 15, e0236787. https://doi.org/10.1371/journal.pone.0236787.

Li, X., Su, X., Liu, J., Li, H., Li, M., Li, W., Luo, X.-J., 2021. Transcriptome-wide association study identifies new susceptibility genes and pathways for depression. Transl. Psychiatry 11, 306. https://doi.org/10.1038/s41398-021-01411-w.

Liu, Z., Wang, D., Zhang, L., Hu, B., 2020. A Novel Decision Tree for Depression Recognition in Speech. ArXiv Prepr. ArXiv200212759.

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., Cai, H., 2017. Detecting depression in speech: comparison and combination between different speech types. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Presented at the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, Kansas City, MO, pp. 1052–1058. https://doi.org/10.1109/BIBM.2017.8217802.

Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y., 2016. DepAudioNet: an efficient deep model for audio based depression classification. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Presented at the MM '16: ACM Multimedia Conference, ACM, Amsterdam The Netherlands, pp. 35–42. https://doi.org/10.1145/2988257.2988267.

Madhavi, I., Chamishka, S., Nawaratne, R., Nanayakkara, V., Alahakoon, D., De Silva, D., 2020. A deep learning approach for work related stress detection from audio streams in cyber physical environments. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). Presented at the 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, Vienna, Austria, pp. 929–936. https://doi.org/10.1109/ETFA46521.2020.9212098.

Mittal, V., Sharma, R.K., 2021. Classification of parkinson disease based on analysis and synthesis of voice signal. Int. J. Healthc. Inf. Syst. Inform. 16 https://doi.org/10.4018/IJHISI.20211001.oa30.

Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., Othmani, A., 2020. AudVowelConsNet: a phoneme-level based deep CNN architecture for clinical depression diagnosis. Mach. Learn. Appl. 2, 100005 https://doi.org/10.1016/j.mlwa.2020.100005.

Negi, H., Bhola, T., Pillai, M.S., Kumar, D., 2018. A novel approach for depression detection using audio sentiment analysis. Int. J. Inf. Syst. Manag. Sci. 1.

Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., Hadid, A., 2021. Towards robust deep neural networks for affect and depression recognition from speech. In: International Conference on Pattern Recognition. Springer, pp. 5–19. https://doi.org/10.1007/978-3-030-68790-8_1.

Pampouchidou, A., Simantiraki, O., Fazlollahi, A., Pediaditis, M., Manousos, D., Roniotis, A., Giannakakis, G., Meriaudeau, F., Simos, P., Marias, K., Yang, F., Tsiknakis, M., 2016. Depression assessment by fusing high and low level features from audio, video, and text. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16. Association for Computing Machinery, New York, NY, USA, pp. 27–34. https://doi.org/10.1145/2988257.2988266.

Pan, W., Wang, J., Liu, T., Liu, X., Liu, M., Hu, B., Zhu, T., 2018. Depression recognition based on speech analysis. Chin. Sci. Bull. 63, 2081–2092. https://doi.org/10.1360/N972017-01250.

Patil, M., Wadhai, V., 2021. Selection of classifiers for depression detection using acoustic features. In: 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA). Presented at the 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), pp. 1–4. https://doi.org/10.1109/ICCICA52458.2021.9697240.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., Othmani, A., 2022. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomed. Signal Process. Control 71, 103107. https://doi.org/10.1016/j.bspc.2021.103107.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M., 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th Annual Workshop on Audio/visual Emotion Challenge, pp. 3–9. https://doi.org/10.1145/3133944.3133953.

Saeedi, A., Saeedi, M., Maghsoudi, A., Shalbaf, A., 2021. Major depressive disorder diagnosis based on effective connectivity in EEG signals: a convolutional neural network and long short-term memory approach. Cogn. Neurodyn. 15, 239–252. https://doi.org/10.1007/s11571-020-09619-0.

Sahu, S., Espy-Wilson, C., 2014. Effects of depression on speech. J. Acoust. Soc. Am. 136, 2312–2312.

Sealock, J.M., Lee, Y.H., Moscati, A., Venkatesh, S., Voloudakis, G., Straub, P., Singh, K., Feng, Y.-C.A., Ge, T., Roussos, P., 2021. Use of the PsycheMERGE network to investigate the association between depression polygenic scores and white blood cell count. JAMA Psychiatry 78, 1365–1374. https://doi.org/10.1001/jamapsychiatry.2021.2959.

Seneviratne, N., Williamson, J.R., Lammert, A.C., Quatieri, T.F., Espy-Wilson, C., 2020. Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression. In: Interspeech 2020. Presented at the Interspeech 2020, ISCA, pp. 4551–4555. https://doi.org/10.21437/Interspeech.2020-2758.

Silva, W.J., Lopes, L., Galdino, M.K.C., Almeida, A.A., 2021. Voice acoustic parameters as predictors of depression. J. Voice S0892199721002058. https://doi.org/10.1016/j.jvoice.2021.06.018.

Simantiraki, O., Charonyktakis, P., Pampouchidou, A., Tsiknakis, M., Cooke, M., 2017. Glottal source features for automatic speech-based depression assessment. In: Interspeech 2017. Presented at the Interspeech 2017, ISCA, pp. 2700–2704. https://doi.org/10.21437/Interspeech.2017-1251.

Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med. 166, 1092–1097. https://doi.org/10.1001/archinte.166.10.1092.

Squarcina, L., Castellani, U., Bellani, M., Perlini, C., Lasalvia, A., Dusi, N., Bonetto, C., Cristofalo, D., Tosato, S., Rambaldelli, G., Alessandrini, F., Zoccatelli, G., Pozzi-Mucelli, R., Lamonaca, D., Ceccato, E., Pileggi, F., Mazzi, F., Santonastaso, P., Ruggeri, M., Brambilla, P., 2017. Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques. NeuroImage 145, 238–245. https://doi.org/10.1016/j.neuroimage.2015.12.007.

Srimadhur, N.S., Lalitha, S., 2020. An end-to-end model for detection and assessment of depression levels using speech. In: Procedia Comput. Sci., Third International Conference on Computing and Network Communications (CoCoNet'19), 171, pp. 12–21. https://doi.org/10.1016/j.procs.2020.04.003.

Taguchi, T., Tachikawa, H., Nemoto, K., Suzuki, M., Nagano, T., Tachibana, R., Nishimura, M., Arai, T., 2018. Major depressive disorder discrimination using vocal acoustic features. J. Affect. Disord. 225, 214–220. https://doi.org/10.1016/j.jad.2017.08.038.

Tjandra, A., Sakti, S., Nakamura, S., 2020. Machine speech chain. IEEEACM Trans. Audio Speech Lang. Process. 28, 976–989. https://doi.org/10.1109/TASLP.2020.2977776.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M., 2016. Avec 2016: depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 3–10. https://doi.org/10.1145/2988257.2988258.

Vázquez-Romero, A., Gallardo-Antolín, A., 2020. Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy 22, 688. https://doi.org/10.3390/e22060688.

Wichers, M., 2014. The dynamic nature of depression: a new micro-level perspective of mental disorder that meets current challenges. Psychol. Med. 44, 1349–1360. https://doi.org/10.1017/S0033291713001979.

World Health Organization, 2017. Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization.

Yadav, U., Sharma, A.K., 2021. Review on automated depression detection from audio visual clue using sentiment analysis. In: 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, pp. 1462–1467. https://doi.org/10.1109/ICESC51422.2021.9532751.

Zhao, Y., Liang, Z., Du, J., Zhang, L., Liu, C., Zhao, L., 2021. Multi-head attention-based long short-term memory for depression detection from speech. Front. Neurorobotics 111. https://doi.org/10.3389/fnbot.2021.684037.