

**Automatic depression recognition by intelligent speech signal processing
A systematic survey**

Wu, Pingping; Wang, Ruihao; Lin, Han; Zhang, Fanlong; Tu, Juan; Sun, Miao

DOI

[10.1049/cit2.12113](https://doi.org/10.1049/cit2.12113)

Publication date

2022

Document Version

Final published version

Published in

CAAI Transactions on Intelligence Technology

Citation (APA)

Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., & Sun, M. (2022). Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*, 8(3), 701-711. <https://doi.org/10.1049/cit2.12113>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



REVIEW

Automatic depression recognition by intelligent speech signal processing: A systematic survey

Pingping Wu¹ | Ruihao Wang² | Han Lin¹ | Fanlong Zhang² | Juan Tu³ | Miao Sun⁴

¹Jiangsu Key Laboratory of Public Project Audit, School of Engineering Audit, Nanjing Audit University, Nanjing, China

²School of Information Engineering, Nanjing Audit University, Nanjing, China

³Key Laboratory of Modern Acoustics (MOE), School of Physics, Nanjing University, Nanjing, China

⁴Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, Delft, The Netherlands

Correspondence

Han Lin, Jiangsu Key Laboratory of Public Project Audit, School of Engineering Audit, Nanjing Audit University, Nanjing 211815, China.

Email: linhan@nau.edu.cn

Juan Tu, Key Laboratory of Modern Acoustics (MOE), School of Physics, Nanjing University, Nanjing 210093, China.

Email: juantu@nju.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 61701243, 71771125; Major Project of Natural Science Foundation of Jiangsu Education Department, Grant/Award Numbers: 19KJA180002, 21KJA630001

Abstract

Depression has become one of the most common mental illnesses in the world. For better prediction and diagnosis, methods of automatic depression recognition based on speech signal are constantly proposed and updated, with a transition from the early traditional methods based on hand-crafted features to the application of architectures of deep learning. This paper systematically and precisely outlines the most prominent and up-to-date research of automatic depression recognition by intelligent speech signal processing so far. Furthermore, methods for acoustic feature extraction, algorithms for classification and regression, as well as end to end deep models are investigated and analysed. Finally, general trends are summarised and key unresolved issues are identified to be considered in future studies of automatic speech depression recognition.

1 | INTRODUCTION

Depression has become a global health crisis in recent years with younger and faster growth and wider coverage. According to the data of World Health Organization, it is estimated that 5% of adults suffer from the disorder globally. Over 300 million people in the world have depression while over 54 million in China [1]. However, only 10% of depressed patients seek medical treatment in China while COVID-19 pandemic brings more challenges [2–4]. Depression can increase the risk of suicide in severe cases. People with depression are 20 times more likely to commit suicide [5]. In addition, depression has become the fourth leading cause of death among people aged 15–29 [6]. Accordingly, depression not only burdens patients

with a heavy financial burden, causing huge losses to individuals but also affects families and communities, and hinders the sustainable development of nations.

At present, the diagnosis of depression is usually made by questionnaires such as the Hamilton Rating Scale for Depression (HAM-D) [7], the Beck Depression Inventory-II (BDI-II) [8], the Patient Health Questionnaire (PHQ) [9], the Quick Inventory of Depressive Symptomatology [10], the Youth Mania Rating Scale (YMRS) [11], the Montgomery Åsberg Depression Rating Scale (MADRS) [12], and the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) [13]. Besides, objective physiological indicators are supplemented. However, such diagnostic methods rely on patient's cooperative attitude, expressiveness, and familiarity with the questionnaire. At the

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

same time, a large amount of clinical data is also required to support the diagnosis. What's more, there are other types of depression in addition to the most common major depressive disorder (MDD), further increasing the difficulty of diagnosis. Misdiagnosing the depression type or using the wrong treatment may delay or even worsen the patient's condition. Therefore, it is of great significance to find accurate, effective and objective diagnostic features for different types of depression.

Recently, automatic depression recognition and analysis has received extensive attention in the fields of medicine, psychology, and computer science. Since depressed patients and normal people behave differently in facial expressions, body postures, speech signal, physiological signals, and audio, researchers have tried to collect and analyse above information of depressed patients to predict the level of depression. In this article, we focus on automatic depression recognition by intelligent processing of speech signal as speech signal can reflect the depression tendency of subjects with slowed speech rate, prolonged pause, and different pitch changes [12, 14]. Specifically, in acoustics, the fundamental frequency F0 of the patient has a limited change while the second formant F2 is significantly reduced and the degree of variation in low frequency spectrum is decreased. Hence, the correlation of speech acoustic features with depression makes it a reliable objective marker for depression assessment [15]. Moreover, speech signal can be acquired non-intrusively, remotely, and the cost is relatively low, speech depression recognition (SDR) is easier to start than the research based on other modalities. In specific, SDR has undergone a transition from the early traditional methods based on hand-crafted features to the application of architectures of deep learning, from the usage of only acoustic features to the current application of multiple features [16–18]. However, there is still a lack of a research review to systematically and precisely sort out methods of automatic depression recognition by intelligent speech signal processing so far.

On the basis of extensive literature reading, this paper makes a systematic and in-depth summary of SDR, and gives an overview of the development history of the methods used in different stages. The main contributions of this paper are: (1) Introduce and sort out the most prominent and up-to-date literature in SDR in recent years in chronological order; (2) Investigate major trends in SDR and analyse their corresponding pros and cons; (3) Explore promising research directions for SDR in the future. The rest of this paper is organised as follows: Section II provides a detailed description and discussion of research evolution; Section III is about public datasets employed for automatic depression recognition. Section V gives conclusion and future work of SDR.

2 | RESEARCH EVOLUTION

Acoustic signal processing and machine learning technology jointly push the development of SDR. Figure 1 shows the development history accordingly. As proved in previous works, the acoustic features of depressed patient are different from

healthy individuals [19–21]. Therefore, in the early stage of the studies of SDR, the main work is to learn acoustic features related with depression and explore feature set for better performance [22, 23]. In the meantime, traditional machine learning algorithms are employed in SDR such as Support Vector Machine (SVM) [24–27], Hidden Markov Model [28], Gaussian Mixture Model (GMM) [27, 29, 30], K-means [31, 32], Boosting Logistic Regression [33–35], multi-layer perceptron [30, 35], etc.

In recent years, deep learning methods have made breakthroughs in the research fields of both Computer Vision (CV) and Nature Language Processing (NLP). Therefore, many studies have shifted from the traditional hand crafted acoustic features to the framework based on deep learning for SDR [36]. There are two application ways for deep learning methods in this field. One is to extract hand-crafted features from speech signals and then input them into deep neural network [37], where deep framework is only used as classifier. The other is to apply an end-to-end deep architecture, which feeds the original audio signal or spectrum to deep network to learn high-level features automatically [38]. As it could solve the problems encountered in hand-crafted features, such as high threshold, labour cost and low feature utilization rate, deep learning slowly becomes the leader in the field of machine learning. In addition, different neural network architectures are employed such as Convolutional Neural Networks (CNN) [39], Recurrent Neural Networks (RNN) [40], Long Short-Term Memory networks (LSTMs) [41], and Transformer [42].

However, in recent study, speech signal processing has received renewed attention because vocal features capture psychomotor activity associated with depression. Specifically, depressed patients have worse vocal tract coordination, so vocal tract variables and articulatory coordination features can be represented by channel delay correlation matrix, improving recognition performance effectively in SDR [43, 44]. Accordingly, acoustic features combined with deep learning become the most popular architectures in SDR.

2.1 | Speech depression recognition based on hand-crafted features

2.1.1 | Extraction of hand-crafted features

As mentioned in previous section, depressed patients have cognitive and psychomotor differences compared to normal people. Owing to the sensibility of speech, slightly physical or cognitive change could result in obvious acoustic change [20, 45, 46]. In earlier research of SDR, low-level acoustic features are regularly used together with statistical features while some feature extraction tools are employed to extract features directly such as COVAREP, OpenSMILE. The commonly used acoustic features are as follows:

Prosodic features include changes in pitch and loudness, as well as changes in the length of syllables, words and phrases [33, 47–49]. Among them, fundamental frequency (F0) and energy are used to represent pitch and loudness perception characteristics [50].

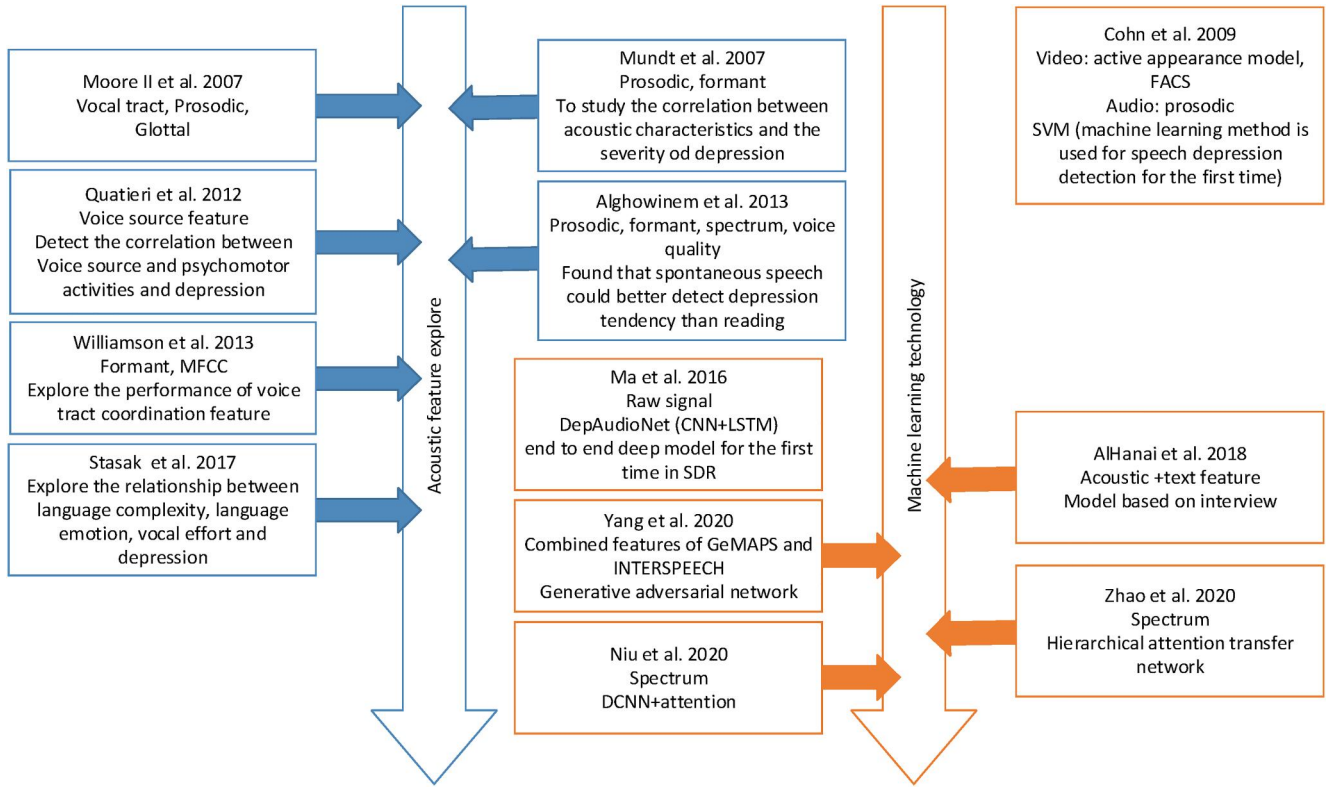


FIGURE 1 Development history of speech depression recognition (SDR)

Voice quality features capture characteristic information of speech generating sources. These features can parameterise the airflow from lung to glottis and vocal tract motion including: Normalised Amplitude Quotient (NAQ), Quasi Open Quotient, Harmonic Difference H1-H2 and H2-H4, spectrum perturbation and amplitude perturbation [33, 43, 49, 51].

Formant features include information about vocal tract resonance and pronunciation efforts, which reflect the characteristics of physical vocal tract. The first three formants (F1-F3) are usually used as formant characteristics in SDR [47, 50].

Spectral features represent the correlation between vocal tract shape changes and articulator movement including spectral flux, energy, slope and flatness [49, 52], Mel-Frequency Cepstrum Coefficient (MFCC) [33] and Linear Predictive Cepstrum Coefficient [49, 52, 53].

In recent years, acoustic features regain researchers' attention as it is found that fusion of acoustic features can improve the performance of SDR. As shown in Table 1, typical acoustic Low Level Descriptors (LLDs) and their statistic features in SDR are enumerated. Besides, Articulatory Coordination Features (ACF) has achieved great success in SDR by quantifying the time change of pronunciation action [43, 54]. By investigating the correlation between MFCC and formant, Williamson et al. achieve excellent recognition result in SDR [54]. Besides, some following studies show good prospect of ACF-based vocal tract variable features in SDR [55, 56].

2.1.2 | Classification algorithms

In early research of SDR, traditional classification or regression algorithms are employed after feature extraction such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest, Decision Tree, Gaussian Mixture Model (GMM), K-means, etc., which are as shown in Table 2.

Support Vector Machine is a classical machine learning algorithm based on statistic, which shows excellent performance in high-dimensional, small sample and nonlinear problems. Due to the high-dimensional extracted acoustic features, small-scale depressive speech dataset, SVM became the most popular classification algorithm in early research of SDR [24–27, 58]. For example, Gong et al. employed a topic modelling-based approach to explore context-relevant information in depressive data (audio, video, text) using a Support Vector Regressor (SVR) with three kernels (linear, polynomial, radial basis functions) for his prediction task [25]. However, this algorithm is slow to train and its performance is affected by the combination of kernel functions and model parameters, and results are slightly less interpretable.

Logistic Regression (LR) is also a common classification algorithm based on statistics. Despite similar to SVM, LR is usually applied in large-scale dataset and can only process discrete features. Due to its discretisation feature, logistic regression is chosen to use for its fast speed, strong robustness, easy crossover and feature combination [33–35]. Zaremba et al. designed an integrated algorithm based on logistic regression [33], which can preserve the diversity of feature subspace and

TABLE 1 Typical acoustic Low Level Descriptors (LLDs) and their statistic features in SDR

LLDs	Statistic features
Fundamental frequency (F0), energy, intensity, harmonic noise ratio (HNR), speech speed, mel frequency cepstrum coefficient (MFCC), formant amplitude, formant bandwidth, formant frequency, linear predictive cepstrum coefficient (LPCC), spectral slope, normalised amplitude quotient (NAQ), spectral perturbation (jitter), amplitude perturbation (shimmer)	Extreme value, maximum value, minimum value, average value, standard deviation, variance, kurtosis, skewness, percentage, percentage range, quartile, centre, deviation, slope, mean square error and duration

TABLE 2 Some traditional classification and regression algorithms applied in speech depression recognition (SDR)

Method	Paper	Dataset	Performance	Application scene
GMM	Helfer et al. 2013 [27]	Mundt-35	AUC 0.76	Suitable for no-label, large-scale dataset
	Williamson et al. 2013 [29]	AVEC2013	MAE/RMSE 5.75/7.42	
	Williamson et al. 2014 [60]	AVEC2014	MAE/RMSE 6.52/8.50	
SVM	Cummins et al. 2013 [26]	Mundt-35	Accuracy 66.9%	Deal with high dimensional, small sample issues
	Nasir et al. 2016 [24]	DAIC-WOZ	F1 0.63	
	Gong et al. 2017 [25]	DAIZ-WOZ	MAE/RMSE 3.96/4.99	
LR	Jan et al. 2017 [34]	AVEC2014	MAE/RMSE 6.14/7.43	Suitable for large-scale dataset, discrete features
	Jayawardena et al. 2020 [35]	DAIC-WOZ	RMSE 6.84	
Decision tree	Pampouchidou et al. 2016 [57]	DAIZ-WOZ	F1 (D/N)0.52/0.81	Not sensitive on the errors of the dataset

extract more discriminative features, which shows better performance than GMM, SVM, random forest, decision tree and AdaBoost.

Gaussian mixture model (GMM) uses Gaussian distribution as the parameter model and is trained by expectation maximum, which shows outstanding performance in no-label, large-scale dataset. As a clustering algorithm, it is employed in early research of SDR [27, 30, 59, 69]. Moreover, GMM-based regression methods such as Gaussian Staircase Regression (GSR) are proposed, where each GMM consists of an ensemble of Gaussian classifiers [29, 54, 61, 62]. In specific, firstly, speech features are mapped to different partitions of clinical depression score, then the mapping results are used as the basis of regression analysis.

2.2 | Speech depression recognition based on deep learning

Due to the successful application in CV and NLP, deep learning is introduced to SDR. Compared with traditional methods, no human intervention is needed after the model and parameters are determined. The essence of deep learning is to learn high-level abstract features automatically by building more hidden layer models to improve the accuracy of classification or score prediction.

There are two ways to employ deep learning in SDR: (1) Build a structure combined acoustic features with deep learning method. Traditional acoustic features or deep acoustic features are then put into the deep classifier for training, recognition or prediction. When used as a feature extractor, deep learning can avoid high labour cost and large-scale loss of

feature, and the extensibility is better than traditional method. When used as a classifier, deep classifiers have many advantages, including dealing with complex structures and functions, and unlabelled and incorrectly labelled data. (2) Build an end-to-end deep architecture and then push raw signal or spectrogram into deep architecture to let model learn high-level features by itself.

2.2.1 | Deep learnt features

In Speech Emotion Recognition, deep speech features through pre-trained deep network have made remarkable performance and are robust to noise changes [63, 64]. Accordingly, deep features are employed in SDR. Yang et al. [65] used DCNN-DNN to forecast depression severity score. Firstly, push multi-modal features in DCNN to learn high global features with tight dynamic information. Then lead these features in DNN to forecast PHQ-8 score. Finally, the PHQ-8 scores of each mode are fused to obtain the final result. Dong et al. learnt the deep feature from the original signal and spectrum through ResNET, and then calculated the correlation coefficient of delay multi-channel change with the Feature Variation Coordination Measurement algorithm to obtain the coordination feature and learn the time information of the deep feature [66]. Seneviratne et al. designed a double-layer neural network architecture of dilated CNN-LSTM [43]. In the first layer, dilated convolution neural network (dilated CNN) was used to extract articulatory co-ordination features (ACF). In the second layer, the channel delay matrix was constructed to solve the problems of repeated sampling in traditional methods and discontinuity on

the boundary of adjacent sub matrix. Also, Huang et al. used the model of all channel coordination convolution neural network (FVTC-CNN) to predict depression, in which the expanded convolution neural network was used to extract the characteristics of channel coordination [44].

Recently, auto-encoder shows its good prospect in SDR as a deep feature extractor [38, 65, 67–70]. Auto-encoder consists of two parts: encoder and decoder, encoder is used to learn the abstract features of the input data, and the function of the decoder is to remap the abstract features back to the original space to obtain the reconstructed data. The optimization goal is to optimise the model by minimising the reconstruction error to learn the abstract features of the input data. The advantage of automatic encoder is that it belongs to unsupervised learning and does not rely on annotation of data. Therefore, automatic encoder can be regarded as an unsupervised nonlinear dimensionality reduction feature extraction method. Then several improved models are proposed for different scenarios, including denoising auto-encoder [71], sparse auto-encoder [72], convolution auto-encoder [73], variational auto-encoder [74], adversarial auto-encoder [75]. For example, Sardari S., et al. extracted deep features from speech depression data by convolution auto-encoder [76]. Due to the outstanding performance of processing local data, convolution auto-encoder has stronger feature learning ability than auto-encoder, which also solves the problem of sample imbalance in the data set by resampling method based on clustering. The experimental results show that the recognition effect is better than the previous ensemble CNN, DepAudioNet, SVM and other methods.

2.2.2 | Deep classifiers

In SDR, the commonly used deep classifier algorithms including Recurrent Neural Network (RNN), Deep Belief Network, Convolution Neural Network (CNN), etc. As Table 3 shows.

In specific, CNN could capture spatial properties of features and has the ability of parallel computing. Therefore, CNN can be used as a classifier for MFCC, spectrum or some other deep learnt features [38, 44, 68, 77, 83–86]. Aiming at characteristics of depressed patients with more speech pauses and slower speech speed, and the problem that LSTM does not perform well in long sequences, Haque et al. proposed causal convolution neural network (C-CNN) to deal with audio [78], text and video data to get multi-modal sequence-level feature instead of LSTM. In addition, ensemble learning can improve recognition performance by combining multiple models and the performance of ensemble convolutional neural network model (integrating 50 one-dimensional convolutions) is also utilised in SDR [85]. The research shows that the effect of integrating CNN is significantly better than normal CNN method when the convolution kernel size is appropriate ((1, 3), (1, 5), (1, 7)). In recent, Niu et al. proposed a CNN model based on attention mechanism, namely time-

frequency channel attention (TFCA) block [86], which is used to emphasise the timestamp, frequency band and channel related to depression detection. TFCA block solved the problem that CNN global pooling cannot consider time domain information of data. Although CNN is favoured by researchers because of its excellent characteristics such as local connection, weight sharing, pooling operation and multi-layer structure, but at the same time, it should also realize its training difficulty and performance problems in very deep networks.

RNN is a network based on sequence information, where adjacent information is interdependent. Normally, this interdependence is useful in predicting the future state. Like CNN, it was born at the end of the last century. The great brilliance of RNN in deep learning originated from [87] while LSTMs is the most common RNN model in SDR. It avoids problems such as gradient disappearance to a certain extent, and can relatively learn information of long time series, so it is suitable for time series data such as speech. Since deep learning methods have been popularised in the field of SDR, a number of RNN-based studies have been carried out [37, 38, 77, 88, 89]. Alhanai et al. employed LSTM to detect depression with Audio/Text feature and came to a conclusion that the performance of context-free model is better than context-weighted model [37]. Du et al. proposed a novel LSTM module, namely IncepLSTM [88], by combining inception module and LSTM to adapt to the situation that bipolar disorder occurs irregularly in different time periods. However, RNN-related algorithms have high time cost due to their poor parallel ability, and RNN cannot be able to cope with data that is too long.

2.2.3 | End to end deep architectures

Compared with methods of performing feature extraction and classification separately, an end to end deep architecture pushes raw signal or spectrogram into its model to learn and give results as shown in Figure 2. End to end deep architecture have advantages like that it does not require scholars to have a priori knowledge, deep networks can learn better features and give better classification result. However, there are a few issues which limit end-to-end deep architectures, such as large-scale data supporting, overfitting easily and poor interpretability.

Ma et al. designed a deep model combined DCNN with LSTM instead of previous SDR method based on acoustic feature [36], named DepAudioNet. In their research, CNN is applied to extract high-level feature from raw wave while LSTM is used to learn the temporal change of Mel scale filter feature, which achieved good results on the DAIC-WOZ data set, and also strongly promoted the follow-up end-to-end model research.

Othmani et al. designed a deep neural network architecture called EmoAudionet [77], input the preprocessed spectrum into DCNN network and combine it with CNN network based on MFCC features to improve performance. The results show that the accuracy of EmoAudionet training in DAIC-WOZ is 73.25%, while the F1 score is 82%.

TABLE 3 Some deep classifiers applied in speech depression recognition (SDR) and their performance

Method	Paper	Dataset	Performance	Application scene
LSTM	Alhanai et al. 2018 [37]	DAIC-WOZ	MAE/RMSE 4.97/6.27	Suitable for time series issues
	Du et al. 2018 [88]	BD	UAR/UAP/Accuracy 0.651/0.678/65.0%	
	Salekin et al. 2018 [89]	DAIC-WOZ	F1/Accuracy 0.901/90%	
	Othmani et al. 2021 [77]	DAIC-WOZ	F1 (D/N) 0.49/0.82 accuracy 73.35%	
	Zhang et al. 2021 [38]	DAIC-WOZ	MAE/RMSE 5.48/6.31	
CNN	Yang et al. 2017 [84]	DAIC-WOZ	MAE/RMSE 5.163/5.974	Deal with spatial–temporal issues
	Haque et al. 2018 [78]	DAIC-WOZ	F1/Precision/Recall 0.769/71.4%/83.3%	
	He et al. 2018 [79]	AVEC2013/14	MAE/RMSE 8.78/10.90	
	Huang et al. 2020 [44]	DAIC-WOZ	F1/Accuracy 0.700/82.9%	
	Muzammel et al. 2020 [83]	DAIC-WOZ	Accuracy/Precision/Recall/F1 86.06%/81%/73%/77%	
	Vázquez-Romero et al. 2020 [85]	DAIC-WOZ	F1/Accuracy/Precision/Recall 0.65/74%/55%/79%	
RNN	Niu et al. 2021 [86]	AVEC2013/14	MAE/RMSE (AVEC2013/2014) 6.01/8.15 7.00/8.96	Suitable for temporal sequence data
	Chao et al. 2015 [80]	AVEC2014	MAE/RMSE 7.91/9.98	
	Al et al. 2018 [81]	AVEC2013/14	MAE/RMSE 7.37/9.28	
GAN	Yang et al. 2020 [82]	DAIC-WOZ	MAE/RMSE 4.634/5.520	Generate additional data to avoid unbalanced or small samples
Transformer	Sun et al. 2021 [68]	E-DAIC	RMSE 3.783	Suitable for very long sequence data
	Zhang et al, 2021 [38]	DAIC-WOZ	MAE/RMSE 4.75/5.73	

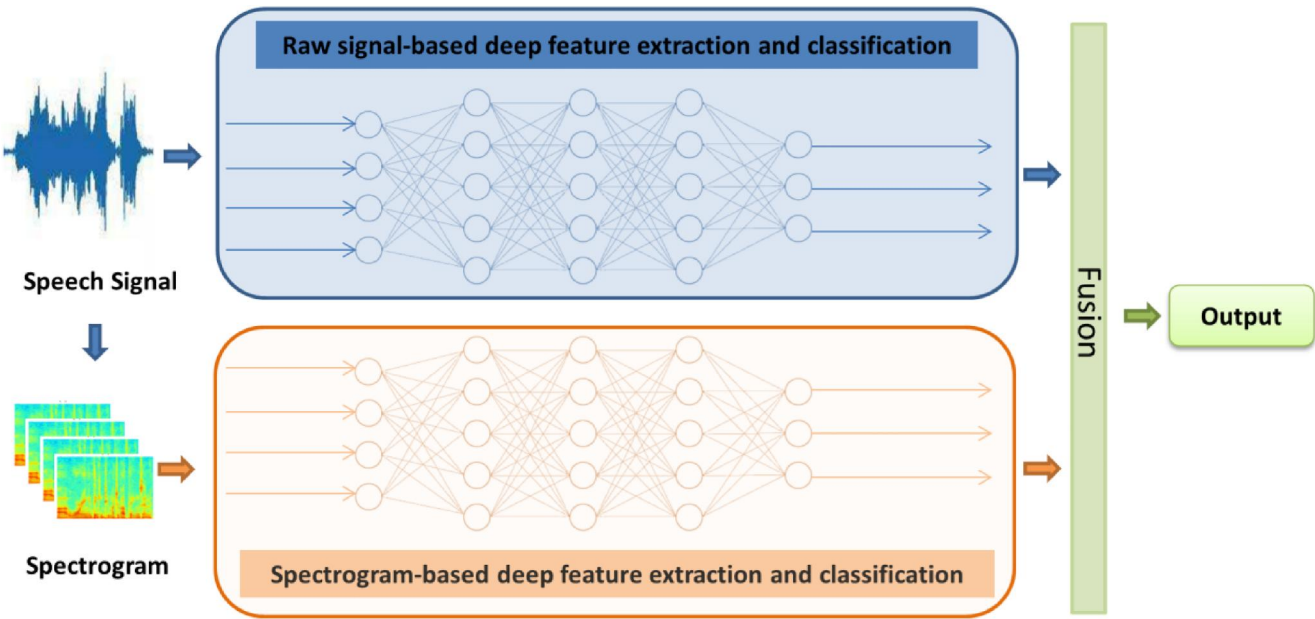


FIGURE 2 Framework of an end to end deep architecture

End-to-end deep model is difficult to determine the contribution of each module in the architecture due to its end-to-end characteristics, limiting further performance improvement. In a word, end-to-end deep architectures have not yet been widely used in the field of SDR because of its poor interpretability, flexibility and current limited dataset scale. For now, the most popular method is still the combination of acoustic features and deep classifiers.

3 | DATASETS

Different from speech emotion datasets, acted and evoked datasets are difficult to apply directly to SDR. Generally, speech depression datasets are recorded during the conversation from clinical doctor with depressed patients by face-to-face, telephone interviews or virtual interviewers. In some data collection processes, other modality information, such as depression scale data, facial expression data, physiological dynamic information, etc., are also recorded at the same time for auxiliary analysis.

Normally, a SDR dataset is composed of three parts: interactive interviews with subjects, descriptions of pictures, and recitations. Study finds that results of gender-specific SDR are affected by different parts of the data. For male subjects, descriptions of pictures performed best for SDR. However, for female subjects, interactive interviews performed best. Therefore, designing different data acquisition schemes for gender is an option worth considering. Moreover, it has been studied whether positive, neutral, or negative speech affects the result of SDR, but a unified conclusion has not yet been reached. Jiang et al. [90] believed that these three affective states had no significant effect for SDR. However, the results of the study [67] showed that the overall accuracy of SDR was reduced after removing negative speech. Therefore, further research is required to verify the association of different affective states with depression.

3.1 | Representative datasets

The collection of speech depression data is the basis for conducting research of SDR. Table 4 lists representative datasets in SDR.

AVEC2013 dataset and AVEC2014 dataset are the subset of Audio-Visual Depression Language Corpus. Particularly, the Audio-Visual Emotion Challenge (AVEC) is a competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and audio-visual emotion analysis. For AVEC2013, there are 340 videos in German, which are recorded when participants performed human-computer interaction tasks in front of webcam and microphone. Video files include free speech, reading, singing, and picture-seeing association tasks while BDI-II is used to annotate depression severity score of participant's interview records. For AVEC2014, it is a subset of AVEC2013, consisting of 300 videos in German, where duration of each video clip is shorter than the clip in AVEC2013.

Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) is a part of Distress Analysis Interview Corpus annotated by PHQ-8, employed for AVEC2016 & AVEC2017. Distress Analysis Interview Corpus - Wizard of Oz adopts a virtual interviewer as it is considered that being confronted with a virtual interviewer makes subjects more willing to speak out than a real person and emotion status of an interviewer needs to be strictly controlled during the interview. Audio, video and deep sensor modalities are collected in the dataset.

Besides, it also contains information of galvanic skin response (GSR), electrocardiogram (ECG), participants' respiratory data.

E-DAIC is an extended version of DAIC-WOZ which is collected from semi-clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety and depression [97]. The dataset contains 163 development samples, 56 training samples and 56 test samples, and the participants' data are marked with age, gender and PHQ-8 score is labelled. This database is employed for AVEC2019 [98].

The Bipolar Corpus, a new Turkish Audio-Video Bipolar Disorder Corpus, is collected by Elvan et al. used for effective computing and psychiatric research. This corpus is also employed for bipolar disorder sub-challenge of AVEC2018, which is annotated by Youth Mania Rating Scale (YMRS). Videos of the dataset are recorded under seven tasks to describe the state of bipolar disorder, such as explaining why you went to the hospital, attending an activity, describing happy and sad memories, counting to 30, describing two pictures that evoke emotions, etc.

MODMA is the first Chinese multi-modality depression database available to our knowledge, including participant's audio and EEG information from 23 patients and 29 healthy. Audio information is recorded during the tasks of an 18-question interview from the Depression Scale, read aloud and a description of emotional pictures. Table 5 shows the results of some benchmark works on the representative datasets measured by F1 score, mean absolute error (MAE) and root mean square error (RMSE).

3.2 | Dataset annotation

Depression dataset annotation is an important and difficult task, the accuracy of which has a direct impact on the follow-up research. A complete annotation of speech depression database is normally consisted of three parts: transcription, analysis and annotation. Transcription is to transcribe audio and linguistic information into text form; Analysis is to further mark the acoustic features such as prosodic information, speech speed, volume and tone changes on the basis of transcription; Annotation is to mark the depression score of a sentence. AVEC 2013 is annotated by BDI-II from the two dimensions of Arousal and Valence. In DAIC-WOZ, the ELAN tool is used in the transcription part, where psychological states of the interviewee as well as the content of the dialogue and non-verbal behaviours are analysed and annotated; For MODMA, all recordings were manually segmented and annotated by Statistical Manual of Mental Disorders (DSM-IV). Table 6 includes the range of score and the corresponding depression level for different questionnaires.

3.3 | Existing issues

Although SDR has made progress with existing datasets, the following issues with datasets hinder the further development.

TABLE 4 Representative datasets in speech depression recognition (SDR)

Dataset	Modality	Label	Number of subjects	Number of clips	Duration
Mundt-35 (2007) [91]	Audio	HAMD QIDS	35 patients	-	-
AVEC2013 [92]	Audio/Video	BDI-II	84 patients	150 chips	20–50m
AVEC2014 [93]	Audio/Video	BDI-II	84 patients	300 clips	6s–4 m
DAIC-WOZ (2014) [94]	Audio/Video/ECG/GSR	PHQ-8	189 patients	189 clips	Wizard-of-Oz 5–20m, automated agent 15–25m
E-DAIC (2014) [97]	Audio/Video	PHQ-8	351 patients	275 clips	-
Bipolar corpus (2018) [95]	Audio/Video	YMRS MADRS	46 depressed 49 control	218 clips	At most 3. 7m
MODMA (2020) [96]	Audio/EEG	HRSD DSM-IV	23 depressed 29 control	1508 clips	At most 2.45 m

Datasets	Methods	F1	MAE	RMSE
AVEC2013	Correlation structure features + GSR (2013) [29]	-	5.75	7.42
AVEC2014	Hand/Deep features + DCNN (2018) [38]	-	8.1919	9.9998
	Spectrogram + STA (2020) [99]	-	7.65	9.13
DAIC-WOZ	DepAudioNet (2016) [36]	0.52	-	-
	Audio/Text LSTM with topic modelling (2018) [37]	-	4.97	6.27
	Spectrum features + HATN (2020) [100]	-	4.28	5.66
	DCGAN generated features + DCNN(2020) [82]	-	4.634	5.520
E-DAIC	Multi-layer attention network on A/V/T features (2019) [101]	-	-	4.28
MODMA	Multi-head time-dimension attention-based LSTM (2021) [102]	0.987	-	-

TABLE 5 Performance of different methods on the representative datasets**TABLE 6** Depression rating for different questionnaires

	Normal	Mild	Moderate	Severe	Very severe
HAM-D [7]	0–7	8–13	14–18	19–22	≥23
BDI-II [8]	0–13	14–19	20–28	29–63	-
PHQ-8 [103]	0–4	5–9	10–14	15–19	20–24
PHQ-9 [9]	0–4	5–9	10–14	15–19	20–27
PHQ-15 [104]	1–4	5–9	10–14	15–30	-
QIDS [10]	0–5	6–10	11–15	16–20	≥21
YMRS [11]	0–5	6–12	13–19	20–29	≥30
MADRS [12]	0–11	12–22	23–30	31–35	≥36
DSM-IV [13]	-	-	11–15	16–20	≥21

- (1) Objectivity of database annotation: Data annotation is the basis of further work, however, the cognition of annotators is not completely accurate, and the distribution of depression scores will affect the performance of the constructed model.
- (2) Unavailability and small in scale: Due to the sensibility of depression speech and the ethics problem, most institutions could not obtain sufficient samples. At present, public depression databases available are AVEC2013, AVEC2014, DAIC-WOZ and BD, which are far from needs of scientific research. Addressing ethical issues is important for the publication of datasets.

- (3) Non- universality: Currently, datasets employed in SDR research come normally from interactive clinical interview, in which the questions are carefully designed and there is no noise interference like in real life. Therefore, these data cannot fully reflect the normal life state of patients with depression. Besides, the issue of cross language and cultural has not yet been considered.

4 | CONCLUSION AND FUTURE WORK

Depression is a common mental disorder, the effective and accurate diagnosis of which requires coordinated efforts among clinical psychology, brain science, affective computing and other fields. It is of great significance for both academic research and clinical care to develop an automatic and objective evaluation system. This paper systematically and comprehensively sorts out depression recognition based on intelligent speech signal processing. As stated in the paper, it can be found that the research of SDR has undergone a shift from exploring acoustic features to deep model research. At present, CNN and LSTM have become the most popular deep models due to their advantages in processing spatiotemporal features. In order to better apply methods of deep learning, it is increasingly important to collect large-scale unified data. Although great progress has been made in the field of SDR, there is still a long way for it to be put into practical use. To achieve a breakthrough, the following challenges must be considered and overcome.

- (1) Availability and limitations of the baseline dataset: Database building is the basis of the research. However, there are some restrictions of the existing databases caused by different collection scenarios and methods, inconsistent labelling, small data scale, and non-disclosure due to privacy. It is a key to breakthroughs in depression analysis based on speech signal to create a large-scale database with open standards, accurate and consistent labelling, cross-cultural and cross-language.
- (2) Model generalization: Most studies are limited to a single or a few small-scale datasets, which makes the models perform poorly when faced with other datasets or data from other languages. Therefore, it is also a necessary study to improve the model generalization and robustness across corpora, cultures, languages, and under noisy environments.
- (3) Unknown underlying correlation mechanism of acoustic information: The medical mechanism of depression on speech is a prerequisite for machine learning-based depression analysis research. To further improve the recognition accuracy, it is necessary to collect and extract clinical information on depression. Therefore, the following research should increase the communication and cooperation with other relevant professionals. It is a long-term and important topic to explore the underlying acoustic mechanism of speech in depression.
- (4) Different types of depression: As mentioned before, there are different types of depression. For example, bipolar disorder differs from the most common MDD in terms of pathogenesis and performance. So far, little research has been done on the difference between the two through the speech signal.
- (5) Multi-modality fusion mechanism: Combining multiple modalities for accurate and effective depression analysis is an inevitable trend in future research, because different modalities can effectively complement each other. However, the success of multimodal research is based on an effective and appropriate fusion mechanism.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (NSFC, no.61701243, 71771125) and the Major Project of Natural Science Foundation of Jiangsu Education Department (no.19KJA180002).

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Han Lin  <https://orcid.org/0000-0001-5136-5059>

REFERENCES

1. World Health Organization: Depression. <https://www.who.int/zh/news-room/fact-sheets/detail/depression>
2. World Health Organization: Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization (2017)
3. Lu, J., et al.: Prevalence of depressive disorders and treatment in China: a cross-sectional epidemiological study. *Lancet Psychiatr.* 8(11), 981–990 (2021)
4. Santomauro, D.F., et al.: Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet.* 398(10312), 1700–1712 (2021)
5. Lépine, J.P., Briley, M.: The increasing burden of depression. *Neuropsychiatric Dis. Treat.* 7, 3–7 (2011). <https://doi.org/10.2147/ndt.s19617>
6. Department of Health, et al.: Healthy People 2010: Understanding and Improving Health. US Department of Health and Human Services, Washington (2000)
7. Hamilton, M.: The Hamilton rating scale for depression. In: *Assessment of Depression*, 143–152. Springer (1986)
8. Beck, A.T., Steer, R.A., Brown, G.K.: Manual for Beck Depression Inventory-II. Psychological Corporation, San Antonio (1996)
9. Kroenke, K., Spitzer, R.L., Williams, J.B.: The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16(9), 606–613 (2001)
10. Rush, A.J., et al.: The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol. Psychiatr.* 54(5), 573–583 (2003)
11. Young, R.C., et al.: A rating scale for mania: reliability, validity and sensitivity. *Brit. J. Psychiatr.* 133(5), 429–435 (1978)
12. Montgomery, S.A., Åsberg, M.A.R.I.E.: A new depression scale designed to be sensitive to change. *Brit. J. Psychiatr.* 134(4), 382–389 (1979)
13. Edition, F.: Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc.* 21, 591–643 (2013)
14. Pampouchidou, A., et al.: Facial geometry and speech analysis for depression detection. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 1433–1436. IEEE (2017)
15. Beck, A.T., et al.: Comparison of Beck depression inventories-IA and-II in psychiatric outpatients. *J. Pers. Assess.* 67(3), 588–597 (1996)
16. Cummins, N., et al.: A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* 71, 10–49 (2015). <https://doi.org/10.1016/j.specom.2015.03.004>
17. Morales, M.R., Scherer, S., Levitan, R.: A cross-modal review of indicators for depression detection systems. In: *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology*, 1–12. ACL (2017)
18. He, L., et al.: Deep learning for depression recognition with audiovisual cues: a review. *Inf. Fusion.* 80, 56–86 (2021). <https://doi.org/10.1016/j.inffus.2021.10.012>
19. Scherer, S., et al.: Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. In: *Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 4789–4793. IEEE (2015)
20. Mundt, J.C., et al.: Vocal acoustic biomarkers of depression severity and treatment response. *Biol. Psychiatr.* 72(7), 580–587 (2012)
21. Tolkmitt, F., et al.: Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *J. Commun. Disord.* 15(3), 209–222 (1982)
22. Degottex, G., et al.: COVAREP—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, 960–964. IEEE (2014)
23. Huang, Z., Epps, J.: An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech. *IEEE. T. Affect. Comput.* 11(4), 653–668 (2018)
24. Nasir, M., et al.: Multimodal and multiresolution depression detection from speech and facial landmark features. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 43–50. ACM (2016)
25. Gong, Y., Poellabauer, C.: Topic modeling based multi-modal depression detection. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 69–76. ACM (2017)

26. Cummins, N., Epps, J., Ambikairajah, E.: Spectro-temporal analysis of speech affected by depression and psychomotor retardation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 7542–7546. IEEE (2013)
27. Helfer, B.S., et al.: Classification of depression state based on articulatory precision. In: *Interspeech*, vol. 2013, pp. 2172–2176. ISCA (2013)
28. Yang, T.H., et al.: Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. *J. Ambient Intell. Hum. Comput.* 8(6), 895–906 (2017)
29. Williamson, J.R., et al.: Vocal biomarkers of depression based on motor incoordination. In: *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 41–48. ACM (2013)
30. Alghowinem, S., et al.: A comparative study of different classifiers for detecting depression from spontaneous speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 8022–8026. IEEE (2013)
31. Sun, Y., et al.: Classification of negative emotion speech intensity based on similarity algorithm. In: 2018 IEEE International Conference on Information Communication and Signal Processing (ICICSP), 4–97. IEEE (2018)
32. Morales, M.R., Levitan, R.: Mitigating confounding factors in depression detection using an unsupervised clustering approach. In: *Computing and Mental Health Workshop*, 1–4. CHI (2016)
33. Jiang, H., et al.: Detecting depression using an ensemble logistic regression model based on multiple speech features. *Comput. Math. Method. M.* 2018, 1–9 (2018). <https://doi.org/10.1155/2018/6508319>
34. Jan, A., et al.: Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE. T. Cogn. Dev. Syst.* 10(3), 668–680 (2017)
35. Jayawardena, S., Epps, J., Ambikairajah, E.: Ordinal logistic regression with partial proportional odds for depression prediction. *IEEE. T. Affect. Comput.* (2020)
36. Ma, X., et al.: DepAudioNet: an efficient deep model for audio based depression classification. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 35–42. ACM (2016)
37. Alhanai, T., Chassei, M., Clss, J.: Detecting depression with audio/text sequence modeling of interviews. In: *Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages*, 1716–1720. ISCA (2018)
38. Zhang, P., et al.: Depa: self-supervised audio embedding for depression detection. In: *Proceedings of the 29th ACM International Conference on Multimedia*, 135–143. ACM (2021)
39. Lecun, Y., et al.: Gradient-based learning applied to document recognition. *Proc. IEEE.* 86(11), 2278–2324 (1998)
40. Elman, J.L.: Finding structure in time. *Cognit. Sci.* 14(2), 179–211 (1990)
41. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* 9(8), 1735–1780 (1997)
42. Vaswani, A., et al.: Attention is all you need. In: *The 31st Conference on Neural Information Processing System*, 5998–6008. NIPS (2017)
43. Seneviratne, N., Espy-Wilson, C.: Speech based depression severity level classification using a multi-stage dilated CNN-LSTM model. *arXiv preprint arXiv:2104.04195* (2021)
44. Huang, Z., Epps, J., Joachim, D.: Exploiting vocal tract coordination using dilated CNNs for depression detection in naturalistic environments. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6549–6553. IEEE (2020)
45. Scherer, S., et al.: Investigating voice quality as a speaker-independent indicator of depression and ptsd. In: *Interspeech 2013 14th Annual Conference of the International Speech Communication Association*, 847–851. ISCA (2013)
46. Hönig, F., et al.: Automatic modelling of depressed speech: relevant features and relevance of gender. In: *Interspeech*, vol. 2014, pp. 1248–1252. ISCA (2014)
47. France, D.J., et al.: Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE. T. Bio-Med. Eng.* 47(7), 829–837 (2020)
48. Cohn, J.F., et al.: Detecting depression from facial actions and vocal prosody. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 1–7. IEEE (2009)
49. Ringeval, F., et al.: Avec 2017: real-life depression, and affect recognition workshop and challenge. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 3–9. ACM (2017)
50. Valstar, M., et al.: Avec 2016: depression, mood, and emotion recognition workshop and challenge. In: *Proceedings of the 6th International Workshop on Audio/visual Emotion Challenge*, vol. 2016, pp. 3–10. ACM (2016)
51. Moore, E., II, et al.: Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE. T. Bio-Med. Eng.* 55(1), 96–107 (2007)
52. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Assessing speaker independence on a speech-based depression level estimation system. *Pattern Recogn. Lett.* 68, 343–350 (2015). <https://doi.org/10.1016/j.patrec.2015.05.017>
53. Moore, E.I.I., et al.: Comparing objective feature statistics of speech for classifying clinical depression. In: *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 17–20. IEEE (2004)
54. Williamson, J.R., et al.: Tracking depression severity from audio and video based on speech articulatory coordination. *Comput. Speech Lang.* 55, 40–56 (2019). <https://doi.org/10.1016/j.csl.2018.08.004>
55. Espy-Wilson, C.Y., et al.: Assessing neuromotor coordination in depression using inverted vocal tract variables. In: *Interspeech*, vol. 2019, pp. 1448–1452. ISCA (2019)
56. Seneviratne, N., et al.: Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression. In: *Interspeech*, vol. 2020, pp. 4551–4555. ISCA (2020)
57. Pampouchidou, A., et al.: 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*, pp. 27–34. ACM (2016)
58. Cummins, N., et al.: Diagnosis of depression by behavioural signals: a multimodal approach. In: *Proceedings of the 3rd International Workshop on Audio/visual Emotion Challenge*, 11–20. ACM (2013)
59. Low, L.S.A., et al.: Detection of clinical depression in adolescents’ speech during family interactions. *IEEE. T. Bio-Med. Eng.* 58(3), 574–586 (2010)
60. Williamson, J.R., et al.: Vocal and facial biomarkers of depression based on motor incoordination and timing. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 65–72. ACM (2014)
61. Williamson, J.R., et al.: Detecting depression using vocal, facial and semantic communication cues. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 11–18. ACM (2016)
62. Cummins, N., et al.: Generalized two-stage rank regression framework for depression score prediction from speech. *IEEE. T. Affect. Comput.* 11(2), 272–283 (2017)
63. Cummins, N., et al.: An image-based deep spectrum feature representation for the recognition of emotional speech. In: *Proceedings of the 25th ACM International Conference on Multimedia*, 478–484. ACM (2017)
64. Chen, S., et al.: Multimodal multi-task learning for dimensional and continuous emotion recognition. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 19–26. ACM (2017)
65. Yang, L., et al.: Hybrid depression classification and estimation from audio video and text information. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 45–51. ACM (2017)
66. Dong, Y., Yang, X.: A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing.* 441, 279–290 (2021). <https://doi.org/10.1016/j.neucom.2021.02.019>
67. Dibekioğlu, H., Hammal, Z., Cohn, J.F.: Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE. J. Biomed. Health.* 22(2), 525–536 (2017)

68. Sun, H., et al.: Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sensors*. 21(14), 47–64 (2021)
69. Zhao, Z., et al.: Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. *IEEE. J-STSP*. 14(2), 423–434 (2019)
70. Harati, A., et al.: Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 7273–7277. IEEE (2021)
71. Vincent, P., et al.: Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*, 1096–1103. IMLS (2008)
72. Ng, A.: Sparse autoencoder. *CS294A lect. Notes*. 72, 1–19 (2011)
73. Masci, J., et al.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: *International Conference on Artificial Neural Networks*, pp. 52–59. ENNS (2011)
74. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
75. Makhzani, A., et al.: Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015)
76. Sardari, S., et al.: Audio based depression detection using Convolutional Autoencoder. *Expert Syst. Appl.* 189, 116076 (2022). <https://doi.org/10.1016/j.eswa.2021.116076>
77. Othmani, A., et al.: Towards robust deep neural networks for affect and depression recognition from speech. In: *International Conference on Pattern Recognition*, 5–19. Springer (2021)
78. Haque, A., et al.: Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv preprint arXiv:1811.08592* (2018)
79. He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. *J. Biomed. Inform.* 83, 103–111 (2018)
80. Chao, L., et al.: Multi task sequence learning for depression scale prediction from video. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 526–531. IEEE (2015)
81. Al Jazaery, M., Guo, G.: Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE. T. Affect. Comput.* 12(1), 262–268 (2018)
82. Yang, L., Jiang, D., Sahli, H.: Feature augmenting networks for improving depression severity estimation from speech signals. *IEEE. Access*. 8, 24033–24045 (2020)
83. Muzammel, M., et al.: AudVowelConsNet: a phoneme-level based deep CNN architecture for clinical depression diagnosis. *Mach. Learn. Appl.* 2, 100005 (2020). <https://doi.org/10.1016/j.mlwa.2020.100005>
84. Yang, L., et al.: Multimodal measurement of depression using deep learning models. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 53–59. ACM (2017)
85. Vázquez-Romero, A., Gallardo-Antolín, A.: Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy-Switz*. 22, 688 (2020). <https://doi.org/10.3390/e22060688>
86. Niu, M., et al.: A time-frequency channel attention and vectorization network for automatic depression level prediction. *Neurocomputing*. 450, 208–218 (2021). <https://doi.org/10.1016/j.neucom.2021.04.056>
87. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent Neural Network Regularization (2014). *arXiv preprint arXiv:1409.2329*
88. Du, Z., et al.: Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 23–30. ACM (2018)
89. Salekin, A., et al.: A weakly supervised learning framework for detecting social anxiety and depression. *Proc. ACM. Interact. Mob. Wearable. Ubiquitous. Technol.* 2(2), 1–26 (2018)
90. Jiang, H., et al.: Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Commun.* 90, 39–46 (2017). <https://doi.org/10.1016/j.specom.2017.04.001>
91. Mundt, J.C., et al.: Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J. Neurolinguist.* 20(1), 50–64 (2007)
92. Valstar, M., et al.: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10. ACM (2013)
93. Valstar, M., et al.: AVEC 2014: 3d dimensional affect and depression recognition challenge. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10. ACM (2014)
94. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 3123–3128. ELRA (2014)
95. Cift, E., et al.: The turkish audio-visual bipolar disorder corpus. In: 2018 1st Asian Conference on Affective Computing and Intelligent Interaction, pp. 1–6. (2018)
96. Cai, H., et al.: MODMA dataset: a multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv:2002.09283* (2020)
97. DeVault, D., et al.: SimSensei Kiosk: a virtual human interviewer for healthcare decision support. In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 1061–1068. IFAAMAS (2014)
98. Ringeval, F., et al.: AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 3–12. ACM (2019)
99. Niu, M., et al.: Multimodal spatiotemporal representation for automatic depression level detection. *IEEE. T. Affect. Comput.* (2020)
100. Zhao, Z., et al.: Hierarchical attention transfer networks for depression assessment from speech. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7159–7163. IEEE (2020)
101. Ray, A., et al.: Multi-level attention network using text, audio and video for depression prediction. In: *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pp. 81–88. ACM (2019)
102. Zhao, Y., et al.: Multi-head attention-based long short-term memory for depression detection from speech. *Front. Neurorobotics*. 15, 1–11 (2021)
103. Kroenke, K., et al.: The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disorders*. 114(1–3), 163–173 (2009)
104. Kroenke, K., Spitzer, R.L., Williams, J.B.: The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. *Psychosom. Med.* 64(2), 258–266 (2002)

How to cite this article: Wu, P., et al.: Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Trans. Intell. Technol.* 1–11 (2022). <https://doi.org/10.1049/cit2.12113>