Hindawi
Computational Intelligence and Neuroscience
Volume 2023, Article ID 6670869, 9 pages
https://doi.org/10.1155/2023/6670869

Research Article

Speech Deception Detection Based on EMD and Temporal Neural Network

Youjun Jiang, Haibo Chen, Shusen Yuan, Hongbo Xing, Yewen Cao, Deqiang Wang, and Hailiang Xiong

School of Information Science and Engineering of Shandong University, Qingdao, Shandong, China

Correspondence should be addressed to Yewen Cao; ycao@sdu.edu.cn and Deqiang Wang; wdq_sdu@sdu.edu.cn

Received 6 February 2023; Revised 10 April 2023; Accepted 18 April 2023; Published 29 May 2023

Academic Editor: Jatinderkumar R. Saini

Deceptive behaviour is a common phenomenon in human society. Research has shown that humans are not good at distinguishing deception, so studying automated deception detection techniques is a critical task. Most of the relevant technologies are susceptible to personal and environmental influences: EEG-based technologies need large and expensive equipment, facial-based technologies are sensitive with the camera's perspective, and these reasons have somewhat limited the development of applications for deception detection technologies. In contrast, the equipment required for speech deception detection is cheap and easy to use, and the capture of speech is highly covert. Based on the application of signal decomposition algorithms in other fields such as EEG signals and speech emotion recognition, this paper proposed a signal decomposition and reconstruction method based on EMD to process the speech signal and a better deception detection performance was obtained by improving the speech quality. The comparison results with other decomposition algorithms showed that the EMD decomposition algorithm is the most suitable for our method. Across many different classification algorithms, accuracy improved by an average of 2.05% and the $F1$ score improved by an average of 1.7%. In addition, a new deception detector, called the TCN-LSTM network, was proposed in this paper. Experiments showed that this network organically combines the processing capability of TCN and LSTM for time series data; the recognition rate of deception detection was greatly improved, with the highest accuracy and $F1$ score reaching 86.2% and 86.0% under the EMD-based signal decomposition reconstruction method. Based on the research in this paper, the signal decomposition algorithms need to be further optimised for speech signals and more classification algorithms not used for this task should be tried.

1. Introduction

The study of deception detection is a work of great significance, especially the act of deceiving someone to avoid the punishment for crime, which has been widely studied and applied in the legal, military, and forensic fields [1]. Deception is a deliberate act of misleading others to gain some advantage or avoid punishment [2] and does not include, for example, self-deception, pathological behaviour, and whether an act represents deception or not depends on what the intention of the person acting is. In psychological terms, a person is deceiving when he exhibits subconscious or conscious behaviours, including shortening of speech, flushing, change in speech frequency, eye avoidance, change in pupil diameter, and change in body posture [3].

Compared to automatic deception detection systems, it is more difficult to rely on humans themselves to recognize deception, it is a challenging task for nonspecialists to accurately detect deceptions [4], and humans themselves are highly subjective, so automatic deception detection methods have considerable research value [5]. Current research on deception detection focuses on the following areas: physiological signals (e.g., electroencephalogram, electromyogram, and so on), facial expressions, and body posture. Changes in physiological signals as indicators of deception detection have been used throughout the recent history of deception detection research. These signals can accurately

reflect changes in a person's mental state and have led to the development of many polygraphs, which have been used in various fields for many years. But there are problems with this method, as the acquisition of these signals requires close contact with the subject, which is an invasive method and has a high probability of causing psychological fluctuations in the subject, leading to inaccurate detection. Deception detection based on face and gesture does not require contact signal acquisition, requiring only a camera as the primary device, reducing the additional stress on the subject, and physical changes such as expressions and gestures have been proved by psychologists to characterize changes in a person's psychological state. However, using this modality for deception detection requires a certain viewpoint, and if the camera angle is faulty, it is difficult to identify the deception.

But if we use speech signals to recognize deceptions, the capture of speech is covert and can significantly reduce the psychological stress on the subject. What is more, where the recording device is located does not affect deception detection. Studies have shown that speech can map the psychological state of the speaker at the moment, and one can easily distinguish the general psychological state of the speaker (happy, sad, or angry) through speech [4]. In addition, researchers have found that people who intend to deceive others often show small changes in a range of behaviours such as vocal pressure, pitch, speech rate, and vocal organs when deceiving [6]. Deception detection using speech is already being investigated in many fields, for example, Duke University's Throckmorton used speech and language analysis methods to identify financial deception [7], so it is feasible to analyse speech signals to detect deceptions.

There have been many studies on deception detection based on speech signals. Machine learning algorithms have achieved good results in speech deception detection. Researchers at Columbia University analysed the effectiveness of machine learning methods and human detection methods on the CSC (Columbia-SRI-Colorado) dataset and proved that human detection methods were inferior to random selection, while methods based on support vector machines and Gaussian mixture models achieved 64.4% accuracy [8]. In Enos's Ph.D. thesis, he conducted a comparative analysis of the performance of five algorithms, including support vector machines, naive Bayes, logistic regression, decision trees, and ripper algorithms; the results showed that decision trees and support vector machines have better performance [9]. Velichko et al. analysed many different machine learning algorithms on the real-life trail dataset, and the most effective random forest algorithm achieved an accuracy of 79.4% [10]. The literature [11] investigated the impact of ensemble learning methods on deception detection performance, achieving a 70% recognition rate with the real-life trail dataset. Bareeda et al. used Mel frequency cepstral coefficients and SVM to detect deceptions and obtained an accuracy of 81% [12].

Neural networks, which have received much attention in recent years, have also played an important role in research on speech deception detection. Xie et al. proposed to replace the multiplicative operation in the LSTM with convolutional operation and achieved an accuracy of 68.4% in the CSC dataset [13]. In 2019, Xie et al. proposed a method that combines speech features with deep learning, and they got an accuracy of 71.4% using a deep belief network [14]. Fu et al. used an improved self-encoder for deception detection and achieved 62.78% and 63.89% accuracy on the CSC corpus and the self-made dataset, respectively [15]; in 2020, they proposed a method based on denoising autoencoder (DAE) and long short-term memory (LSTM) network, with an accuracy of 65.78% on the CSC and 68.89% on the homemade datasets [16]. Hershkovitch Neiterman et al. developed a deception detection recognition system based on MLP and LSTM and conducted experiments on cross-lingual datasets [17]. Chou and Lee proposed a BLSTM (bidirectional long and short-term memory network) architecture with dense layers that incorporate an attention mechanism [18], feature fusion [19], and a multitask architecture [20], all of which achieved excellent performance on their own corpus.

In addition to the abovementioned classification methods, many other networks are not used to detect deception, considering that the speech signal is a time-series data; in this paper, TCN (temporal convolutional network) is used to detect deception speech signals. This network performed well in dealing with time series and achieved better than other networks on many application scenarios [21] but has not been used to detecting deception.

Regarding speech deception detection, most studies have focused on the classifier and feature level but ignored the speech itself, which is very critical for speech deception detection. Speech signals contain multiple components, so it makes sense to decompose and analyse the signal. Similar attempts have been made in the study of EEG (electroencephalogram) signals since there are many different waves in the EEG (mainly consisting of α, β, γ, δ,, and θ), and the application of EEG signals to solve some practical problems requires the analysis of different waves: the literature [22] developed a scheme to automatically identify schizophrenia by decomposing the EEG signal through EMD (empirical mode decomposition) and calculating 22 each feature from it; Reference [23] proposed a computer-aided clinical decision support system (CACDSS) to detect and diagnose Parkinson's disease through EEG by combining automatic variational modal decomposition (AOVMD) and automatic extreme learning machine (AOELM) classifiers; the literature [24] developed an EEG rhythm separation (VHERS) based on variational modal decomposition (VMD) and Hilbert transform (HT) to help experts detect attention deficit hyperactivity disorder (ADHD) in a real-time situation. Reference [25] proposed the robust tuneable Q wavelet transform (TQWT) for the automatic selection of optimal tuning parameters to accurately decompose nonsmooth EEG signals and identify motor imagery (MI) tasks with low complexity.

What is more, in some studies of speech emotion recognition, classical signal decomposition algorithms were used to improve the performance of emotion recognition. Reference [26] proposed a feature extraction method based on VMD (variational modal decomposition) for speech emotion recognition, and they also conducted a comparative

validation of EMD (empirical mode decomposition) and LMD (local mean decomposition). Kerkeni et al. [27] used EMD and Teager–Kaiser Energy Operator (TKEO) and obtained high accuracy both in the Spanish sentiment database and in the Berlin database. Krishnan et al. [28] proposed to use EMD to classify signals into high, medium, and low frequencies and then calculate five entropy values at the three frequencies and achieved good performance by using these entropy features. However, all these algorithms have not dealt with speech deception detection. These articles showed the usefulness of signal decomposition techniques in speech signals.

So, based on combining EMD and signal reconstruction, a novel deception detection system was proposed, where the combination of TCN and LSTM is used as the classifier. Overall, this paper has two contributions as follows:

(1) First, according to the general characteristics of the signal, the signal decomposition algorithm, which is rare in speech signal processing, was used. The proposed method effectively improved the performance of speech deception detection.

(2) Second, two networks that can process time series data were concatenated, and they greatly enhanced their respective capabilities. The network retains the time series features of speech signals as much as possible so that the results of speech deception detection are greatly improved.

2. EMD and TCN-LSTM Deception Detection System

As shown in Figure 1, first, the speech signal is decomposed and reconstructed, and then the reconstructed signal is used to extract MFCC (Mel frequency cepstral coefficients) features, and finally, the features are sent to the TCN-LSTM network to train the deception detection classifier.

2.1. EMD (Empirical Mode Decomposition). It is an adaptive signal decomposition method proposed by Huang et al. [29], which is useful in nonsmooth and nonlinear signals, and the speech signal is exactly this kind of signal. The EMD algorithm decomposes the signal into imfs (intrinsic mode functions), and an imf must satisfy the following two conditions: first, the difference between the number of extreme points and zero crossing points is 0 or 1; second, at any time, the average of the upper envelope formed by the local maximums and the lower envelope formed by the local minimums is 0.

The specific steps of EMD are as follows:

(1) We plot the upper and lower envelopes, respectively, based on the local maximum and minimum values of the original signal.

(2) We calculate the mean value of the upper and lower envelopes to obtain the mean envelope.

(3) We let the original signal subtract the mean envelope to obtain the intermediate signal.

(4) If the intermediate signal meets the two conditions of the imf, then this signal is an imf, and the steps (1) to (4) will be repeated using the mean envelope as the original signal; otherwise, the intermediate signal will be used as the original signal, and the steps (1) to (4) are repeated. We iterate this process until the stopping condition is satisfied.

Some of the imfs obtained by EMD processing may be ineffective or even inhibitory in distinguishing deceptive speech. So, some imfs could probably be removed to improve the discrimination of deceptive speech. This observation is the motivation for our proposed scheme.

2.2. Speech Signal Decomposition and Reconstruction. In our proposed scheme, the original speech signal needs to be preprocessed, i.e., first resampling the speech and then following a preemphasis operation (made on the sampled signal to weigh the high-frequency part of the speech) to remove the effects of lip radiation.

Next, the aboveprocessed signal is decomposed using EMD to obtain the different imf components ($\mathrm{imf}_1, \ldots, \mathrm{imf}_i, \ldots, \mathrm{imf}_n$), and the main components are selected based on a correlation threshold as follows. The correlation coefficient R_i of each imf with the preprocessed signal is calculated as follows:

$$R_i = \frac{\sum_{m=1}^{M-1}\left[\mathrm{imf}_i(m) - \overline{\mathrm{imf}_i}\right]\left[y(m) - \overline{y}\right]}{\sqrt{\sum_{m=1}^{M-1}\left[\mathrm{imf}_i(m) - \overline{\mathrm{imf}_i}\right]^2}\sqrt{\sum_{n=1}^{N-1}\left[y(m) - \overline{y}\right]^2}}, \tag{1}$$

where M is the total number of samples of the speech signal, $\mathrm{imf}_i(m)$ is the mth sample value of the ith subsignal obtained by EMD decomposition, $y(m)$ is the mth sample value of the signal before decomposition, $\overline{\mathrm{imf}_i}$ and $\overline{y}$ represent the average value of these two signal sampling points, respectively, and the value of R_i is in the range of [−1,1].

The correlation threshold value was obtained according to these coefficients. This threshold is calculated by

$$\text{Threshold} = \frac{\sqrt{\sum_{i=1}^{n}\left(R_i - \overline{R}\right)^2}}{n}, \tag{2}$$

where $\overline{R}$ is the mean value of all correlation coefficients and n is the total number of imfs. Finally, the components are selected according to the relationship between the magnitude of the correlation coefficients and the threshold value, and the selected components are recombined to obtain the reconstructed signal. The complete process is shown in Figure 2(a).

2.3. Feature and Classifier. Following the previously mentioned signal processing, the features are extracted and a classifier is trained, as shown in Figure 2(b). In this paper, MFCC, a cepstrum feature that has been proven effective by many researchers, was chosen, and it was proposed based on the auditory characteristics of the human ear [30]. The standard MFCC reflects only the static characteristics of speech parameters; the dynamic characteristics of speech can

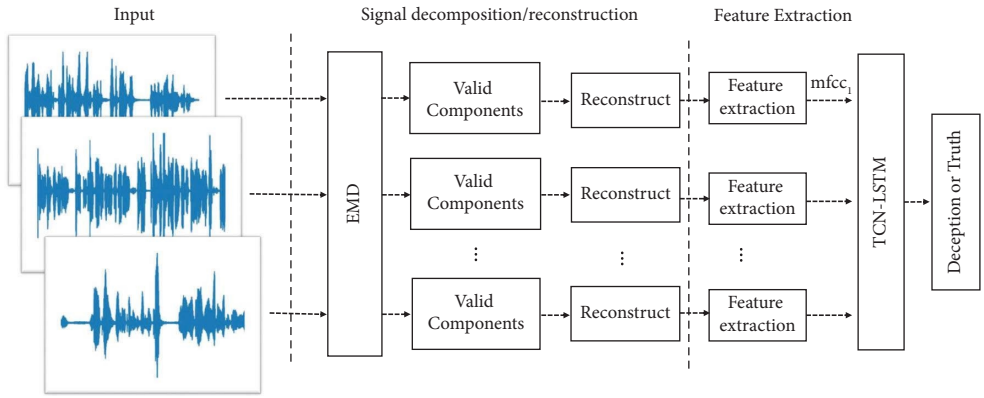

Figure 1: A system based on EMD and TCN-LSTM for speech deception detection.

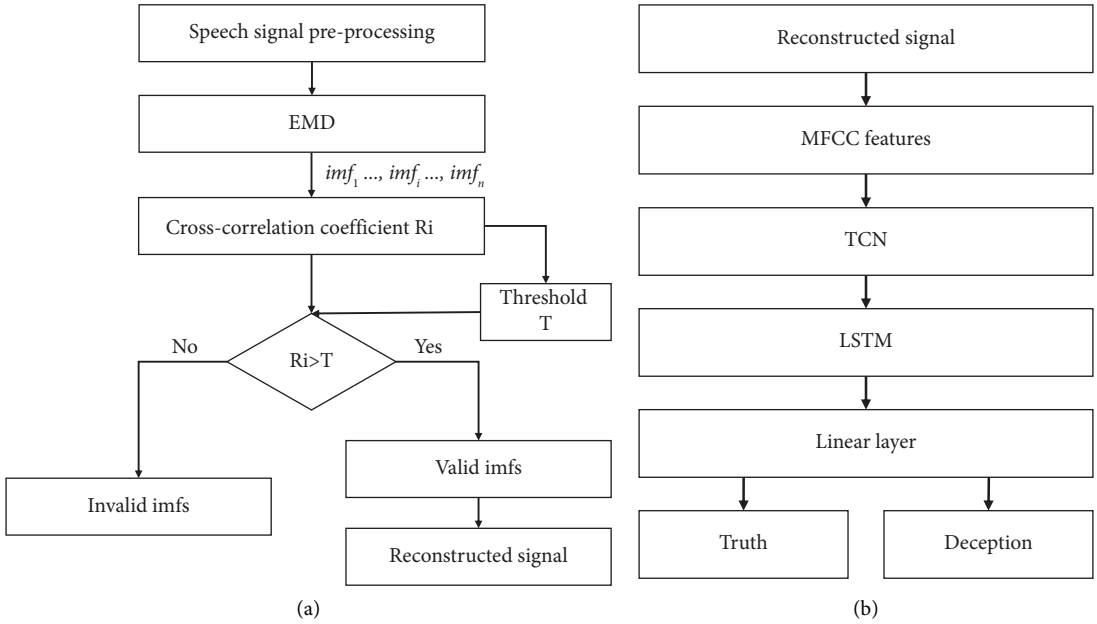

Figure 2: (a) EMD-based speech reconstruction method and (b) basic classifier structure of the TCN-LSTM concatenation model.

be described by the differential spectrum of these static features [31]. The complete process is shown in Figure 3, and the specific steps are described as follows:

(1) We split the speech signal into a frame-level representation and multiply each frame with a Hamming window

(2) FFT (fast Fourier transform) is applied to each frame to convert the time domain signal into a frequency domain representation

(3) Each frame is passed through a Mel filter bank, and the logarithmic energy output from each filter bank is calculated

(4) The standard MFCC coefficients are obtained by DCT (discrete cosine transformation)

(5) Finally, the first and second-order difference coefficients are calculated and combined with the standard coefficients to get the required MFCC features

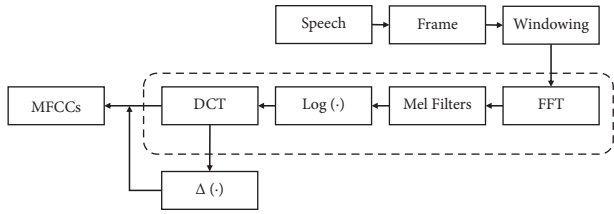

Figure 3: MFCCs' feature extraction block diagram.

As for classification algorithms, because TCN based on convolutional neural network structure has powerful deep feature extraction ability and LSTM based on the recurrent neural network has good modelling prediction ability for time-series data, the concatenation of TCN and LSTM is used to match the speech signal feature and need of deception detection in our proposed scheme. The MFCC features are then fed into TCN and LSTM in turn, which not only effectively extracts the deep information but also

improves the processing efficiency of LSTM, and then a linear neural network is used to output the final prediction result.

3. Dataset and Experiment Configurations

The corpus is a very critical issue in deception detection research. Currently, many deception detection researchers have constructed some datasets based on different approaches: researchers at Columbia University took the form of interviews to build the Columbia-SRI-Colorado (CSC) database [32]; the Idiap Research Institute in Switzerland recorded the Idiap Wolf dataset in the context of the werewolf game [33]; for the Chinese corpus, the Soochow University researchers constructed the SUSP deception detection dataset considering three cases, including induced deceptions, deliberately imitative deceptions, and natural deceptions [5].

However, many corpora are not open source, and only a small number of them are easily accessible. In this paper, our method is evaluated by using the real-life trial dataset [34], which is a dataset based on a real courtroom trial session. Researchers searched public multimedia sources to get data, including public court trials, and these sources should meet the condition that truthful and deceptive statements in them are easily detected and verified. The defendant and witnesses in the video should be visible, and the audio quality should be sufficient to be heard to understand what was being said.

In the real-life trial dataset, three different verdict results are considered: guilty, not guilty, and exonerated. Thus, the deceptive data were collected from the defendant or the suspect, while the truthful ones were collected from witnesses or from videos of the suspect answering certain facts (verified by the police). The final dataset consists of 121 videos, including 61 deceptive videos and 60 truthful videos, with an average length of 28.0 seconds (27.7 and 28.3 seconds for deceptive and truthful, respectively). The speakers in the dataset contained 21 females and 35 males, aged between 16–60 years old.

In our proposed scheme, for the computational convenience of EMD, an upper limit is set for the number of imfs, if the number reaches 20, no further decomposition will be made.

For the dimension of features, the first 13 dimensions of MFCCs were taken as the features in the experiments of this paper. The first-order and second-order difference features of these static features were extracted to describe dynamic features, and finally, the 39-dimensional MFCC features were used in experiments.

The specific experiments are as follows: first, considering the possible effect of the speech signal sampling rate on deception detection, the results of different sampling rates (4 kHz, 8 kHz, and 16 kHz) are compared. Second, the effectiveness of the signal decomposition reconstruction method is verified, including a comparison of various other decomposition algorithms. Finally, a TCN-LSTM network is trained for deception detection. A 10-fold cross-validation technique was used to ensure that the results obtained were more stable.

In this paper, the accuracy rate and $F1$ score are used to measure the performance of the system. Accuracy is the percentage of all correct predictions. The $F1$-score is a statistical measure of the performance of a binary classification model; it takes into account both the precision and recall of a classification model and can be seen as a weighted average of the precision and recall of the model. The formulas for calculating the accuracy and $F1$ are as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ F1 &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \end{aligned} \tag{3}$$

The definition of TP, TN, FP, and FN are shown in Table 1.

4. Results and Discussion

Table 2 contains the results when different sampling rates are used. Overall, the $F1$ scores do not differ significantly from the accuracy, representing that the model performance during the experiments was reliable. Specifically, there is no similar situation where the deception recognition rate is high and the truth recognition rate is low. It is shown that the results are not good at 4 kHz. Although the calculation speed is faster at a low sampling rate, the sampling rate of 4 kHz is difficult to retain enough information, resulting in some useful information being discarded. In addition, there is no significant difference between the results at 8 kHz and 16 kHz, indicating that enough useful information is retained at the two sampling rates. However, the signal decomposition speed of 8 kHz is faster, and the results under 16 kHz do not far exceed 8 kHz, so the 8 kHz sampling rate is set for the following results.

Table 3 shows the results of the signal decomposition reconstruction method for speech deception detection. In addition to using EMD as the decomposition algorithm, the results were compared with those of two other decomposition algorithms: one is the LMD (local mean decomposition), which decomposes a complex multicomponent signal into the sum of several product functions (PF); the other is the VMD (variational modal decomposition), which assumes that the signal consists of a series of subsignals with a specific centre frequency and finite bandwidth, and subsignals are obtained by constructing and solving a variational problem.

Overall, speech deception detection is improved by applying the signal decomposition reconstruction method. However, the VMD method causes a significant loss in the recognition rate. The reason may be as follows: VMD is computed by solving a variational problem, which cannot restore the signal by summing all subsignals as EMD does. This process is likely to produce changes in the internal structure of the signal, which has a negative effect on deception detection, a task that relies on signal depth information.

For more intuitive analysis, the results of EMD and LMD algorithms are statistically analysed and the histograms

Table 1: Definition of TP, TN, FP, and FN.

True value	Predicted value	
	1	0
1	TP (True positive)	FN (False positive)
0	FP (False negative)	TN (True negative)

Table 2: Deception detection results of different speech sampling rates.

	4 kHz		8 kHz		16 kHz	
	Acc (%)	*F*1 (%)	Acc (%)	*F*1 (%)	Acc (%)	*F*1 (%)
KNN	75.1	74.3	77.3	77.0	78.4	78.3
SVM	74.2	74.0	77.5	77.2	79.5	79.5
Decision tree	68.6	68.1	70.0	70.7	73.7	73.9
Random forest	75.5	76.0	78.6	79.0	79.5	79.3
Naïve Bayes	66.6	66.5	68.8	69.1	69.0	69.2
AdaBoost	70.1	71.0	72.8	73.1	73.6	73.5
Ensemble learning	77.2	77.0	79.2	79.5	79.4	79.4
MLP	70.9	70.2	72.1	72.3	73.1	72.7
RNN	58.2	59.0	64.7	64.2	64.3	65.1
Autoencoder	65.1	65.5	67.4	68.1	68.9	68.8

Table 3: Results with signal decomposition reconstruction using different decomposition algorithms.

Methods	Original signal		With EMD		With LMD		With VMD	
	Acc (%)	*F*1 (%)	Acc (%)	*F*1 (%)	Acc (%)	*F*1 (%)	Acc (%)	*F*1 (%)
KNN	77.3	77.0	78.8	78.5	78.2	78.5	76.3	76.0
SVM	77.5	77.2	79.6	79.5	79.3	79.2	76.0	76.2
Decision tree	70.0	70.7	70.8	71.0	71.0	70.5	69.4	69.0
Random forest	78.6	79.0	79.6	79.3	79.0	78.8	78.1	78.3
Naïve Bayes	68.8	69.1	70.3	70.1	70.1	70.2	66.9	66.6
AdaBoost	72.8	73.1	72.9	73.1	73.4	73.2	72.0	71.3
Ensemble learning	79.2	79.5	81.2	80.7	80.5	80.1	78.0	78.1
MLP	72.1	72.3	76.6	76.2	75.2	75.5	70.5	70.6
RNN	64.7	64.2	66.2	66.0	65.8	65.7	64.1	63.2
Autoencoder	67.4	68.1	72.9	72.8	70.8	71.1	66.6	65.8

shown in Figure 4 are drawn according to the difference of results between them and the original signal.

The histogram clearly shows the performance of the two signal decomposition algorithms, with EMD outperforming LMD on average, with both having a larger standard deviation, due to differences in the sensitivity of the different classification algorithms to reconstructed signals. Overall, the EMD algorithm improves the accuracy by an average of 2.05% and the *F*1 score by 1.7%, and subsequent experiments will be based on the EMD signal decomposition algorithm only.

Table 4 shows a comparison experiment of different parameter settings of the TCN-LSTM. The number of hidden layers of the TCN, the number of layers of the LSTM, and whether a bidirectional LSTM was used were compared. Based on some experience and other research, 3- and 4-layer TCN as well as 1- and 2-layer LSTM were verified.

According to the results in the table, first, the bidirectional operation of the LSTM is better than the unidirectional one, but the effect is not very large. It shows that although the past information in speech is influenced by the future, it is not very obvious. Second, the 2-layer LSTM not only increases the computational cost significantly, but the

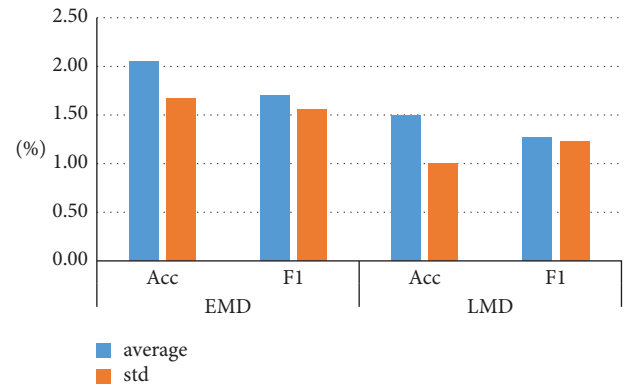

Figure 4: Analysis of the improvement of deception detection results by using EMD and LMD.

effect is instead reduced, which shows that the 1-layer LSTM is good enough for modelling the data. What is more, 2-layer LSTM gets the biggest difference between accuracy and the *F*1 score in this table which represents that model reliability is influenced. Finally, 4-layer TCN is more effective than a 3-layer TCN, suggesting that deeper features of the data are

Table 4: TCN-LSTM deception detection results with different parameter settings.

Number of TCN layers	Number of LSTM layers	Bi-LSTM	Acc (%)	$F1$ (%)
3	1	Yes	85.7	85.5
3	2	Yes	83.2	82.2
3	1	No	85.6	85.3
3	2	No	83.0	82.0
4	1	Yes	86.2	86.0
4	2	Yes	83.4	82.1
4	1	No	86.0	85.7
4	2	No	83.2	81.7

Table 5: Comparison results about TCN-LSTM, TCN, and LSTM.

	Acc (%)	$F1$ (%)
TCN	76.6	76.2
LSTM	69.2	69.6
TCN-LSTM	86.2	86.0

Table 6: Comparison with other speech deception detection research studies.

	Dataset	Classifier	Feature	Best result (%)
This paper	Real-life trail dataset	TCN-LSTM	MFCC	86.2
[12]	Real-life trail dataset	SVM	MFCC	81.5
[35]	Real-life trail dataset	Boosting models	IS16 + IS13 + IS11	85.6
[36]	CSC	Hybrid-RNN	Acoustic and lexical features	84.1
[37]	Self-made dataset	SVM	Prosodic features and MFCC	82.5

being mined, but it is more difficult to identify which depth of features is most appropriate. Moreover, the results show that the improvement in the recognition rate of the 4-layer TCN is not very large, so the research of more layers of TCN will not be made.

Table 5 shows the results using TCN and LSTM alone compared to the TCN-LSTM network. The accuracy and the $F1$ score have been greatly improved, which fully proves the superiority of TCN-LSTM. There must be a complementary relationship between TCN and LSTM. TCN effectively extracts deep features of speech, but its ability to learn the relationship between deception and deep features is not good. In contrast, the LSTM is weaker in extracting depth features but is considered to have good modelling and prediction capabilities for time series data. So in the TCN-LSTM, the LSTM effectively learns the relationship between deception and depth features generated by TCN.

5. Comparison with Other Studies

To more fully evaluate the work in this paper, a comparative analysis of deception detection studies conducted in recent years was carried out with the method in this paper. The comparators selected were all speech-based studies. It is shown in Table 6.

Compared with the research under the same dataset (the real-life trail), the final scheme of this paper has obvious advantages, but few studies are using deep learning to train the dataset, and more attempts are needed. The recognition rate of this paper is also higher compared to studies with other datasets. Compared to general research, the key element of this thesis is that most deception detection studies do not focus on the temporal information in speech and the additional processing of speech.

6. Conclusions

In this paper, a novel system for speech deception detection is proposed. The use of EMD decomposition to reconstruct the speech signal improves the quality of the original speech and increases the recognition rate under a variety of classical classification algorithms, with an average improvement of 2.05% accuracy and 1.70% of $F1$ score. In addition, the new network architecture TCN-LSTM organically combines the features of TCN and LSTM and has extremely strong temporal data processing capability, achieving 86.2% accuracy and 86.0% $F1$ scores under the real-life trail dataset. Moreover, the method of this paper has great advantages compared to similar studies.

However, there are still some shortcomings in this paper: first, the paper does not focus on the effect of other features; second, there is no suitable modification of EMD for speech signals; and finally, it is difficult to validate in other datasets due to the low sharing of datasets in this domain.

So, in future work, the first thing is to experiment with more combinations of features, and the second thing is to further investigate improvements in signal decomposition algorithms for speech signals. As for the deception detection dataset, the plan is to produce a small Chinese dataset drawing on existing datasets, but this will only provide a limited contribution and will not fully solve the problem.

Data Availability

The data used to support the findings of the study can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key R and D Program of China under grant no. 2020YFC0833201 and in part by the Natural Science Foundation of Shandong Province under grant no. ZR2020MF004.

References

[1] Y. Zhou, H. Zhao, and X. Pan, "Lie detection from speech analysis based on K-SVD deep belief network model," in *Proceedings of the International Conference on Intelligent Computing*, pp. 189–196, Springer, Fuzhou, China, August 2015.

[2] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychological Bulletin*, vol. 129, no. 1, pp. 74–118, 2003.

[3] Z. Labibah, M. Nasrun, and C. Setianingsih, "Lie detector with the analysis of the change of diameter pupil and the eye movement use method Gabor wavelet transform and decision tree," in *Proceedings of the IEEE Int. Conf. Internet Things Intell. Syst. (IOTAIS)*, pp. 214–220, Bali, Indonesia, November 2018.

[4] C. Kirchhuebel, *The Acoustic and Temporal Characteristics of Deceptive Speech*, University of York, York, UK, 2013.

[5] C. Fan, H. Zhao, and X. Chen, "Distinguishing deception from non-deception in Chinese speech," in *Proceedings of the 2015 Sixth International Conference on Intelligent Control and Information Processing (ICICIP)*, pp. 268–273, IEEE, Wuhan, China, November 2015.

[6] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Invited article: face, voice, and body in detecting deceit," *Journal of Nonverbal Behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[7] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decision Support Systems*, vol. 74, pp. 78–87, 2015.

[8] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing*, Toulouse, France, May 2006.

[9] F. Enos, *Detecting Deception in Speech*, The Graduate School of Arts and Sciences, Columbia University, New York, NY, USA, 2009.

[10] A. Velichko, V. Budkov, and I. Kagirov, "Comparative analysis of classification methods for automatic deception detection in speech," in *Proceedings of the 20th International Conference on Speech and Computer SPECOM-2018*, pp. 737–746, Springer, Leipzig, Germany, September 2018.

[11] A. Velichko, V. Budkov, and I. Kagirov, "Applying ensemble learning techniques and neural networks to deceptive and truthful information detection task in the flow of speech," in *Intelligent Distributed Computing XIII*, pp. 477–482, Springer International Publishing, Berlin. Germany, 2020.

[12] E. Bareeda, B. Mohan, and K. Muneer, "Lie detection using speech processing techniques," *Journal of Physics: Conference Series*, vol. 1921, Article ID 012028, 2021.

[13] Y. Xie, R. Liang, H. Tao, Y. Zhu, and L. Zhao, "Convolutional bidirectional long short-term memory for deception detection with acoustic features," *IEEE Access*, vol. 6, pp. 76527–76534, 2018.

[14] Y. Xie, R. Liang, and Y. Bao, "Deception detection with spectral features based on deep belief network," *Journal of Acoustics*, vol. 44, no. 2, pp. 214–220, 2019.

[15] H. Fu, P. Lei, H. Tao, L. Zhao, and J. Yang, "Improved semi-supervised autoencoder for deception detection," *PLoS One*, vol. 14, no. 10, Article ID e0223361, 2019.

[16] H. Fu and P. Lei, "Speech deception detection algorithm based on denoising auto-encoder and long short-term memory network," *Computer Applications*, vol. 40, no. 2, pp. 589–594, 2020.

[17] E. Hershkovitch Neiterman, M. Bitan, and A. Azaria, "Multilingual deception detection by autonomous agents," in *Proceedings of the Companion Web Conference*, pp. 480–484, Taipei Taiwan, April 2020.

[18] H. Chou and C. Lee, "Your behavior makes me think it is a lie': recognizing perceived deception using multimodal data in dialog games," in *Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Auckland, New Zealand, December 2020.

[19] H. Chou, Y. Liu, and C. Lee, "Joint learning of conversational temporal dynamics and acoustic features for speech deception detection in dialog games," in *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, November 2019.

[20] H. Chou, Y. Liu, and C. Lee, "Automatic deception detection using multiple speech and language communicative descriptors in dialogs," *APSIPA Transactions on Signal and Information Processing*, vol. 10, no. 1, 2021.

[21] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, https://arxiv.org/abs/1803.01271.

[22] S. Siuly, S. K. Khare, V. Bajaj, H. Wang, and Y. Zhang, "A computerized method for automatic detection of schizophrenia using EEG signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 11, pp. 2390–2400, 2020.

[23] S. K. Khare and V. Bajaj, "A CACDSS for automatic detection of Parkinson's disease using EEG signals," in *Proceedings of the 2021 International Conference on Control, Automation, Power and Signal Processing (CAPS)*, pp. 1–5, Jabalpur, India, December 2021.

[24] S. K. Khare, N. B. Gaikwad, and V. Bajaj, "VHERS: a novel variational mode decomposition and Hilbert transform-based EEG rhythm separation for automatic ADHD detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, Article ID 4008310, pp. 1–10, 2022.

[25] S. K. Khare, N. Gaikwad, and N. D. Bokde, "An intelligent motor imagery detection system using electroencephalography with adaptive wavelets," *Sensors*, vol. 22, no. 21, p. 8128, 2022.

[26] Y. Liu, X. Zhang, and G. Chen, "Feature extraction of emotional speech based on improved GFCC with VMD,"

Computer Engineering and Design, vol. 41, no. 8, pp. 2265–2270, 2020.

[27] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech Communication*, vol. 114, pp. 22–35, 2019.

[28] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features: speech emotion recognition," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1919–1934, 2021.

[29] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[30] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[31] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 34, no. 1, pp. 52–59, February 1986.

[32] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic, lexical, and cepstral systems for deceptive speech detection," in *Proceedings of the 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, pp. 1033–1036, Toulouse, France, May 2006.

[33] H. Hung and G. Chittaranjan, "The Idiap Wolf corpus: exploring group behaviour in a competitive role playing game," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 879–882, Firenze, Italy, October 2010.

[34] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 59–66, SA, USA, November 2015.

[35] A. N. Velichko and A. A. Karpov, "Automatic detection of deceptive and truthful paralinguistic information in speech using two-level machine learning model," in *Proceedings of the Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pp. 698–704, June 2021.

[36] S. Desai, M. Siegelman, and Z. Maurer, *Neural Lie Detection with the CSC Deceptive Speech Dataset*, Stanford University, Stanford, CA, USA, 2017.

[37] H. Tao, P. Lei, M. Wang, J. Wang, and H. Fu, "Speech deception detection algorithm based on SVM and acoustic features," in *Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 31–33, Dalian, China, October 2019.