LETTER

Detecting Depression from Speech through an Attentive LSTM Network

Yan ZHAO[†], **Yue XIE**[††], ***Nonmembers***, **Ruiyu LIANG**[††], ***Member***, **Li ZHANG**[†††], **Li ZHAO**[†a)], ***and*** **Chengyu LIU**[††††b)], ***Nonmembers***

SUMMARY Depression endangers people's health conditions and affects the social order as a mental disorder. As an efficient diagnosis of depression, automatic depression detection has attracted lots of researcher's interest. This study presents an attention-based Long Short-Term Memory (LSTM) model for depression detection to make full use of the difference between depression and non-depression between timeframes. The proposed model uses frame-level features, which capture the temporal information of depressive speech, to replace traditional statistical features as an input of the LSTM layers. To achieve more multi-dimensional deep feature representations, the LSTM output is then passed on attention layers on both time and feature dimensions. Then, we concat the output of the attention layers and put the fused feature representation into the fully connected layer. At last, the fully connected layer's output is passed on to softmax layer. Experiments conducted on the DAIC-WOZ database demonstrate that the proposed attentive LSTM model achieves an average accuracy rate of 90.2% and outperforms the traditional LSTM network and LSTM with local attention by 0.7% and 2.3%, respectively, which indicates its feasibility.

key words: *depression detection, LSTM, attention mechanism, vocal expression*

1. Introduction

Depression is a mental disorder and recognized as a major factor for the increase in suicide. It has negative effect on a person's mental state, which may lead to a persistent low mood and a loss of interest in daily life. The deleterious effect of depression makes it a burden of society and affects the social order. Thus, detection and treatment of depression is urgent.

Traditional approaches [1], [2] for depression detection mainly rely on the interviews with patients and the mental status test, which generates a depression level point from the evaluation index. However, such an approach has drawbacks. For example, approaches employed by [1], [2] have substantial costs in time and sometimes can be affected by subjective factors. To solve out this issue, an effective and reliable automatic depression detection method is in urgent need to help assist general practitioners. Wen [3] explored facial region visual-based nonverbal behavior analysis for automatic depression diagnosis. Zhu [4] proposed a deep neural network for automated depression diagnosis based on facial appearance and dynamics. Zhang [5] used multi-agent strategy with fusion of electroencephalography and vocal expression for multimodal depression detection. Their studies demonstrate that depression classification is a task which can be tackled by machine-based automatic detection and is expected to improve detection efficiency. Though facial features are widely used for classification, the image data acquisition process will make participants more nervous, which could influence the result. Therefore, using vocal features is a more convenient and relaxed way to go.

Considerable researches have shown that depression is associated with vocal prosody and expression. Moore [14] analyzed variations in prosodic feature statistics, showing how someone was speaking as well as what they were saying were considerable for depression evaluation. Cummins [15] investigated several speech features, demonstrating that a combination of Mel Frequency Cepstral Coefficients (MFCC) and formant based features was effiective for depression detection. Zhang [5] made a point that depression was related with low extraversion, which was manifested in slow speech and sensitive to negative emotions. Le [16] directly extracted low level descriptors (LLDs) from speech segments as acoustic features. However, various acoustic features differ in detecting negative emotions and the emotional saturation shows its diversity in speech fragment [18], while most previous studies did not focus on it. On the basis, for mining negative-emotion-specific information in speech, this paper proposes an attention based LSTM network, which has achieved great success for speech emotion recognition [6], [7]. Besides, for the reason that the frame-level features can be considered as a two-dimensional matrix with respect to time and feature dimensions, the multi-dimensional attention mechanism to weight the LSTM output is employed to distinguish the difference of depression and non-depression among the speech segments, for better performance of depressive speech classification.

Manuscript received October 12, 2020.
Manuscript revised June 11, 2021.
Manuscript publicized August 24, 2021.

[†]The authors are with Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Jiangsu Nanjing, 210096, P.R. China.
[††]The authors are with School of Communication Engineering, Nanjing Institute of Technology, Jiangsu Nanjing, 211167, P.R. China.
[†††]The author is with Computational Intelligence Group, Northumbria University, Newcastle upon Tyne, UK.
[††††]The author is with School of Instrument Science and Engineering, Southeast University, Nanjing, 210096, P.R. China.
a) E-mail: zhaoli@seu.edu.cn
b) E-mail: chengyu@seu.edu.cn
DOI: 10.1587/transinf.2020EDL8132

2. Attentive LSTM Network

Network with attention mechanism firstly achieved the state-of-art results in the field of image processing [8], [9]. The main idea is to assign different weights based on tasks rather than to provide the same weight for all the distinctions. Inspired by such strategies, this study employed a self-attention mechanism to improve the final performance. Moreover, the frame-level features used in the study contained not only the time-domain but also feature-domain information, to enhance the classification performance.

2.1 LSTM Network

In LSTM cell, a forget gate determines whether the prior information should be discarded or not. LSTM cell state is related to the present input and the previous output in the LSTM, as stated by Hochreiter [10]. The update formulas are summarized as follows:

$$\begin{aligned} i_t &= \sigma\left(W_i \times [C_{t-1}, h_{t-1}, x_t] + b_i\right) \\ C'_t &= \tanh\left(W_c \times [C_{t-1}, h_{t-1}, x_t] + b_c\right) \\ f_t &= \sigma\left(W_f \times [C_{t-1}, h_{t-1}, x_t] + b_f\right) \\ C_t &= f_t \cdot C_{t-1} + i_t \cdot C'_t \end{aligned} \tag{1}$$

where C_{t-1} represents the prior cell state, h_{t-1} represents the prior hidden layer output, x_t is the present input while C'_t is the candidate updating value, W represents weight and b represents bias, symbol '$\cdot$' represents the Hadamard product, the parameter σ denotes logistic sigmoid function. i_t and f_t represent the input and the forget gate calculation, respectively.

2.2 Multi-Dimensional Attention Algorithm

The proposed model used frame-level features, which capture the temporal information of emotion speech as the input of LSTM network. Then, the output of LSTM layers was passed on to the attention layer. The model employed the attention layer to calculate the score, verifying the abilities of different features for emotions to obtain final weighted features. In this study, attention mechanism is applied in both time and feature domains.

The contribution of each time step to the final depression detection is different, which means that the weights of frames could describe the degree of contribution. Mirsamadi [6] proposed an attention mechanism for the calculation of the frame weights, as defined in formula (2).

$$\alpha_t = \frac{e^{(u^H y_t)}}{\sum_{j=1}^{J} e^{(u^H y_j)}} \tag{2}$$

where parameter u denotes the attention parameter vector, H is the transpose operator, α_t represents the weight of the output y_t at the time step t.

The information is accumulated in the last output of LSTM network for its historic memory ability. Therefore, we use attention mechanism to make sure the last output could get a large weight. At last, we apply those weighted coefficients on both time and feature dimensions. The weight coefficients were applied to the output of all time on time domain and were summed up as the output:

$$\begin{aligned} s_T &= \text{softmax}\left(o_{\text{last}} \times (o_{\text{all}} \times w_t)^H\right) \\ O_T &= s_T \times o_{\text{all}} \end{aligned} \tag{3}$$

We employed multiple features for speech classification. Different features showed diverse abilities for different tasks. To figure out the difference, attention weight formula was calculated as:

$$\begin{aligned} s_F &= \text{softmax}\left(\tanh\left(o_{all} \times w_F\right) \times v_F\right) \\ O_F &= \textstyle\sum_{time} s_F \cdot o_{all} \end{aligned} \tag{4}$$

where w_F and v_F are trainable parameters. The summation was calculated on frames to obtain the statistical data of features from time domain. Thus, O_F can be viewed as the statistical value of the feature in the time dimension.

2.3 Deep Feature Representation Generation

Through the attention layers with respect to time and feature dimension, we obtain two feature matrix ($O_F \in R^{B,1,Z}$ and $O_T \in R^{B,1,Z}$). B represents the batch size, Z represents the selected feature's dimension and 1 here represents the last time step. Then, they are put into the Unsequeeze function. The new matrix with a dimension of $(B, 1, Z, 1)$ is obtained from the original matrix with $(B, 1, Z)$ dimension. Concat function is used for composing the two matrix into a new one. Finally, we apply an average pooling function to it for the deep feature representation.

$$\begin{aligned} O_{tf} &= \text{Concat}[O_T, O_F], O_{tf} \in R^{B,1,Z,2} \\ O_{last} &= \text{Average pooling}[O_{tf}], O_{last} \in R^{B,1,Z,1} \end{aligned} \tag{5}$$

The calculated deep feature representations strengthen the key information from both time and feature dimension and abandon irrelevant information for better depression detection performance.

Figure 1 exhibits the proposed model architecture. We

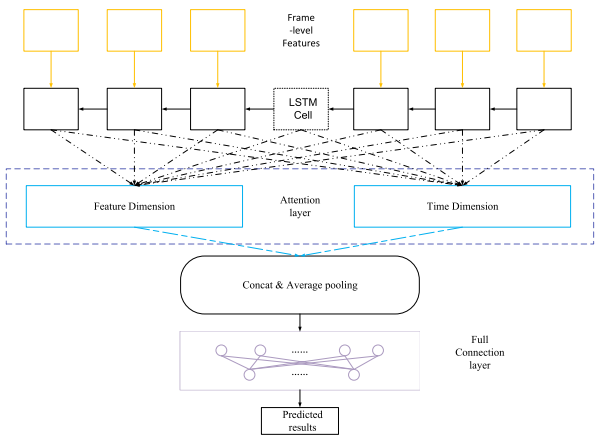

Fig. 1 Structure of attention-based LSTM

used the frame-level feature as the input data for the LSTM. The model applied attention to time and feature domains, respectively. Then, weight results were passed on to form the deep feature representations, which are used as the input of the fully connected layer. The fully connected layer connected each neural network unit of its output layer to all neural network units. At last, the output of softmax layer is the outcome of depression recognition.

3. Experiment and Discussion

3.1 Database

We apply the Distress Analysis Interview Corpus (DAIC-WOZ) [11] data corpus, which contains clinical interviews designed to support the diagnosis of psychological distress conditions. To analyze the depressive speech, we segment recordings into utterances automatically by audio processing tools. The sampling rate is 16000 Hz. The number of depressive speeches is 42 and we divide the depressive speeches into 2156 segments, while the number of normal speeches is 47 with 2245 segments. To ensure the effectiveness of the fragments, segments less than 3 seconds that contain litter information and those larger than 20 seconds that are difficult to handle are discarded in this research. At last, we use 3401 and 1000 speech segments randomly chosen by the software from the original segments as the train set and the test set, respectively.

3.2 Feature Extraction

Various acoustic features have been applied for depression prediction, such as prosodic features [5], [14], Mel spectrograms [17], spectral features [5], [15], [16]. According to the above characteristics, the voicing related features, energy and spectral related features are used.

Voicing features: Delay in expressive communication can be seen as a clinical sign of depression. The voicing probability is extracted in the study. Besides, the voice stability has a certain relation with mental state [5]. On the basis, we apply the jitter, shimmer and F0 as part of voicing features.

Energy and Spectral related features: The spectral features have been widely used in the field of speech emotion recognition [12]. The Mel spectral features, MFCC as well as its delta regression are emploied. Loudness is used for the reason that depression can also be embodied in the intensity of speech. We use the harmonic energy (harmonicERMS) and the glottal noise energy (noiseERMS) to preserve the difference between depresssive speeches and normal ones. The harmonics-to-noise ratio (HNR) can objectively reflect the characteristics of depressive speech.

In summary, based on previous researches of depression detection and speech emotion recognition, we employ the frame-level features extracted by openSMILE [13] in this research. Table 1 shows 7 voicing related features as well as 86 energy and spectral related features, where the numbers inside the brackets represent the corresponding feature dimensions.

Table 1 Frame-level features

Voicing related (7)	
voiceProb(1)	shimmerLocal(1)
jitterLocal(1)	jitterDDP(1)
F0(1) F0raw(1)	F0env(1)
Energy and Spectral related (86)	
pcm_loudness_sma(1)	pcm_loudness_sma_de(1)
harmonicERMS(1)	noiseERMS(1)
HNR(1)	pcm_zcr(1)
mfcc_sma(15)	mfcc_sma_de(15)
pcm_Mag(26)	logMelFreqBand(8)
lpcCoeff(8)	lspFreq(8)

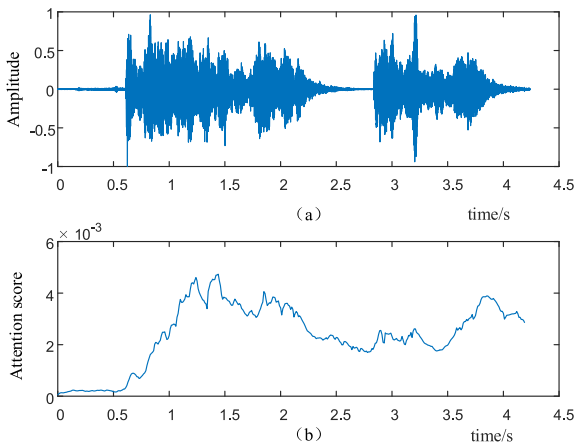

Fig. 2 (a) shows a depressive speech wave while (b) demonstrates its attention score in the time dimension

3.3 Experimental Setup

The proposed models include the LSTM network, the attentive LSTM on time dimension (LSTMT), the attentive LSTM on feature dimension (LSTMF) and the attentive LSTM on both time and feature dimension (LSTMTF). The initial learning rate is 0.0001. The network settings of all models are the same to ensure the effectiveness of experimental results.

3.4 Attentive LSTM Output

We apply the attention mechanism to the output of the LSTM to assign weights for the key information related to depression. The attention applied to the feature dimension does not have physical meanings and we consider it to assign weights to the LSTM output directly. Therefore, we only analyze the attention mechanism in the time dimension. Figure 2 demonstrates a depressive speech wave with its attention score in the time dimension.

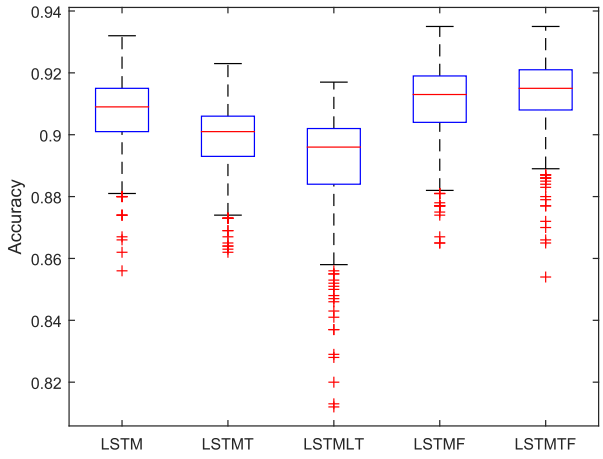

Fig. 3 Stability of five models on the test set

As is illustrated in Fig. 2, the attention score of the silent segment at the beginning is around zero while that of the silent segment around 2.5s achieves a better score. The phenomenon is caused by the memory ability for historical information of LSTM. The silent segment at the begining contains little information, which makes the attention score close to zero. Key information obtained between 1.5s and 2s will be accumulated to the following silent segment around 2.5s, leading to a better score rather than zero. This proves that not only the present information is calculated in the LSTM but also the prior information, indicating the importance of applying the attentive LSTM network.

3.5 Models Results

We employ the attention mechanism on the LSTM output to weigh the key information. We conducted a comparison between LSTM, LSTMF, LSTMT, and LSTMTF. Figure 3 establishes the stability of models in the test set. We also re-implementation Mirsamadi's local attention mechanism module (LSTMLT) [6] for comparison with our proposed model in Fig. 3. The vertical axis indicates the accuracy. The blue edges indicate 25 and 75 percentiles, denoting the stability of a model, and the red line show the median. The range of the black solid edges, which are connected to the blue rectangular box by a line, means the distribuction range of the model's accuracy except the outliers. As shown in Fig. 3, LSTMF and LSTMTF achieve the best performance. Compared with LSTMF, LSTMTF obtains a better unweighted average recall (UAR), which is defined as the accuracy per class averaged by total number of classes, with a similar stability.

Table 2 describe the results of different models in DAIC-WOZ corpus. 'accuracy' in Table 2 represents the average accuracy rate with the standard deviation. We calculate UAR and f1-score from the best performance of all models for depression detection. Compared with LSTM model, attentive LSTM models on time dimension (LSTMT and LSTMLT) show a negative impact on detection accuracy with a decrease around 1%. However, with a decrease about 2% of standard deviation, attentive LSTM models on time dimension exhibit a more stable detection accuracy range. The results of LSTMF prove the contrary. Compared with LSTM, LSTMF shows a higher UAR and a lower standard deviation. The above two situations indicate that the two dimension attention focus on different key information, leading to the diverse advantages and drawbacks. Our proposed model (LSTMTF) achieves an average accuracy rate of 90.2% and outperforms the traditional LSTM network and LSTM with local attention by 0.7% and 2.3%, respectively, which indicates its feasibility. In short, by concating the attentive LSTM output from time and feature dimensions, the proposed model captures high-level feature representations from different dimensions for superior detection performance.

Table 2 Results of different models (%)

Methods	accuracy	UAR	f1-score
LSTM	89.5±3.8	93.2	93.62
LSTMT	88.7±1.8	92.3	92.70
LSTMLT	87.9±1.4	91.7	91.70
LSTMF	90.0±3.4	**93.5**	93.96
LSTMTF	**90.2±1.6**	**93.5**	**94.01**

4. Conclusions

In this research, we propose an attentive LSTM network for depression detection from speech. The approach captures the key information from the frame-level features and distinguishes the difference of depression and non-depression among the speech segments by applying attention weighting on the time and the feature dimension. We carry out evaluations on the DAIC-WOZ database and the proposed model achieves the state-of-art performance of depression detection.

Further work is identified from the following aspects. This research seems to indicate that attention on the feature dimension has much more influence than that on the time dimension. Further studies are required to verify this observation. Besides, we only conduct the experiment on DAIC-WOZ corpus. Experiments should be carried out on other databases to verify the generalization of the proposed attention-based LSTM model for different depressive groups or languages.

Acknowledgments

This research was founded in part by the Distinguished Young Scholars of Jiangsu Province (BK20190014), and the Natural Science Foundation of China (No.81871444, 61673108, 61571106, 61633013).

References

[1] K. Kroenke, R.L. Spitzer, and J.B.W. Williams, "The PHQ-9: validity of a brief depression severity measure," Journal of General

Internal Medicine, vol.16, pp.606–613, 2001.
[2] K. Kroenke, T.W. Strine, R.L. Spitzer, J.B.W. Williams, J.T. Berry, and A.H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," Journal of Affective Disorders, vol.114, no.1-3, pp.163–173, 2009.
[3] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated Depression Diagnosis Based on Facial Dynamic Analysis and Sparse Coding," IEEE Trans. Inf. Forensics Security, vol.10, no.7, pp.1432–1441, 2015.
[4] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics," IEEE Transactions on Affective Computing, vol.9, no.4, pp.578–584, 2018.
[5] X. Zhang, J. Shen, Z. ud Din, J. Liu, G. Wang, and B. Hu, "Multimodal Depression Detection: Fusion of Electroencephalography and Paralinguistic Behaviors Using a Novel Strategy for Classifier Ensemble," IEEE J. Biomed. Health Inform., vol.23, no.6, pp.2265–2275, 2019.
[6] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in international conference on acoustics, speech, and signal processing, pp.2227–2231, 2017.
[7] Y. Xie, R. Liang, Z. Liang, and L. Zhao, "Attention-Based Dense LSTM for Speech Emotion Recognition," IEICE Trans. Inf. & Syst., vol.E102-D, no.7, pp.1426–1429, 2019.
[8] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified Visual Attention Networks for Fine-Grained Object Classification," IEEE Trans. Multimedia, vol.19, no.6, pp.1245–1256, 2017.
[9] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual Attention Network for Image Classification," in Proceedings of computer vision and pattern recognition, pp.6450–6458, 2017.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol.9, no.8, pp.1735–1780, 1997.
[11] J. Gratch, R. Artstein, G.M. Lucas, et al., "The distress analysis interview corpus of human and computer interviews," Proceedings of Language Resources and Evaluation, pp.3123–3128, 2014.
[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in Proceedings of Interspeech, pp.2794–2797, 2010.
[13] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in Proceedings of acm multimedia, pp.835–838, 2013.
[14] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Analysis of prosodic variation in speech for clinical depression," in Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.2925–2928, 2003.
[15] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," Proceeding of INTERSPEECH 2011 12th Annual Conference of the International Speech Communication Association, pp.2997–3000, 2011.
[16] L. Yang, H. Sahli, X. Xia, E. Pei, M.C. Oveneke, and D. Jiang, "Hybrid Depression Classification and Estimation from Audio Video and Text Information," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp.45–51, 2017.
[17] L. Lin, X. Chen, Y. Shen, and L. Zhang, "Towards automatic depression detection: a bilstm/1d cnn-based model," Applied Sciences, vol.10, no.23, p.8701, 2020.
[18] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," IEEE/ACM Trans. Audio, Speech, Language Process., vol.27, no.11, pp.1675–1685, 2019.