Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.033971
Article

Tech Science Press

An Optimal Method for Speech Recognition Based on Neural Network

Mohamad Khairi Ishak[1], Dag Øivind Madsen[2,*] and Fahad Ahmed Al-Zahrani[3]

[1]School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Nibong Tebal, 14300, Malaysia
[2]University of South-Eastern Norway, Bredalsveien 14, 3511, Hønefoss, Norway
[3]Computer Engineering Department, Umm Al-Qura University, Mecca, 24381, Saudi Arabia
*Corresponding Author: Dag Øivind Madsen. Email: Dag.Oivind.Madsen@usn.no
Received: 02 July 2022; Accepted: 26 August 2022

Abstract: Natural language processing technologies have become more widely available in recent years, making them more useful in everyday situations. Machine learning systems that employ accessible datasets and corporate work to serve the whole spectrum of problems addressed in computational linguistics have lately yielded a number of promising breakthroughs. These methods were particularly advantageous for regional languages, as they were provided with cutting-edge language processing tools as soon as the requisite corporate information was generated. The bulk of modern people are unconcerned about the importance of reading. Reading aloud, on the other hand, is an effective technique for nourishing feelings as well as a necessary skill in the learning process. This paper proposed a novel approach for speech recognition based on neural networks. The attention mechanism is first utilized to determine the speech accuracy and fluency assessments, with the spectrum map as the feature extraction input. To increase phoneme identification accuracy, reading precision, for example, employs a new type of deep speech. It makes use of the exportchapter tool, which provides a corpus, as well as the TensorFlow framework in the experimental setting. The experimental findings reveal that the suggested model can more effectively assess spoken speech accuracy and reading fluency than the old model, and its evaluation model's score outcomes are more accurate.

Keywords: Machine learning; neural networks; speech recognition; signal processing; learning process; fluency and accuracy

1 Introduction

Human speech contains different types of information. Given the diversity of vocal tract structures, each person's speech data has specific information for that speaker. Extracting speaker-specific information from complex speech data is trivial for humans, but a challenging task for computers. With the development of artificial intelligence, deep learning was introduced into the field of speech recognition in 2009 [1–4]. In just a few years, it has been widely used in speech recognition, speaker recognition, text recognition, emotion recognition and other related fields. In view of the differences and their expressions, scholars have achieved remarkable results in the field, but few scholars have conducted in-depth research on minors' reading voices. This paper focuses on the research of children's voices in reading scenarios. The children in this article range in age from 5 to 12, with the majority of them in primary school. Reading aloud is an essential learning skill for all children. Reading is the core link of primary school Chinese education, according to the primary school Chinese syllabus. Reading aloud is the most crucial and regular instruction in the subject [5–9]. It is clear from this that the importance of reading aloud in primary school education makes it difficult for instructors to properly train and evaluate each student's reading effect owing to the limited number and duration of language courses available to pupils [10–13].

By introducing deep learning technology, designing and mixing a variety of traditional models, attention mechanism and speech recognition technology, and using machines to automatically evaluate reading speech, it provides a good idea for students' after-school reading training.

In order to address the above issues, this paper proposed novel method based on neural networks. The main contributions are as follows:

- To recognize the speech and improve the fluency and accuracy of the learning process.
- It uses the children's reading speech corpus, extracts numerous fixed features from a huge amount of reading data, develops a model suited for reading speech assessment, and creates reading effect evaluation criteria.
- Then, adds an attention mechanism to describe speech recognition and spectrogram channels, respectively, with the goal of evaluating children's speech reading aloud based on the speech corpus.

Simulation results further validated the effectiveness of the proposed algorithm.

The remaining of the paper is organized as follows. Section 2 provides the related work. Section 3 explained the proposed model. Section 4 provides the experimentation results while Section 5 concludes the paper.

2 Related Work

Existing scientific research mainly focuses on speech recognition, speaker recognition, speech emotion analysis and natural language processing. For the blank field of children's reading speech, the basic premise of this work is to extract some judgmental values from reading speech. A specific feature, on this basis, appropriate selection of targeted models.

In recent years, three primary approaches for analyzing speech signals have emerged: 1) Extract signal features from original audio files [14] to capture the most authentic acoustic features; 2) run the deep learning model directly on the original audio waveform [15]; 3) convert speech to text using automatic speech recognition (ASR) technology, and then use a traditional text-based analysis system [16]. At the moment, the aforementioned three methodologies are being investigated in parallel, and there are limited cross-research topics.

2.1 Original Acoustic Feature Extraction Technology

The first method of speech analysis requires converting speech signals into speech feature vectors that computers can process, namely low-level descriptors (LLDs). It's divided into three categories: prosodic features, spectral characteristics, and timbre features [17]. Prosodic features are closely related to syntax, discourse, information structure, etc., mainly including pitch, formant, duration, fundamental frequency, energy, etc. The spectral feature is the measurement feature produced by the original signal excited in the vocal tract. At present, the common extraction methods are: Linear Prediction Coefficient (LPC), Mel Frequency Cepstral Coefficient (MFCC) [18] and Linear Prediction Cepstral Coefficient (LPCC) [19]. Breathing noises, guttural sounds, phonemes, word boundaries, brightness, and other sound quality characteristics are among the most common [20].

The speech acoustic and language characteristics of children aged 5–12 are different from those of adults [21]. Children's speech, for example, has a higher pitch [22], and formants occur at a higher frequency [23]. Through trials, reference [24] demonstrated that bandwidth has a significant impact on children's speech reading and speech appraisal.

2.2 Spectrogram

The core idea of the second speech analysis method is to retain the complete features of the speech signal and only convert it into the original audio waveform, that is, the speech spectrogram, which is derived from processing the received time-domain signal and correlates with it. The spectrograms include amplitude graphs, energy graphs, logarithmic energy graphs, etc.

Reference [25] adopted the spectrogram, fused it with the convolutional neural network (CNN) and attention mechanism, and used two different convolution kernels to extract time-domain features and frequency-domain features respectively, and saved the spectrogram as an image directly. After normalization processing, the accuracy of speech emotion evaluation is high.

Reference [26] used the spectrogram as the input, and a lot of their work focused on the preprocessing of cutting speech, and evaluated the immunity of the spectrogram to noise through experiments for different cutting durations, frequency resolutions and models.

2.3 Recurrent Neural Network Model and Attention Mechanism

Since the information extracted from a single speech signal must depend on the content of its context, speech features are generally fed into deep learning models such as recurrent neural networks (RNNs). For example, reference [27] extracted high-level features from the original spectrogram, fused CNN and long short-term memory (LSTM) architectures, designed a neural network for speech emotion recognition, and used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset to verify its effectiveness. Reference [28] combined the spectrogram and a three-layer LSTM to judge the robustness of the model to noisy data on the basis of comparing and analyzing whether the data is denoised or not. Reference [29] used the Gated Recurrent Unit (GRU) to recognize speech emotion, and achieved results comparable to LSTM on the basis of adding noise, but it can be applied to embedded devices.

According to the complexity of the speech signal in the existing scene, the attention model can be used to analyze and extract the effective features to achieve the best emphasis effect. For example, reference [30] used a combination of CNN and an attention mechanism to emphasize emotional content in music. Reference [31] connected the local attention mechanism to the RNN and achieved automatic identification of the speaker's mood by centrally extracting short-term frame-level acoustic information associated to emotion.

2.4 ASR Technology

A third method of speech analysis converts speech to text by recognizing each word spoken by the speaker in the audio and changing them into word embeddings, using some techniques from natural language processing such as word frequency-inverse file frequency (TF-IDF) and bag of words (BOW) model [32]. The results are not always accurate, since the accuracy of emotion detection depends on whether accurate pronunciation can be reliably detected in spoken language [33]. In addition, some emotion-related signal features are also lost when speech is converted to text, resulting in lower accuracy of emotion classification. At present, in the field of Chinese recognition, many domestic companies have opened speech recognition interfaces to the outside world, such as iFLYTEK, Yunzhisheng, Baidu, etc., but few companies have open sourced speech recognition engines, Deep-Speech is an open source implementation library developed by

Baidu, which uses sophisticated and cutting-edge machine learning technology to create a speech-to-text engine [34], which facilitates developers to migrate subtasks in specific fields.

2.5 Reading Voice Evaluation Criteria

This paper mainly establishes the evaluation standard of reading speech from the two dimensions of accuracy and fluency. The accuracy is mainly measured from the phoneme similarity of the speaker's pronunciation, and the fluency is mainly evaluated from the fluency and coherence of the phonetic expression. In this paper, the proportion of accuracy is set in the range of 0.6 to 1, and the proportion of fluency is in the range of 0 to 0.4. According to this evaluation system, a reading speech evaluation measurement method is established [35], assuming that the accuracy score is A, and the fluency score is F, the total score of reading voice is R, W_1 and W_2 are the coefficients of accuracy and fluency respectively [36], and the formula is as follows:

$$R = A * W_1 + F + W_2 \tag{1}$$

Among them, $\sum_{i=1,2} W_i = 1$, assuming that W_1 is linearly related to A, and the value interval of W_1 is [0.6, 1], and expressed in Eq. (2):

$$R = -\lambda A^2 + \lambda A \times F + A \tag{2}$$

where, $\lambda = 1 - W_1$.

3 Proposed Model

In this paper, a reading speech evaluation model based on the combination of DeepSpeech and LSTM is designed. The model converts the read speech waveform file into a spectrogram as the unified input of the model.

When $\lambda = 0.4$, the overall frame diagram of the network model is shown in Fig. 1.

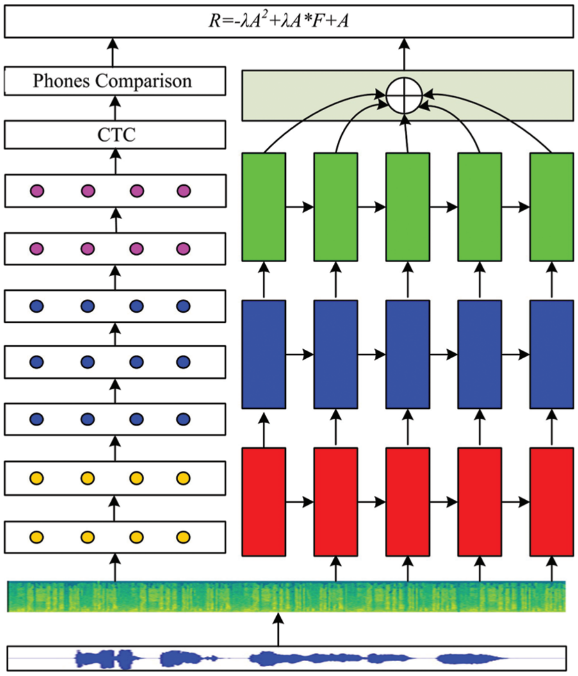

Figure 1: Proposed framework

The DeepSpeech branch model is first used to input the map into the convolutional neural network, and then passes through the GRU network layer, the multi-layer fully convolutional network, and then enters the Connectionist Temporal Classification (CTC) to obtain the phoneme sequence, which is compared with the training samples. The value of similarity between 0~1.

The fluency evaluation model is to input the graph into the three-layer LSTM network, and then enter the attention mechanism layer to obtain the fluency evaluation result. Finally, the accuracy and fluency are combined to obtain the final evaluation result.

3.1 Deep Speech Model

The DeepSpeech model is Baidu's open source Deepspeech2 neural network model based on Paddle. Its functions include feature extraction, data enhancement, model training, language model, decoding module, etc. It is powerful and easy to use. It is based on the end-to-end speech recognition technology of LSTM-CTC, and introduces LSTM modeling and CTC training in the field of machine learning into the traditional speech recognition framework.

3.1.1 Spectrogram Input

The spectrogram's abscissa represents time, the ordinate represents frequency, and the coordinate point value represents the energy of voice data. Because the two-dimensional plane is used to transmit three-dimensional information, the depth of the color represents the size of the energy value. The darker the hue, the stronger the point's communicating intensity.

The voice file will be read aloud at 20 ms/frame in this paper, and the windowing procedure will be done through the Hamming window. In each frame, the energy value of each frequency is determined using the fast Fourier transform, with a step size of 10 ms, and the spectrogram of each frame is generated independently. Finally, all of the spectrograms are spliced together in chronological sequence as the input to the two neural network models in this article, corresponding to this reading voice spectrogram. Fig. 2 depicts the spectrogram's procedure.

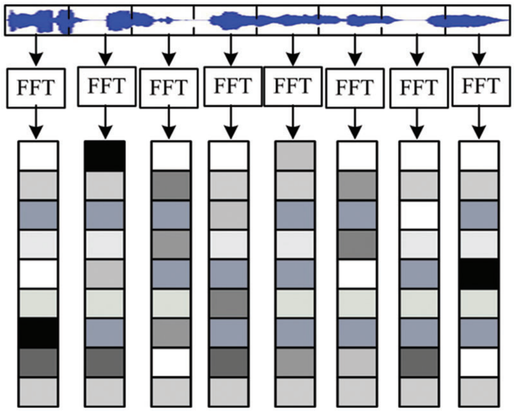

Figure 2: Voice spectrum generation

3.1.2 Gated Recurrent Unit

GRU is an enhanced version of LSTM network. Compared with LSTM network, its structure is simpler and it is a popular network at present. Since GRU is a variant of LSTM, it can also solve the long dependency problem in RNN network. Unlike LSTM, there are only two gates in GRU model: update gates and reset gates. The structure of the GRU model is shown in Fig. 3.

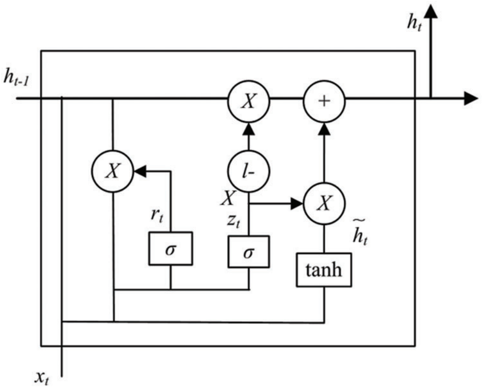

Figure 3: GRU model structure

where *zt* and *rt* in the figure represent the update gate and the reset gate, respectively. The update gate is used to govern how much state information from the previous instant is brought into the present state; the bigger the update gate's value, the more previous state information is brought in. The reset gate controls how much previous state information is written to the current candidate set $\tilde{h}_t$. The smaller the value of the gate, the less information from the previous state is written.

3.1.3 Connectionist Temporal Classification

The Connectionist Temporal Classification algorithm can be understood as a neural network-based time series classification. The output sequence y is obtained from the input sequence x. For example, the distribution $p(I/x)$ of the output sequence can be obtained, and the one with the highest probability is selected. As the output sequence, as shown in Eq. (3).

$$h(x) = \arg_{I\in L\leq T}^{\max} p(I/x) \tag{3}$$

Many practical sequence learning tasks may contain noise, i.e., serialized data without prior alignment. The CTC is to calculate a loss value and the main advantage is that it can automatically align unaligned data.

3.1.4 Phoneme Similarity Judgment

This paper measures the accuracy of reading aloud by comparing the similarity between the phone sequence of the phonetic conversion into Chinese and the correct phone sequence.

In this paper, a similarity comparison matrix is constructed, which consists of 23 initials and 24 finals. Because it involves the measurement of pronunciation similarity, the tones of the finals are not considered. The designed similarity comparison is listed in Table 1.

Table 1: Similarity comparison matrix

	b	p	m	F	…	ing	ong
B	0	10	24	100	…	100	100
P	10	0	18	30	…	100	100
M	24	18	0	16	…	100	100
F	100	30	16	0	…	100	100
…	…	…	…	…	…	…	…
Ing	100	100	100	100	…	0	30
ong	100	100	100	100	…	30	0

The similarity between initials and finals is represented by a number between 0 and 100. The larger the number, the lower the similarity. After identifying the phonemes through the model, the reader goes to the confusion matrix to search. The corresponding position is 1, the other positions are 0, the generation matrix B, the phoneme accuracy matrix $C =$ A.B, the sum of all elements in C is the accuracy value of the phoneme, and the average of all phoneme accuracy of a speech is the whole speech accuracy evaluation value.

3.2 Fluency Assessment Model

The fluency evaluation model uses a three-layer LSTM network, the spectrogram is used as the input, the hidden nodes of each layer of LSTM are set to 256, and the output of the third layer is connected to an attention layer to obtain the final fluency evaluation result.

3.2.1 LSTM

The traditional RNN is prone to the problem of gradient disappearance or gradient explosion. Therefore, on this basis, Hochreiter and Schmidhuber proposed the LSTM model, which increased the units for storing long-term valid data, thereby overcoming the gradient problem and improving the prediction ability.

As a more powerful RNN neural network model, LSTM introduces a mechanism of long-term information validity in LSTM, and this information is selectively controlled and preserved. The strategy adopted by LSTM is to add in each neuron: input gate, output gate and forget gate. Select the error function feedback weight, and decide whether the memory unit is cleared through the forget gate. The default LSTM method is expressed in Eq. (4).

$$
\begin{aligned}
f_t &= \sigma\left(W_f[h_{t-1},\ x_t] + b_f\right) \\
i_t &= \sigma(W_i[h_{t-1},\ x_t] + b_i) \\
\tilde{C}_t &= \tanh\left(W_{\tilde{C}}[h_{t-1},\ x_t] + b_{\tilde{C}}\right) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
o_t &= \sigma(W_o[h_{t-1},\ x_t] + b_o) \\
h_t &= o_t * \tanh(C_t)
\end{aligned}
\tag{4}
$$

Among them, W_f, W_i, $W_{\tilde{C}}$, W_o are the weight parameters; b_f, b_i, $b_{\tilde{C}}$, b_o are biases; x_t is taken as the input sequence, combined with the state of the previous hidden layer h_{t-1}, to form the forget gate f_t through the activation function. The input gate layer i_t and the output gate o_t are also calculated by x_t and h_{t-1}. The forget gate f_t is combined with the previous cell state C_{t-1} to determine whether to discard the information.

3.2.2 Attention Mechanism

Considering that the human brain's perception of things is a process of selective concentration, this attention mechanism can be applied to the field of deep learning, and attention can be described as a "selection mechanism for allocating limited information processing capabilities". It helps to quickly analyze the target data, and cooperate with the information screening and weight setting mechanism to improve the computing power of the model.

For each vector x_i in the sequence of input x, the attention weight α_i can be calculated according to Eq. (5):

$$\alpha_i = \frac{e^{(f(x_i))}}{\sum_j e^{(f(x_i))}} \tag{5}$$

where $f(x_i)$ is the scoring function.

The output of the attention layer, $attentive_x$ is the weighted sum of the input sequence. As shown in Eq. (6).

$$attentive_x = \sum\nolimits_i \alpha_i x_i \tag{6}$$

4 Experimental Results

4.1 Data Collection

The data from the "Export Chengzhang" software's children's reading speech corpus is utilized to examine the effect of the speech reading evaluation model in this study. The researchers gathered 400 participants (5–12 years old, on average 9 years old, 50 percent male, 50 percent female), who each read the necessary text aloud in a calm location and collected their varied vocal signals. At the same time, six broadcasting specialists are asked to assess the volunteers' fluency after listening to their speech data. Otherwise, experts must repeat the score in the second round, and if no result is obtained in the third round, the sample data will be dismissed without further consideration.

4.2 Model Parameter Settings

The model parameters involved are mainly related to LSTM and attention mechanism. Among them, the model adopts one-way three-layer LSTM, the Batch_size is 150, the maximum Epochs is 10000, the learning rate is 0.001, and the Dropout is 0.5.

The initialization weight method is RandomUniform, the neuron activation function is Tanh, the optimizer is Adam, and the loss function is the mean square error.

4.3 Result Evaluation

This paper uses the Tensorflow framework to build the network model structure, so as to evaluate the children's reading voice. Taking the original DeepSpeech model + traditional LSTM model as the baseline, compared to Model 1: Original DeepSpeech model + two-layer LSTM, model 2: Original DeepSpeech model + three-layer LSTM + attention mechanism, model 3: Improved DeepSpeech model + three-layer LSTM, model 4: In the current system (improved DeepSpeech model + three-layer LSTM + attention mechanism), the relative error between the prediction results of each model and the real score is calculated respectively. The calculation formula is shown in Eq. (7):

$$\sigma = \sqrt{\frac{1}{n-1}\sum\nolimits_{i=1}^{n} (y_i - t_i)^2} \tag{7}$$

Among them, σ represents the root mean square of the error between the scoring result and the true result, y_i is the model scoring value, and t_i is the true value.

Fig. 4 shows the scoring data of each model whose voice ID is 1001–1005. It can be seen from Fig. 4 that the difference between the score value of Model 4 and the true value is relatively the smallest.

After the experimental verification, the root mean square error (RMSE) of the prediction results of different models is listed in Table 2.

Table 2 shows that the technique described in this study outperforms all other models and has the highest accuracy. The second goal is to enhance the DeepSpeech model + three-layer LSTM model; the original DeepSpeech model's accuracy is poor, and the baseline model's accuracy is the worst. It can be determined that the improved DeepSpeech model has a great impact on the results, and the attention mechanism also improves the accuracy to a certain extent, which is basically consistent with the reading speech evaluation standard.

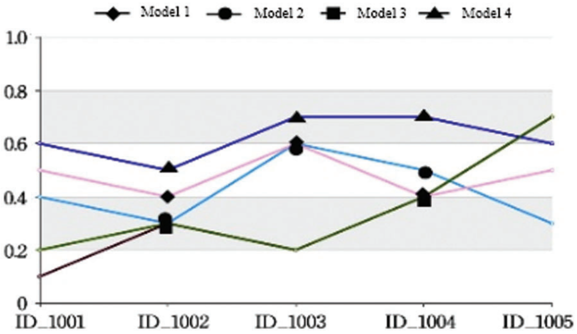

Figure 4: Statistics of predicted values and expert scores

Table 2: Comparison of the RMSE of algorithms

Model	RMSE (%)
Baseline: Original DeepSpeech Model + Traditional LSTM	18.55
Original DeepSpeech Model + Double Layer LSTM	15.52
Original DeepSpeech model + three-layer LSTM + attention mechanism	15.3
Improved DeepSpeech model + three-layer LSTM	12.25
Current system (improved DeepSpeech model + three-layer LSTM + attention mechanism)	11.89

5 Conclusion

Aiming at the evaluation of children's speech reading aloud, based on the speech corpus, this paper designs an improved model combining DeepSpeech and LSTM neural networks, and adds an attention mechanism to characterize speech recognition and spectrogram channels respectively. Extract and use the reading speech evaluation model to form a complete set of regression problems to solve.

Through the experimental verification, the model proposed in this paper has high accuracy, and the mean square error value is easy to converge. More specifically, the scoring and RMSE of the proposed method are superior to the existing algorithms, which makes it a significant candidate to be deployed in speech recognition applications.

The limitation of the proposed method is that it lacks the opportunity to consider the entire database, which will increase the algorithm complexity.

In the future, we will further integrate children's reading voice information to obtain more authoritative data and establish a reading voice database to provide corresponding interfaces.

Acknowledgement: The authors would like to thanks the editors and reviewers for their review and recommendations. The correspondence should be addressed to: Mohamad Khairi Ishak, khairiishak@usm.my.

Funding Statement: The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4170008DSR06).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Zhang, “The implementation of an english word learning system feedback system and smartphone app,” *Computer Systems Science and Engineering*, vol. 35, no. 3, pp. 207–214, 2020.

[2] M. Mustaqeem, M. Ishaq and S. Kwon, “Short-term energy forecasting framework using an ensemble deep learning approach,” *IEEE Access*, vol. 9, no. 4, pp. 94262–94271, 2021.

[3] S. Isobe, S. Tamura, S. Hayamizu, Y. Gotoh and M. Nose, “Multi-angle lipreading with angle classification-based feature extraction and its application to audio-visual speech recognition,” *Future Internet*, vol. 13, no. 7, pp. 1–18, 2021.

[4] Y. Song, D. Zhang, Q. Tang, S. Tang and K. Yang, “Local and nonlocal constraints for compressed sensing video and multi-view image recovery,” *Neurocomputing*, vol. 406, no. 3, pp. 34–48, 2020.

[5] D. Zhang, S. Wang, F. Li, S. Tian, J. Wang *et al.,* “An efficient ECG denoising method based on empirical model decomposition, sample entropy, and improved threshold function,” *Wireless Communications and Mobile Computing*, vol. 3, no. 5, pp. 1–11, 2020.

[6] F. Li, C. Ou, Y. Gui and L. Xiang, “Instant edit propagation on images based on bilateral grid,” *Computers, Materials & Continua*, vol. 61, no. 2, pp. 643–656, 2019.

[7] Y. Song, Y. Zeng, X. Y. Li, B. Y. Cai and G. B. Yang, “Fast CU size decision and mode decision algorithm for intra prediction in HEVC,” *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2001–2017, 2017.

[8] D. Zhang, J. Hu, F. Li, X. Ding, A. K. Sangaiah *et al.,* “Small object detection via precise region-based fully convolutional networks,” *Computers, Materials & Continua*, vol. 69, no. 2, pp. 1503–1517, 2021.

[9] J. Wang, Y. Zou, P. Lei, R. S. Sherratt and L. Wang, “Research on recurrent neural network based crack opening prediction of concrete dam,” *Journal of Internet Technology*, vol. 21, no. 4, pp. 1161–1169, 2020.

[10] J. Zhang, J. Sun, J. Wang and X. G. Yue, “Visual object tracking based on residual network and cascaded correlation filters,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 8, pp. 8427–8440, 2021.

[11] S. He, Z. Li, Y. Tang, Z. Liao, F. Li *et al.,* “Parameters compressing in deep learning,” *Computers, Materials & Continua*, vol. 62, no. 1, pp. 321–336, 2020.

[12] S. R. Zhou and B. Tan, “Electrocardiogram soft computing using hybrid deep learning CNN-ELM,” *Applied Soft Computing*, vol. 86, no. 3, pp. 1067–1078, 2020.

[13] S. R. Zhou, M. L. Ke and P. Luo, “Multi-camera transfer GAN for person re-identification,” *Journal of Visual Communication and Image Representation*, vol. 59, no. 1, pp. 393–400, 2019.

[14] W. Wang, H. Liu, J. Li, H. Nie and X. Wang, “Using CFW-net deep learning models for X-ray images to detect COVID-19 patients,” *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 199–207, 2021.

[15] W. Wang, Y. Yang, J. Li, Y. Hu, Y. Luo *et al.,* “Woodland labeling in chenzhou, China, via deep learning approach,” *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 1393–1403, 2020.

[16] S. Ezzat, N. Gayar and M. Ghanem, “Sentiment analysis of call centre audio conversations using text classification,” *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 4, no. 1, pp. 619–627, 2012.

[17] S. Byun and S. Lee, “A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms,” *Applied Sciences Journal*, vol. 11, no. 4, pp. 1–15, 2021.

[18] K. Pilaro, M. Shafiee, Y. Cao, L. Lao and S. Yang, “A review of kernel methods for feature extraction in nonlinear process monitoring,” *Processes Journal*, vol. 8, no. 1, pp. 1–17, 2020.

[19] A. Sophokleous, P. Christodoulou, L. Doitsidis and S. Chatzichristofis, “Computer vision meets educational robotics,” *Electronics Journal*, vol. 10, no. 6, pp. 1–18, 2021.

[20] M. Chowdary, T. Nguyen and D. Hemanth, “Deep learning-based facial emotion recognition for human-computer interaction applications,” *Neural Computing and Applications*, vol. 8, no. 2, pp. 982–993, 2021.

[21] M. Gerosa, S. Lee and D. Giuliani, “Analyzing childrens speech: An acoustic study of consonants and consonant-vowel transition,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, New York, pp. 1393–1396, 2006.

[22] A. Holm, O. Reyk, S. Crosbie, S. Bono, A. Morgan *et al.,* “Preschool childrens consistency of word production,” *Clinical Linguistics & Phonetics*, vol. 7, no. 4, pp. 1759–1768, 2021.

[23] A. Hagen, B. Pellom and R. Cole, "Highly accurate childrens speech recognition for interactive reading tutors subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[24] S. Shahnawazudding, W. Ahmad, N. Adiga and A. Kumar, "Childrens speaker verification in low and zero resource conditions," *Digital Signal Processing*, vol. 116, no. 2, pp. 1031–1045, 2021.

[25] H. Zhang, H. Huang and H. Han, "A novel heterogeneous parallel convolution bi-lstm for speech emotion recognition," *Applied Sciences Journal*, vol. 11, no. 21, pp. 1–16, 2021.

[26] A. Badshah, J. Ahmad and N. Rahim, "Speech emotion recognition from spectrogram with deep convolutional neural network," in *IEEE Int. Conf. on Platform Technology & Service*, Ottawa, Canada, pp. 53–59, 2017.

[27] M. Farooq, F. Hussain, N. Baloch, F. Raja, H. Yu *et al.,* "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors Journal*, vol. 20, no. 21, pp. 1–14, 2020.

[28] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Sing *et al.,* "Tow-way feature extraction for speech emotion recognition using deep learning," *Sensors Journal*, vol. 22, no. 6, pp. 1–19, 2022.

[29] J. Kang, W. Zhang and J. Liu, "Gated recurrent units based hybrid acoustic models for robust speech recognition," in *IEEE 10th Int. Symp. on Chinese Spoken Language Processing (ISCSLP)*, Shanghai, China, pp. 1–6, 2016.

[30] M. Cara, C. Lobos, M. Varas and O. Torres, "Understanding the association between musical sophistication and well-being in music students," *International Journal of Environmental Research and Public Health*, vol. 19, no. 7, pp. 1–19, 2022.

[31] S. Mirsamad, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE ICASSP Conf.*, Hong Kong, China, pp. 1–5, 2017.

[32] N. Passalis and A. Tefas, "Neural bag-of-features learning," *Pattern Recognition*, vol. 64, no. 3, pp. 277–294, 2017.

[33] L. Kaushik, A. Sangwan and J. Hansen, "Automatic sentiment detection in naturalistic audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1668–1679, 2017.

[34] C. Yu, Y. Chen, Y. Li, M. Kang, S. Xu *et al.,* "Cross-language end-to-end speech recognition research based on transfer learning for the low-resource tujia language," *Symmetry Journal*, vol. 11, no. 2, pp. 1–15, 2018.

[35] M. Mustaqeem and S. Kwon, "Att-net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, no. 3, pp. 1071–1083, 2021.

[36] M. Mustaqeem and S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Systems with Applications*, vol. 167, no. 5, pp. 1147–1158, 2021.