Journal of Multimedia Information System VOL. 10, NO. 3, September 2023 (pp. 249-258): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2023.10.3.249

A Robust Online Korean Teaching Support Technology Based on TCNN

Shunji Cui[1*]

Abstract

The emergence and development of multimedia forms provide technical support for online Korean language teaching. However, in many aspects, there are still many problems in online Korean teaching, such as noise interference, inaccurate translation, and unstable translation models. In this paper, we propose a Korean speech enhancement model based on temporal convolutional neural network and GRU neural network. We explore a Korean speech enhancement technology based on deep neural network, to make Korean speech teaching clearer and smoother, and to provide a robust support technology for online Korean teaching. First, we construct a temporal convolutional neural network to process and extract temporal feature in language data. Second, we introduce the sliding window mechanism and the maximum pooling structure to extract the feature in the speech time series data effectively and reduced the data scale. Third, we employ the Bi-GRU neural network and encoder-decoder for temporal data enhancement, which effectively avoids the problem that the hidden layer information cannot be effectively used in the traditional model, thereby improving the prediction accuracy and speed of speech data. The experimental outcomes demonstrate the effective evaluation performance of the method proposed in this paper.

Key Words: Online Korean Teaching, Bi-GRU, Neural Networks, Data Scale, Evaluation Performance.

I. INTRODUCTION

Language and culture always go hand in hand and complement each other [1-2]. Language is the carrier of cultural dissemination and development, and culture endows language with different backgrounds and connotations. Through the relevance of language and culture, people's understanding and digestion of other languages can be deepened. However, due to their geographical proximity, China and South Korea are both within the eastern cultural circle and have similar language and cultural foundations. Bilingual teaching plays an essential role in the formation and development of the Korean language. It reflects the collision and fusion of Chinese and foreign cultures as a brand-new teaching method.

Information-based teaching is a new form of modern teaching [3-4], which is characterized by the support of information technology. Informatization teaching design is to promote the development of students' problem-solving ability under the guidance of today's educational concepts and according to the new characteristics of the times. Based on multimedia and network media, the teaching strategy is designed with the problem situation as the core to realize the teaching planning. The purpose is to encourage students to use the information environment to cooperate in inquiry, practice, and problem-solving thinking activities, to cultivate students' autonomous learning ability and practical ability. Compared with the traditional teaching design, the teaching design under the information environment pays more attention to the main role of the learners. It is not limited to classroom teaching form and subject knowledge system, but combines teaching objectives into new teaching activity units, which requires teachers to change their roles.

In today's increasingly interconnected world, language is no longer a solitary barrier, but rather a bridge that connects diverse cultures and societies. Especially in the era of digitization and remote learning trends, more individuals seek opportunities to learn new languages transcending geographical boundaries [5-6]. Korean, as a highly sought-after Asian language, captivates learners worldwide who aspire to grasp not only the language but also gain deeper insights into Korean culture, history, and society [7-8]. Nevertheless, mastering a new language remains a multifaceted challenge, particularly within the realm of online learning. Deep learning models play a very important role in language processing and translation research, and they provide the technical foundation and support for online teaching of many languages.

Manuscript received July 14, 2023; Revised August 29, 2023; Accepted September 01, 2023. (ID No. JMIS- 23M-07-029)
Corresponding Author (*): Shunji Cui, +86-15904333372, cunsj188@21cn.com
[1]Jilin Agricultural Science and Technology University, Jilin, China, cunsj188@21cn.com

Addressing these challenges, the robust online Korean teaching support technology based on Temporal Convolutional Neural Network (TCNN) emerges as a solution. With rapid advancements in artificial intelligence and natural language processing [9,10], coupled with the successful integration of neural networks in education, a unique opportunity arises to revolutionize online Korean education. This technology goes beyond mere substitution of traditional teaching methods—it represents a novel paradigm, fusing the expertise of educators with the computational power of neural networks. Through deep learning models, the system can intricately analyze learners' needs, weaknesses, and learning styles, tailoring personalized learning paths and instructional content. This promises to significantly enhance learning efficiency and outcomes, empowering learners to grasp Korean with greater confidence and autonomy.

Hence, the motivation behind this paper lies in exploring and developing a Korean online teaching support technology grounded in TCNN, aimed at providing a potent, flexible, and personalized learning platform for the vast community of Korean learners. By amalgamating artificial intelligence with education, we stand to overcome the limitations of traditional instruction, rendering the conquest of language barriers feasible and facilitating the exchange and understanding of cultural diversity. To enhance the online teaching experience for Korean learners, we present a novel approach in this study: a Korean speech enhancement model leveraging the power of TCNN [11,12] and GRU Neural Network [13,14]. Our proposed model consists of two key components. Firstly, we employ a temporal convolutional neural network comprising a temporal convolutional layer and a max pooling layer. By incorporating a sliding window mechanism and a maximum pooling structure, we are able to effectively extract the temporal patterns present in the speech time series data while reducing the overall data scale. Secondly, we integrate the GRU neural network for time series prediction. This approach addresses the challenge faced by traditional models, where the utilization of hidden layer information may be limited. By leveraging GRU's ability to capture temporal dependencies, our model significantly enhances the accuracy and speed of speech data prediction. Through the fusion of TCNN and GRU, our Korean speech enhancement model offers a powerful solution to optimize online teaching experiences.

The subsequent sections of this paper are structured as follows. In Section 2, we delve into the realm of deep learning neural networks by exploring related research. Section 3 outlines our proposed method for enhancing Korean speech, encompassing preprocessing and feature extraction techniques, alongside the architecture design involving TCNN and Bi-LSTM models. Moving on to Section 4, we present the experimental studies and outcomes that substantiate the efficacy of our proposed model for enhancing Korean speech. Lastly, Section 5 encapsulates our conclusions drawn from the entirety of this study.

II. RELATED WORK

Within this section, we aim to provide a foundation of fundamental knowledge concerning language processing models. We will explore several prominent deep learning models utilized in the field, specifically focusing on Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Gated Recurrent Unit (GRU) network models. To begin, Convolutional Neural Networks (CNNs) have gained significant popularity in language processing. Originally developed for image recognition tasks, CNNs have been adapted to effectively process sequential data, such as text. By leveraging their ability to capture local patterns through convolutional operations, CNNs have proven valuable in tasks such as text classification, sentiment analysis, and machine translation. Moving on, Recurrent Neural Networks (RNNs) offer a unique approach to language processing. They are designed to model sequential data by incorporating recurrent connections that allow information to persist across different time steps. RNNs excel in tasks involving sequential dependencies, making them suitable for tasks such as language modeling, speech recognition, and named entity recognition. Furthermore, Gated Recurrent Unit (GRU) network models provide an enhancement to traditional RNNs. GRUs introduce gating mechanisms that control the flow of information within the network, allowing them to better capture long-term dependencies and mitigate the vanishing gradient problem. GRUs have demonstrated success in various language processing tasks, including machine translation, text summarization, and sentiment analysis. By familiarizing ourselves with these language processing models, we can establish a solid understanding of the underlying principles and techniques employed in the field. This knowledge forms the basis for exploring more advanced topics and applications in language processing.

The Convolutional Neural Network (CNN) was initially proposed by LeCun [15] in 1998 as a neural network model. CNNs are widely recognized and extensively utilized in the field of computer vision as a classic deep learning model. In comparison to traditional neural networks, CNNs introduce convolutional layers and pooling layers. The primary purpose of the convolutional layer within a CNN is to perform convolution operations on both the input image feature map and the convolution kernel. By employing a sliding window with a smaller size than the original input, local features are extracted from the input. These local features are then combined at higher layers to generate global features representing the entire input image. The convolutional

layer leverages a sparse connection and weight-sharing mechanism, effectively reducing the number of parameters involved. In addition to the convolutional layer, CNNs incorporate pooling layers, typically applied after the convolution operation. The pooling layer primarily serves the purpose of downsampling the data. By reducing the dimensionality of the data, pooling layers facilitate the extraction of essential features while simultaneously reducing computational complexity. Overall, CNNs have proven to be highly effective in computer vision tasks due to their ability to exploit local correlations and hierarchical representations. Through the integration of convolutional and pooling layers, CNNs excel at capturing intricate patterns within images and significantly outperform traditional neural networks in various computer vision applications. The purpose of pooling is to calculate sufficient local statistics of features, thereby reducing the number of overall features, preventing overfitting, and reducing the amount of computation. At the same time, by increasing the observation window of the convolutional layer, a hierarchy of spatial filters is introduced.

The Temporal Convolutional Neural Network (TCNN) is a distinctive iteration of the convolutional neural network paradigm, utilizing a one-dimensional convolution kernel [16]. Tailored for processing time series data within the temporal domain, TCNN fundamentally considers the temporal dimension akin to a spatial dimension. Illustrated in Fig. 1, TCNN's framework and operation involve segmenting time series data into localized sequential segments, followed by applying a one-dimensional convolution kernel to compute on neighboring signals within each segment. This design choice facilitates efficient analysis of sequential patterns within the input sequence. This process generates an output tensor that captures the learned temporal features. The use of a one-dimensional convolution kernel in TCNN allows for efficient and effective analysis of time series

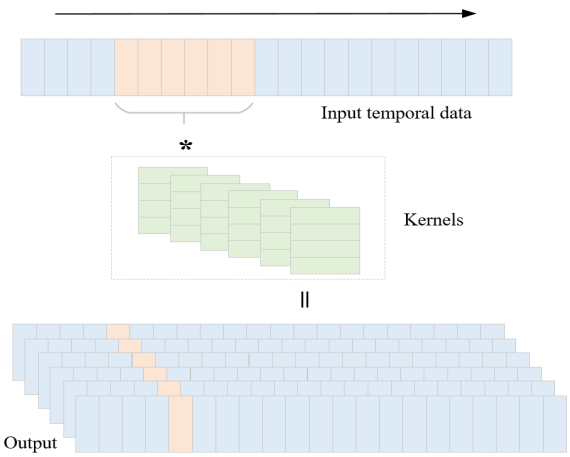

Fig. 1. The working principle of the temporal convolutional neural network.

data. By considering the temporal dimension as a spatial dimension, TCNN can leverage the inherent sequential nature of the data and exploit local dependencies. This approach facilitates the extraction of meaningful patterns and relationships within the time series, enabling accurate and insightful analysis. The temporal convolution is calculated as follows.

$$y_t = \sum_{k=1}^{N} w_k \cdot x_{t-k+1}, \tag{1}$$

where y_t represents the temporal convolution value at time t, and x_{t-k+1} represents the input sequence $x = \{x_1, x_2, \ldots, x_n\}$ the t-k+1-th value, w_k represents the weight parameter in the one-dimensional convolution kernel, and N is the length of the convolution kernel.

In the TCNN, each output tensor is computed by convolving the same discrete convolution kernel with local sequence segments positioned at various locations. Consequently, the parameters of the convolution kernel remain fixed throughout the computation process. Therefore, the translation of temporal features in the time dimension will not affect the time domain CNN. Similar to traditional convolutional neural networks, temporal convolutional neural networks are translation-invariant in the time dimension and can identify local feature patterns implicit in local sequence segments. Since the temporal convolutional neural network does not retain the order information of the input time step when processing the input sequence, in order to add timing sensitivity, we introduce a bidirectional GRU network to further process the output of the time-domain convolutional neural network.

TCNNs have gained prominence for their adeptness in modeling sequential data. They extend CNNs to the time domain, enabling effective processing of sequences like time-series data and sequential text. Pioneering work by [17] harnessed TCNNs for natural language tasks, achieving competitive results in text analysis. Chen et al. (2018) introduced dilated convolutions to capture long-range dependencies in multivariate time-series data [18]. Zhao et al. applied TCNNs to action recognition in videos, showcasing their prowess in spatiotemporal feature modeling [19]. TCNNs have since found utility in speech recognition, music generation, and healthcare [20-21]. This study builds on these achievements by proposing a TCNN-based approach for personalized online Korean teaching support.

RNN are a class of neural networks with internal loops [17]. The sequence processing method is to first traverse the sequence elements and generate a hidden layer state, which contains pattern information related to historical data. It can preserve sequence context and is often used to process se-

quence data. In traditional RNNs, a common challenge arises when dealing with long input sequences—the issue of vanishing gradients. This phenomenon hampers the network's capacity to adeptly capture extended dependencies, commonly termed as the long-term dependencies issue [18]. To tackle this challenge, an encouraging remedy involves incorporating a gating mechanism that governs information flow and aggregation. This mechanism enables the network to selectively incorporate new information while also selectively forgetting previously accumulated information. Networks employing such gating mechanisms are commonly referred to as Gated RNN. By incorporating gates into the network architecture, Gated RNNs can effectively capture and retain relevant information over longer sequences. The gating mechanism allows the network to control the flow of gradients, thereby alleviating the issue of vanishing gradients. This enhanced architecture enables Gated RNNs to better model long-distance dependencies, making them particularly effective in tasks that require capturing contextual information from extensive sequences. Two common recurrent neural networks based on gating mechanisms are LSTM [19] and GRU [20].

LSTM neural network was proposed by Hochreiter and Schmidhuber in 1997, which is a variant of the recurrent neural network [19]. The GRU neural network was proposed by Chung et al., which is further optimized based on LSTM, making it cheaper to run. GRU and LSTM can get the same accurate prediction results, and the GRU model contains fewer network parameters, so its computational performance is better, and the risk of overfitting is reduced. Both LSTM and GRU networks use memory modules or structures called "gates" to control the memory of historical sequence features.

III. PROPOSED METHOD

In this paper, we aim to explore a Korean speech enhancement technology based on deep neural network, to make Korean speech teaching clearer and smoother, and to provide a robust support technology for online Korean teaching. In our method, we propose a Korean speech enhancement model based on TCNN and GRU. We first construct a temporal convolutional neural network, including a temporal convolutional layer and a max pooling layer. Second, we introduce the sliding window mechanism and the maximum pooling structure to extract the feature in the speech time series data effectively and reduced the data scale. Third, we employ the GRU neural network for temporal data enhancement, which avoids that the hidden layer information cannot be used in the traditional model, to improve the prediction accuracy and speed of speech data. Next, we introduce our method in detail. First, we introduce preprocessing methods for text-to-speech data. Second, we describe the TCNN-based language enhancement model design.

3.1. Preprocess and Feature Extraction

In order to simplify the processing, the classical speech processing methods are generally based on the theoretical basis of the linear stationary system, which is based on the relative stability of the short-term speech. However, the speech signal is a typical nonlinear and non-stationary random process, which makes it difficult to further improve the performance of the speech processing system by using the classical processing method, such as the recognition rate of the speech recognition system. With the continuous development of robotics technology, the new applications of voice represented by intelligent voice interaction of robots urgently require the development of new voice processing technologies and means to improve the performance level of voice processing systems. In the past ten years, artificial intelligence technology is developing at an unprecedented speed. New technologies and new algorithms, especially for new neural networks and deep learning technologies, have greatly promoted the development of speech processing. The research provided new methods and technical means, and intelligent speech processing came into being. So far, there is no precise definition of intelligent speech processing. In a broad sense, all or part of the intelligent processing technology or means used in speech processing algorithms or system implementation can be called intelligent speech processing.

At the same time, speech is a rich information signal carrier, which carries a lot of information such as semantics, speakers, emotions, languages, dialects. The separation and perception of these information require a very fine analysis of speech. The discrimination of these information is no longer a simple rule of description. It is unrealistic to use artificial means to analyze the sound mechanism and simple characteristics of the signal. Similar to the idea of human language learning, machine learning means are used to make machines listen to a large amount of speech data and learn the rules contained in the speech data, which is the main means to effectively improve the performance of speech information processing. Different from classical speech processing methods, which are limited to extracting artificially set characteristic parameters, the most important feature of intelligent speech processing is that the idea of learning laws from data is reflected in the speech processing process or algorithm.

Preprocessing is an important link before data augmentation. In practical applications, the audio to be processed is often accompanied by background noise, pauses and blanks. Therefore, before extracting feature parameters, it

is usually important to process the audio data to ensure that the recognition system is not affected by the interference information in the speech signal and accurately capture the voiceprint features in the speech signal.

The physiological characteristics of human vocal organs determine that the energy of speech signals is mainly concentrated in low frequency bands. And when the voice reaches the high frequency of 8,000 Hz, the phenomenon of weakening occurs. In addition, compared with the low-frequency part, the high-frequency part of the speech signal is more susceptible to interference and loss during the propagation process. The function of pre-emphasis is to enhance the high-frequency part of the speech signal to make the spectrum more stable and facilitate the extraction of key speaker information contained in the high-frequency part of the speech signal. Using a first-order high-pass filter to filter speech signals is a common pre-emphasis method, which includes a key parameter, the pre-emphasis filter coefficient μ, whose calculation formula is as follows:

$$H(z) = 1 - \mu z^{-1}, \tag{2}$$

where z is the frequency domain representation of the discrete signal and μ represents set to 0.95.

The researchers believe that the characteristic parameters of the speech signal are basically unchanged within 10−30 ms, so the stable short-term speech signal is usually obtained by framing, which is convenient for Fourier transform analysis. In the framing process, it is necessary to set a fixed frame length and frame shift, that is, the time length of each frame and the time distance between frames. The frame length is generally greater than the frame shift, so there is an overlap between the speech frames, which makes the frequency domain characteristics of the speech signal change more smoothly in the time dimension. To avoid spectral leakage during Fourier transform, we perform windowing during speech preprocessing. We multiply each frame of speech signal by a window function to reduce the amplitude of the two ends of the time window. Window functions include Hamming window function [21] and rectangular window function [22]. These window functions have different shapes and main lobe widths. In speech signal processing, an appropriate window function can be selected according to the actual situation. We employ the Hamming window function in the speech preprocessing, which is defined as follows.

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) & 0 \le n \le M-1 \\ 0 & \text{others} \end{cases}. \tag{3}$$

3.2. Model Design Based on TCNN

Numerous studies have showcased the efficacy of fully convolutional neural networks (FCNNs) in the realm of temporal speech enhancement. Additionally, research efforts have explored the utilization of frequency-domain dropout in training models, aiming to enhance the perceptual quality of augmented speech in the time domain. However, it is important to note that these studies primarily focus on offline augmentation techniques and do not explicitly address real-time augmentation scenarios. This paper combines the TCNN model and encoder-decoder-based architecture [23] to model temporal speech enhancement, and proposes a real-time speech enhancement model, which provides technical support for online Korean language teaching.

After speech preprocessing, we get sequence data $x_1, x_2, x_3, \ldots, x_n$. The basic structure of our proposed temporal prediction model is shown in Fig. 2. Then it is analyzed and processed as the input of the bidirectional GRU network layer. This temporal data prediction model fully combines the speed advantage of time-domain convolutional neural networks in processing temporal data and the temporal sensitivity of gated recurrent units. The model proposed in this paper reduces the magnitude of the input data by introducing a time-domain CNN, and improves the model operation speed. On the other hand, the output of the time-domain convolutional neural network captures temporal features across a broader time span. This enables the bidirectional GRU network to examine previous temporal data and higher-resolution temporal information during

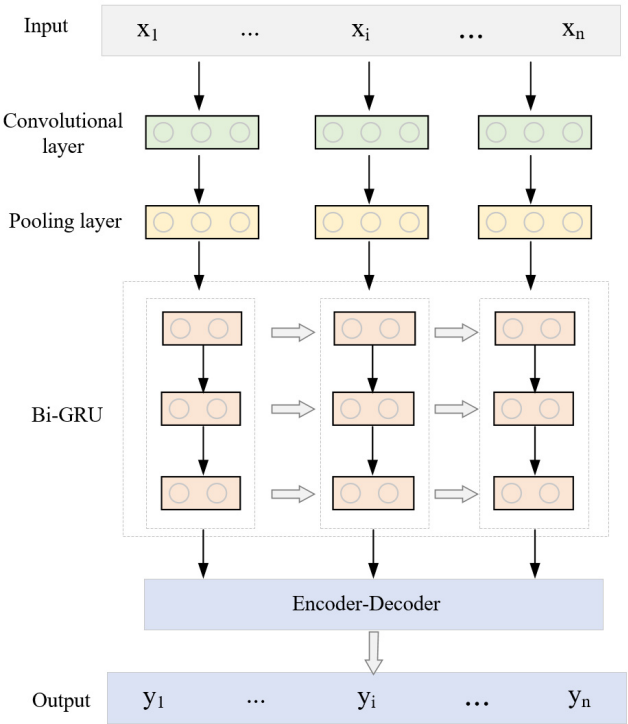

Fig. 2. The basic structure of our proposed temporal prediction model.

subsequent computations. As a result, the accuracy of temporal forecasting is enhanced, as the bidirectional GRU network can leverage this comprehensive temporal context for improved predictions.

In the temporal convolutional neural network processing layer, our approach employs 128 temporal convolution kernels with a size of 24 for convolution calculations. Concurrently, the max pooling layer utilizes a max pooling operator with a size of 3. Within the processing layer of the time-domain convolutional neural network, we introduce the Rectified Linear Unit (ReLU) activation function, a modified variant of the linear unit [24]. The ReLU function introduces sparsity by allowing certain neurons to output 0, promoting network efficiency and reducing parameter interdependencies to combat overfitting [11]. Moreover, the ReLU function exhibits a derivative of 1, effectively addressing the issue of gradient disappearance. Mathematically, the ReLU function is defined as follows:

$$f(x, w, b) = max(0, w^T x + b). \tag{4}$$

This equation signifies that the ReLU function produces an output of 0 when the input (x) is negative, and otherwise returns the input value as is. By incorporating the ReLU activation function, the time-domain convolutional neural network benefits from enhanced sparsity, improved generalization capabilities, and alleviated concerns regarding gradient vanishing. Its widespread utilization stems from its ability to facilitate efficient information flow and combat overfitting challenges, ultimately contributing to more effective and robust network performance.

The input data undergoes scaling and mapping through pooling operations within the model. In this particular architecture, the Max Pooling method is employed for this purpose. Through the application of maximum pooling, the algorithm identifies local maxima within the input features, effectively curtailing the count of learnable parameters and fortifying the model's resilience [25]. While diminishing the output data's dimensionality, this process of maximum pooling conserves the paramount feature details inherent in the input data [26]. Max pooling holds significance in two main aspects: firstly, it effectively reduces the dimensionality of the feature map, thereby minimizing the parameters required for subsequent layers. Secondly, it maintains translation invariance [27]. The definition is as follows.

$$p_i^{l+1}(j) = \underset{(j-1)W+1 \leqslant t \leqslant jW}{max} \{q_i^l(t)\}, \tag{5}$$

where $q_i^l(t)$ represents the value of the t-th neuron in the i-th feature vector of the l-th layer, W is the width of the pooling region, and $p_i^{l+1}(j)$ represents value calculated by the j-th pooling in the $l+1$-th layer.

Since the timing information will continue to decay when it is passed forward in the GRU network, the more advanced the sequence information is, the more serious the attenuation is. To overcome the problem of information decay, we improve the GRU network and train two GRUs in opposite directions at the same time to form a bidirectional GRU model [28]. The bidirectional GRU model is composed of two unidirectional GRUs superimposed together. The input at each time t will be simultaneously provided to two opposite GRU network layers for learning. The ultimate output of the model is a collective result derived from the outputs of the two individual unidirectional GRU network layers. In order to avoid overfitting of the model during training, we introduce a dropout mechanism in the gated recurrent unit processing layer [29]. This mechanism can enhance the generalization of the model because it does not depend too much on some local features. After experimental tests, when the dropout of the gated recurrent unit is set to 0.4, the model performance is optimal.

In our approach, we have developed an encoder-decoder structure to enhance speech data. This encoder-decoder model is comprised of 2D causal convolutional layers, while the intermediate module incorporates a combination of 1D causal convolutional layers and dilated convolutional layers. Fig. 3 illustrates the architecture of the encoder-decoder framework. The output of the encoder is reshaped into a one-dimensional signal with a size of T × 256. We then perform operations on this reshaped output to generate an output of the same size. The intermediate module comprises multiple stacked 1D convolutions. The decoder component mirrors the encoder structure and is composed of a series of 2D causally transposed convolutional (deconvolu-

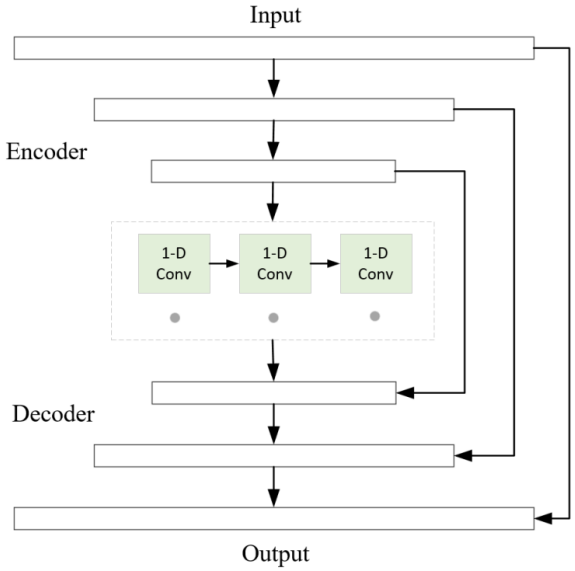

Fig. 3. The encoder-decoder framework for temporal data enhancement.

tional) layers. These deconvolutional layers aim to reconstruct the enhanced speech data by mapping the learned features back to the original dimensions. Overall, the encoder-decoder framework, consisting of 2D causal convolutional layers, intermediate modules with 1D convolutions and dilated convolutions, and the decoder with 2D causally transposed convolutional layers, allows for effective speech data enhancement and restoration.

IV. EXPERIMENTS AND RESULTS

In this section, we experimentally verify the proposed Korean speech enhancement model based on TCNN and GRU in this paper. First, we introduce the metrics for evaluating the model. Second, we analyze and compare the experimental results. On the one hand, in order to verify the model performance, we compare and analyze the method proposed in this paper with other method models. On the other hand, to verify the effectiveness of each module in the model, we conduct ablation experiments on each module.

Before feeding the experimental data into the neural network, we preprocess the speech. Since the numerical variation range of the data in each time series dataset is different, for the convenience of comparison, we standardize each temporal data. The processing technique employed in this approach utilizes the z-score standardization method. The approach encompasses the normalization of the mean and standard deviation of the initial time series data. This normalization procedure transmutes the temporal data to conform to a standard normal distribution with a mean of 0 and a standard deviation of 1. The definition is presented as follows.

$$x_i' = \frac{x_i - \mu_i}{\sigma_i}, \tag{6}$$

where x_i represents the i-th value in the original data, μ_i is the mean value, σ_i represents the standard deviation, and x_i' is the data after standardization.

The experiments were implemented on a server-grade machine equipped with a high-performance GPU, specifically an NVIDIA GeForce RTX 3090, and 32 GB of RAM. The machine was running the Ubuntu 20.04 operating system, which is widely used for deep learning tasks due to its compatibility with popular deep learning frameworks and libraries. For software infrastructure, we utilized Python 3.8 as the primary programming language, along with TensorFlow 2.5 as the deep learning framework. TensorFlow's GPU support was leveraged to expedite the training process by harnessing the parallel processing capabilities of the GPU. Additionally, we employed Keras, a high-level API integrated within TensorFlow, to facilitate model architecture design and training. To ensure efficient experimentation and reproducibility, the experiment code and configurations were managed using version control systems such as Git. This allowed us to track changes, collaborate effectively, and maintain a record of different experimental setups.

4.1. Performance Metrics

In our experiments, we introduce four evaluation indicators, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Squared Error (MSE) [30], and Signal-Noise Ratio (SNR). Through these evaluation indexes, the prediction error and prediction speed of each temporal prediction model on the testing dataset can be fully evaluated. They are respectively defined as follows.

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |(y_i - \hat{y}_i)|. \tag{7}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2. \tag{8}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}. \tag{9}$$

$$SNR = 10\log\frac{s}{n}. \tag{10}$$

Among them, MAE refers to the average distance between the model predicted value $\hat{y}_i$ and the sample real value y_i, and MAE can more accurately reflect the real situation of the predicted value error. The smaller the MAE value, the greater the prediction accuracy. MSE stands for Mean Squared Error, which represents the anticipated value of the squared disparity between the parameter's estimated and actual values. It serves as a metric to gauge the extent of data variation. A lower MSE value corresponds to enhanced accuracy of the prediction model. RMSE represents the sample standard deviation of the difference between the model predicted value $\hat{y}_i$ and the sample true value y_i. RMSE is more sensitive to abnormal data points in the test set, that is, if there is a predicted value that is very different from the true value, the RMSE will be large. The smaller the RMSE value, the greater the prediction accuracy.

4.2. Results and Analysis

In order to verify performance of our proposed method, we compare a series of temporal models, including RNN, LSTM, GRU, and encoder-decoder. Table 1 shows the performance of these methods on the above performance indicators. From the table, we can see that our method achieves the best results, which fully shows that the method proposed in this paper has good performance on the Korean speech

Table 1. The comparison results with other methods on the performance indicators.

Method	MAE	MSE	RMSE	SNR
RNN	0.2664	0.1543	0.3928	1.218
LSTM	0.2159	0.1247	0.3531	1.046
GRU	0.2634	0.1138	0.3373	0.945
Encoder-decoder	0.3167	0.1351	0.3676	1.348
Ours	0.1563	0.0723	0.2689	0.612

enhancement support. Compared with other methods, our proposed method improves SNR to 0.612 dB.

Additionally, to verify the function of each module, we conducted ablation experiments to analyze the function of each module. Specifically, we divide the model into three parts: convolution module, Bi-GRU module, and Encoder-Decoder module. We remove each module separately to compare the experimental results. Fig. 4 depicts the performance comparison of ablation experiments on four metrics. In Fig. 4, CDT 1 represents the removal of the convolution module, CDT 2 represents the removal of the Bi-GRU module, and CDT 3 represents the removal of the encoder-decoder module, respectively.

From the results in the Fig. 4, we can see that each module has a certain role in improving the performance of the model. While the Bi-GRU module and the encoder-decoder module have the greatest impact, which shows that the model proposed in this paper has good effectiveness.

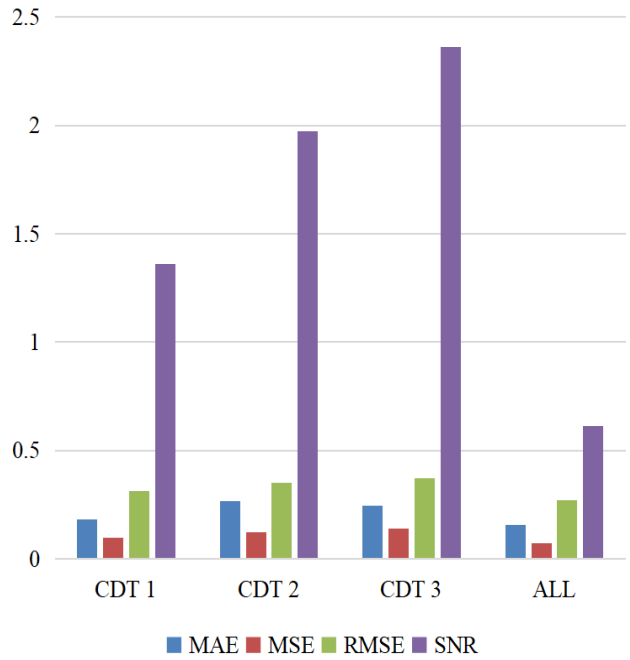

Fig. 4. The performance comparison of ablation experiments on four metrics.

V. CONCLUSION

In this paper, we propose a Korean speech enhancement model based on temporal convolutional neural network (TCNN) and GRU neural network. We explore a Korean speech enhancement technology based on deep neural network, to make Korean speech teaching clearer and smoother, and to provide a robust support technology for online Korean teaching. We first construct a temporal convolutional neural network, including a temporal convolutional layer and a max pooling layer. Second, we introduce the sliding window mechanism and the maximum pooling structure to extract the feature in the speech time series data effectively and reduced the data scale. Third, we employ the Bi-GRU neural network and encoder-decoder for temporal data enhancement, which effectively avoids the problem that the hidden layer information cannot be effectively used in the traditional model, thereby improving the prediction accuracy and speed of speech data.

REFERENCES

[1] M. Battiste, "Language and culture in modern society," *Reclaiming Indigenous Voice and Vision*, vol. 192, 2000.

[2] W. Jiang, "The relationship between culture and language," *ELT Journal*, vol. 54, no. 4, pp. 328-334, 2000.

[3] T. Dewett and G. R. Jones, "The role of information technology in the organization: A review, model, and assessment," *Journal of Management*, vol. 27, no. 3, pp. 313-346, 2001.

[4] J. Y. Bakos and M. E. Treacy, "Information technology and corporate strategy: A research perspective," *MIS Quarterly*, pp. 107-119, 1986.

[5] I. Shah and M. Khan, "Impact of multimedia-aided teaching on students' academic achievement and attitude at elementary level," *US-China Education Review A*, vol. 5, no. 5, pp. 349-360, 2015.

[6] P. F. Velleman and D. S. Moore, "Multimedia for teaching statistics: Promises and pitfalls," *The American Statistician*, vol. 50, no. 3, pp. 217-225, 1996.

[7] C. Warner and B. Dupuy, "Moving toward multiliteracies in foreign language teaching: Past and present perspectives… and beyond," *Foreign Language Annals*, vol. 51, no. 1, pp. 116-128, 2018.

[8] I. A. N. Moodie and A. Feryok, "Beyond cognition to commitment: English language teaching in South Korean primary schools," *The Modern Language Journal*, vol. 99, no. 3, pp. 450-469, 2015.

[9] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3-4, pp. 197-387, 2014.

[10] L. Deng, J. Li, J. T. Huang, K. Yao, D. Yu, and F. Seide,

et al,. "Recent advances in deep learning for speech research at Microsoft," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8604-8608.

[11] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sensing*, vol. 11, no. 5, p. 523, 2019.

[12] A. Pandey and D. L. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain", in *Proceedings of the ICASSP 2019−2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6875-6879.

[13] R. Fu, Z. Zhang, and L. Li. "Using LSTM and GRU neural network methods for traffic flow prediction", in *Proceedings of the 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 2016, pp. 324-328.

[14] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks", in *Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS).* 2017, pp. 1597-1600.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[16] A. Haidar and B. Verma, "Monthly rainfall forecasting using one-dimensional deep convolutional neural network", *IEEE Access,* vol. 6, pp. 69053-69063, 2018.

[17] S. Yadav and A. Rai "Frequency and temporal convolutional attention for text-independent speaker recognition", in *Proceedings of the ICASSP 2020−2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6794-6798.

[18] R. Chen, X. Yan, S. Wang, and G. Xiao, "DA-Net: Dual-attention network for multivariate time series classification", *Information Sciences*, vol. 610, pp. 472-487, 2022.

[19] H. Zhao, W. Xue, X. Li, Z. Gu, L. Niu, and L. Zhang, "Multi-mode neural network for human action recognition", *IET Computer Vision*, vol. 14, no. 8, pp. 587-596, 2020.

[20] Q Wang and C. Li, "Incident detection and classification in renewable energy news using pre-trained language models on deep neural networks", *Journal of Computational Methods in Sciences and Engineering,* vol. 22, no. 1, pp. 57-76, 2022.

[21] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Soujanya, "A review of deep learning techniques for speech processing", *Information Fusion*, p. 101869, 2023.

[22] M. S. Chavan, R. A. Agarwala, and M. D. Uplane, "Interference reduction in ECG using digital FIR filters based on Rectangular window", *WSEAS Transactions on Signal Processing,* vol. 4, no. 5, pp. 340-349, 2008.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, 2017.

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines", *ICML*, 2010.

[25] J. Schmidhuber, "Deep learning in neural networks: An overview", *Neural Networks*, vol. 61, pp. 85-117, 2015.

[26] G. Tolias, R. Sicre, and H. "Jégou, Particular object retrieval with integral max-pooling of CNN activations", *arXiv Prep. arXiv:1511.05879,* 2015.

[27] A. Borovykh, S. Bohte, and C. W. Oosterlee, "Conditional time series forecasting with convolutional neural networks", *arXiv Prep. arXiv:1703.04691*, 2017.

[28] R. Lu and Z. Duan, "Bidirectional GRU for sound event detection", *Detection and Classification of Acoustic Scenes and Events*, 2017, pp. 1-3.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[30] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations", *Transactions of the ASABE*, vol. 50, no. 3, pp. 885-900, 2007

AUTHOR

Shunji Cui received her B.A. degree in Chinese Laguage and Literature from Jilin Normal University in 2007, and her M.A. and Ph.D. degrees in Korean language and Literature from Kangwon University, South Korea, in 2016 and 2020, respectively. She is a full-time instructor specializing in Korean linguistics in JiLin Agriculture Science and Technology University. Her research interests are mainly in Korean Implication Theory and Grammar Theory.