electronics

MDPI

Article

Generative Adversarial Network-Based Neural Audio Caption Model for Oral Evaluation

Liu Zhang, Chao Shu, Jin Guo, Hanyi Zhang, Cheng Xie * and Qing Liu

School of Software, Yunnan University, Kunming 650504, China; zhangliu@mail.ynu.edu.cn (L.Z.); shuchao_shanty@126.com (C.S.); jinguo@mail.ynu.edu.cn (J.G.); miukkazhang@outlook.com (H.Z.); liuqing@ynu.edu.cn (Q.L.)

* Correspondence: xiecheng@ynu.edu.cn

Received: 1 February 2020; Accepted: 2 March 2020; Published: 3 March 2020

Abstract: Oral evaluation is one of the most critical processes in children's language learning. Traditionally, the Scoring Rubric is widely used in oral evaluation for providing a ranking score by assessing word accuracy, phoneme accuracy, fluency, and accent position of a tester. In recent years, by the emerging demands of the market, oral evaluation requires not only providing a single score from pronunciation but also in-depth, meaning comments based on content, context, logic, and understanding. However, the Scoring Rubric requires massive human work (oral evaluation experts) to provide such deep meaning comments. It is considered uneconomical and inefficient in the current market. Therefore, this paper proposes an automated expert comment generation approach for oral evaluation. The approach first extracts the oral features from the children's audio as well as the text features from the corresponding expert comments. Then, a Gated Recurrent Unit (GRU) is applied to encode the oral features into the model. Afterwards, a Long Short-Term Memory (LSTM) model is applied to train the mappings between oral features and text features and generate expert comments for the new coming oral audio. Finally, a Generative Adversarial Network (GAN) is combined to improve the quality of the generated comments. It generates pseudo-comments to train the discriminator to recognize the human-like comments. The proposed approach is evaluated in a real-world audio dataset (children oral audio) collected by our collaborative company. The proposed approach is also integrated into a commercial application to generate expert comments for children's oral evaluation. The experimental results and the lessons learned from real-world applications show that the proposed approach is effective for providing meaningful comments for oral evaluation.

Keywords: oral evaluation; generative adversarial network; neural audio caption; gated recurrent unit; long short-term memory

1. Introduction

Oral evaluation is a language-testing process, which includes pronunciation accuracy, fluency, integrity, logical ability, understanding ability and so on. Among them, the evaluation of logical ability and understanding ability generally requires more personalized expert comments, rather than a single score. Oral evaluation plays an important role in the process of language learning. Now there are some products and standards available in the market for oral evaluation [1–3]. However, most of these approaches are based on the Scoring Rubric [4] that only focuses on the pronunciation characteristics but ignoring the semantic characteristics, such as context, content, logic, or understanding, of the oral expression. It leads to current oral evaluation that can only provide a single ranking score [2]. It cannot provide meaningful comments for the oral evaluation. In order to improve the quality of the oral evaluation, some companies thus hire massive experts to generate comments for the evaluation manually. However, it is expensive and inefficient that not all companies could bear it. Therefore,

automated comment generation in oral evaluation becomes the emerging demand that markets are chasing. The emerging demand requires machines to imitate expert to generate expert comments for the oral expressions. It is a comprehensive problem that combines speech recognition [5], natural language generation [6], and deep learning [7]. This new interdisciplinary study is challenging. It not only requires the machine to recognize the audio features from oral speech but also requires the machine to understand the relationships between audio features and corresponding comments. There is still no mature product/method in the market that can automatically generate expert comments for oral evaluation.

With the rapid development of artificial intelligence technology [8–10], a new possible solution for automated oral evaluation emerges gradually. Our previous work had tried to apply the caption generation model to generate expert comment for the oral evaluation [11]. In this work, we optimize the previous model. In detail, a Neural Audio Caption Model (NACM) is proposed to generate expert comments from the oral audio. In NACM, based on Gate Recurrent Unit [12], an elaborate encoder-decoder structure is designed for the mapping learning between audio features and text features. Afterwards, a recurrent structure is designed by combing Generative Adversarial Network (GAN) [13] with NACM. The new model is called Generative Adversarial Network-based Neural Audio Caption Model (GNACM). Compared with the previous work, GNACM can produce more accurate and complete expert comment for the oral evaluation. Figure 1 shows the overall framework of the proposed approach.

As shown in Figure 1, the input of the model is the oral audio to be evaluated. Section 3.1 will introduce the detail of audio feature extraction. The output of the model is the generated comments according to the input oral audio. The mappings between oral audio and comments are trained in NACM. Section 3.2 will provide the structure of NACM. The generated comments are further trained through the Discriminator to improve the quality of the comments. Section 3.3 will explain the detail of GNACM. In summary, the work has the following contributions:

- We propose a model called NACM that can generate expert comment for the oral audio.
- Based on NACM, we propose an improved model called GNACM that can generate more accurate and complete expert comment for the oral audio.
- Beyond the Scoring Rubric approach, the work is the early try to generate expert comments for the oral evaluation.

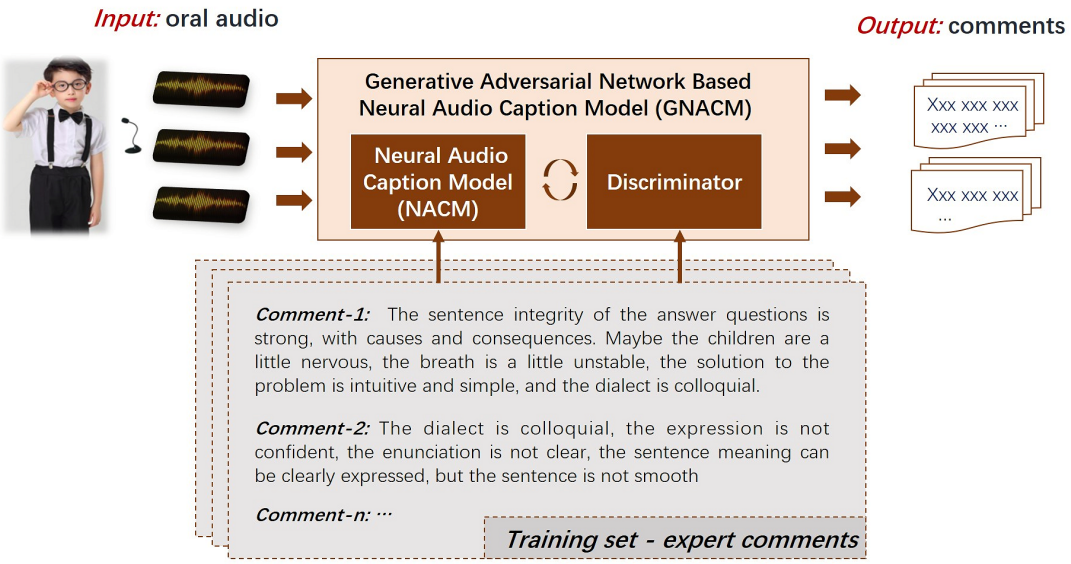

Figure 1. The overview of the proposed approach for the oral evaluation.

2. Related Work

In the related works, we first discuss the typical image caption generation model. The basic idea of image caption generation is also applied in the audio comments generation. Then, related technologies are discussed, including audio feature extraction and text generation approaches.

2.1. Caption Generation Model

Caption generation model is widely used in visual recognition for image description generating [14,15]. In the typical caption generation model, encoder-decoder architecture usually applies to generate captions for each feature vector. In the architecture, the encoder relies on deep neural networks to encode images, audio, or video to generate intermediate vectors. Then, the decoder accepts the intermediate vectors as input, perceives the intermediate vectors in turn, outputs the words one by one, and finally generates captions [16,17]. Vinyals O, et al. [18] proposed an encoder-decoder image caption generation method based on Convolutional Neural Network (CNN) [19] and Long Short-Term Memory (LSTM) [20] network. The approach extracts image features through a convolutional neural network. It then generates the target language through a long short-term memory network, whose objective function is the maximum likelihood estimation of the maximum objective description. Since the classic Neural Image Caption (NIC) model only accepts images as input at the beginning of the LSTM model. As the length of caption grows, LSTM will gradually lose the correspondence between caption and image features. Jia X, et al. [21] uses semantic information to guide the LSTM to generate descriptions at various moments. Here, descriptions indicate image caption. Various moments mean each moment of LSTM for words generation. You Q, et al. proposed a novel semantic attention model [22], which combines the mechanisms of top-down and bottom-up. The model uses responses from intermediate filters of the classification CNN to build a global visual description. In addition, the model runs a set of attribute detectors to obtain a list of visual attributes or concepts that are most likely to appear in the image. The advantage of this semantic attention model is the focus on these aspects and the use of global and local information to generate better caption. The sequence-to-sequence learning model for generating image captions has become popular, but systems for generating audio captions in the speech field are indeed rare. Therefore, this paper study the audio caption model to solve the problem of audio caption generation.

2.2. Audio Feature Extraction Model

The capture of audio feature information is closely related to the generated captions. Therefore, feature extraction is a crucial step in the caption generation task. With the development of deep learning, researchers have proposed a large number of acoustic model (AM) methods based on deep neural networks in speech recognition, which is generally divided into hybrid acoustic models and end-to-end acoustic models. Hinton G, et al. presented a pioneering work that applied deep neural networks in speech recognition tasks, and achieved a significant progress [23]. Alex Graves and Navdeep Jaitly described a system [24]. The system combined a deep bidirectional LSTM network structure and a connectionist temporal classification (CTC) [25] objective function. When the network is used in combination with the baseline system, compared with the training dataset used in the experiment (LDC corpus LDC93S6B and LDC94S13B in the Wall Street Journal corpus), the word error rate is reduced to 6.7%. Yangyang Shi, et al. proposed a method that replaces the traditional projection matrix with a higher-order projection layer [26]. Experimental results show that compared with the traditional LSTM-CTC end-to-end speech model, a higher-order LSTM-CTC model can bring a 3–10% decline in relative word error rate. It can be seen that the field of speech recognition is developing rapidly. Audio caption generation requires speech recognition-related technologies for speech processing, thereby completing the task of audio feature extraction.

2.3. Text Generation Model Based on Deep Learning

Text generation is an essential subtask of natural language processing. According to different inputs, automatic text generation can include text-to-text generation, meaning-to-text generation, data-to-text generation, and image-to-text generation. In the process of image-to-text generation at the generated text, a Recurrent Neural Network (RNN) or a recursive neural network is usually used to model the process of natural language sentence generation [27]. Socher, et al. [28] used recurrent neural networks to model sentences and used syntactic parse trees to highlight the model of actions (verbs). This method jointly optimized the image and text ends to characterize the relationship between objects and actions better. To unify the data of two different modalities under one framework, Chen and Zitnick [29] combined text information and image information into the same recurrent neural network and realized image-to-text and text-to-image bidirectional representation. To improve the quality of the generated text, Fedus W, et al. employed a Generative Adversarial Network (GAN). Compared with the maximum likelihood training model, this method can produce more realistic conditional and unconditional text samples to achieve good results [30]. The audio caption model studied in this paper is an essential branch of natural language text generation. The fidelity of the generated text determines the quality of the model.

3. The Approach

The design details of the method are described in this section. The preprocessing method for audios and comments is designed in Section 3.1. The neural audio caption model (NACM) is described in Section 3.2. A Generative Adversarial Network (GAN) is considered adding to the NACM structure. The GAN-based generative model (GNACM) is designed in Section 3.3.

3.1. Data Preprocessing

Data preprocessing is a critical step in model research. The input of this model is audio data, and the output is natural language. The specific methods for processing audio data and comment data will be described in the section.

3.1.1. Audio Feature Extraction

Audio feature extraction base on Mel Frequency Cepstral Coefficient (MFCC) features [31]. The first step is to divide the original audio into frames in time series, and then extract the MFCC features of each frame. MFCC simulates the processing characteristics of human ear to speech to a certain extent and is designed according to the knowledge of human ear auditory system. It is a general method for feature extraction in speech processing. MFCC feature extraction mainly comprises the following steps: Pre-emphasis, Framing, Windowing, FFT, Mel filter bank, computing DCT. After the audio above processing, each audio can be represented as a two-dimensional matrix ($N_{time} \times N_{mfcc}$), where N_{time} represents the number of frames in each audio, N_{mfcc} represents the feature dimension of MFCC. Next, according to the MFCC features of the first N frames and the next N frames of each frame, we calculate the deltas and delta-deltas for personalized features that preserve dynamic information effectively [32]. The approach through this step increases the connection between the previous and subsequent frames, thereby improving the representation of the feature. The length of different voice files causes the number of frames N_{time} to be different after feature extraction. To be suitable for batch calculation, it needs to uniform the number of frames in each audio. Thus, we set a hyperparameter N_{max} to indicate the maximum number of frames. Therefore, for audio with frames less than N_{max}, we pad zero to its feature matrix until the length reaches N_{max}. Finally, the MFCC feature sequence $\mathbf{A} = (\vec{a_1}, \vec{a_2}, ..., \vec{a_n})$ is obtained after the source audio processing. Where $\vec{a_x}$ represents the MFCC feature vector of each frame of an audio.

3.1.2. Text Preprocessing

Text preprocessing converts natural language into vector form. The comments in the training set are first segmented in the text processing process, and the segmented result is loaded into the thesaurus. All the words appeared in the training set together to be a thesaurus. In the thesaurus, the first word is coded 0, and the second word is coded 1, and so on. The thesaurus holds a series of correspondences between words and codes. Therefore, the sentence corresponding to each audio can be expressed as $\vec{S} = (S_1, S_2, ..., S_n)$. Where S_x represents the x^{th} word's code in a sentence. When the model performs natural language to vector conversion, the corresponding content can be retrieved by directly accessing the thesaurus.

3.2. Neural Audio Caption Model

The model is based on the neural audio caption method. The basic idea of the model is: input the audio MFCC feature sequence $\mathbf{A} = (\vec{a_1}, \vec{a_2}, ..., \vec{a_n})$ and encode it into a fixed-length $\vec{K}$, then decode the fixed-length vector and output the predicted evaluation $\mathbf{R} = (\vec{R_1}, \vec{R_2}, ..., \vec{R_n})$. The encoder of the model encodes MFCC features into learnable feature vectors. Then, these feature vectors are used to learn the correspondences with the training comments. Afterwards, the decoder of the model decodes the feature vector into comments. The complete structure of the neural audio caption method is shown in Figure 2. The model is divided into an encoder part and a decoder part.

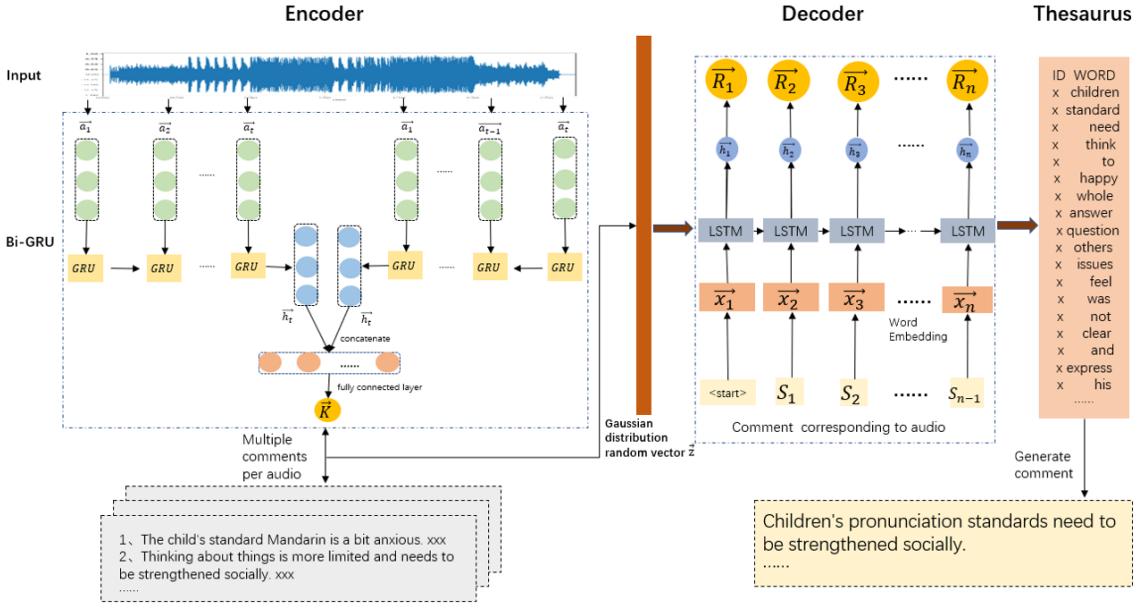

Figure 2. Neural Audio Caption Model. The encoder of the model encodes the MFCC feature sequence into a learning representation. Because one oral audio corresponds to multiple comments, a random vector with a Gaussian distribution is added after audio information encoding. The decoder model gets the input learnable audio features and corresponding comment vectors, and finally outputs comments.

3.2.1. Encoder

The feature encoder part in Figure 2 shows that the information encoding of the NACM model uses our Bi-GRU model. The Bi-GRU model is composed of GRU cells. Given the audio MFCC feature token sequence $\mathbf{A} = (\vec{a_1}, \vec{a_2}, ..., \vec{a_n})$, we use the encoder to encode the sequence information and generate a single representation $\vec{K}$. A GRU unit takes each token as input and outputs a hidden state computed by Equations (1)–(4). Then, a hidden state sequence is generated after the GRU network has computed from left to right along the input sequence. Meanwhile, a counterpart is calculated by another GRU unit computing in the reverse direction. By concatenating the last hidden state vectors and connecting the fully connected layer with activation, we finally get the encoded audio feature $\vec{K}$

with a dimension of 256. In this way, we obtain the holistic information of the sequence forward and backward. This process can be represented by Equation (5).

$$\vec{r}_t = \sigma(\mathbf{W}_r \cdot [\vec{h}_{t-1}, \vec{a}_t]) \tag{1}$$

$$\vec{z}_t = \sigma(\mathbf{W}_z \cdot [\vec{h}_{t-1}, \vec{a}_t]) \tag{2}$$

$$\vec{\tilde{h}}_t = tanh(\mathbf{W}_{\tilde{h}} \cdot [\vec{r}_t \circ \vec{h}_{t-1}, \vec{a}_t]) \tag{3}$$

$$\vec{h}_t = (1 - \vec{z}_t) \circ \vec{h}_{t-1} + \vec{z}_t \circ \vec{\tilde{h}}_t \tag{4}$$

The formula and output of GRU forward propagation are shown in Equations (1)–(4). Where $\vec{a}_t$ represents the current input vector at time t, $\vec{h}_{t-1}$ and $\vec{h}_t$ represent the hidden layer states at time $t-1$ and time t respectively. In Equations (1)–(4), $\mathbf{W}$ terms denote weight matrices and the weight matrices are defined in Equation (13), σ represents the sigmoid function and sigmoid is a non-linear function in neural network, $\vec{z}_t$ and $\vec{r}_t$ represent update gate and reset gate, $\vec{\tilde{h}}_t$ is the candidate hidden state at time t. $\circ$ indicates element-wise multiplication.

$$\vec{K} = tanh(\mathbf{W}_K \cdot \vec{h}_t + \vec{b}_K) \tag{5}$$

where $\mathbf{W}_K$ is weight matrix and $\vec{b}_K$ is bias vector. The bias vector is initialized to be 0. The weight matrices are defined in Equation (13).

3.2.2. Decoder

The decoder part designed in Figure 2 consists of LSTM cells. In practice, the same audio will be evaluated by different experts. In our training set, one audio corresponds to multiple evaluations. Thus, we concatenate a random vector $\vec{z}$ of a Gaussian distribution to the encoded audio feature $\vec{K}$ and use them as the initial hidden state to the LSTM network. Gaussian vector enables Decoder to generate multiple comments with the same input feature vector. During the training, Gaussian vector is initialized by standard normal distribution and we use the embedding of expected output from the training dataset as the input of LSTM network. The procedure can be formulated as Equations (6)–(11). Then we can get a hidden state sequence $\mathbf{h} = (\vec{h}_1, \vec{h}_2, ..., \vec{h}_n)$ after the LSTM unit has finished loop computing. Finally, we use a compositional operation to change the dimension to N_{voc}, and then getting the probability distribution for word in $\vec{R}_n$. The calculation formula is shown in Equation (12).

$$\vec{f}_t = \sigma(\mathbf{W}_f \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_f) \tag{6}$$

$$\vec{i}_t = \sigma(\mathbf{W}_i \cdot [\vec{h}_{t-1}, \vec{x}_t]] + \vec{b}_i) \tag{7}$$

$$\vec{\tilde{c}}_t = tanh(\mathbf{W}_c \cdot [\vec{h}_{t-1}, \vec{x}_t]] + \vec{b}_c) \tag{8}$$

$$\vec{c}_t = \vec{f}_t \circ \vec{c}_{t-1} + \vec{i}_t \circ \vec{\tilde{c}}_t \tag{9}$$

$$\vec{o}_t = \sigma(\mathbf{W}_o \cdot [\vec{h}_{t-1}, \vec{x}_t]] + \vec{b}_o) \tag{10}$$

$$\vec{h}_t = \vec{o}_t \circ tanh(\vec{c}_t) \tag{11}$$

$$\vec{R}_n = softmax(\mathbf{W}_R \cdot \vec{h}_t + \vec{b}_R) \tag{12}$$

The calculation formula and output of LSTM forward propagation are shown in Equations (6)–(11). Where σ is the sigmoid function, $\vec{h}_{t-1}$ is the hidden layer output at time $t-1$, $\vec{x}_t$ is the input at time t, $\vec{c}_{t-1}$ is the cell state at time $t-1$, $\vec{\tilde{c}}_t$ is the candidate value of cell state at time t, $\vec{f}_t$, $\vec{i}_t$ and $\vec{o}_t$ are the

forget gate, input gate and output gate respectively. In Equation (12), $\mathbf{W}_R$ is weight matrix and $\vec{b}_R$ is bias vector. The bias is initialized to be 0 and the weight is initialized as Equation (13).

$$W \sim U[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}] \tag{13}$$

where U denotes the uniform distribution, n is the size of the previous layer.

According to the probability distribution for the output word in $\mathbf{R} = (\vec{R}_1, \vec{R}_2, ..., \vec{R}_n)$ and the coding vector $\vec{S} = (S_1, S_2, ..., S_n)$ of the actual sample, we can calculate the loss via L Equation (14) and update the parameters of this model.

$$L = \frac{1}{n} \sum_{i=1}^{n} -log(R_i[S_i]) \tag{14}$$

where vector $\vec{S} = (S_1, S_2, ..., S_n)$ represents real reviews from experts, each word is numbered. S_i denotes the code of the word whose index is i in the sentence and $R_i[S_i]$ is the probability of the word whose index is S_i in $\vec{R}_i$.

3.3. Generative Adversarial Network-Based Neural Audio Caption Model

The neural audio caption method based on data transformation can generate understandable and appropriate evaluations. To make the generated comment closer to the expert comment, a Generative Adversarial Network (GAN) is combined [33,34].

GAN-Based Neural Audio Caption Model is composed of two neural networks, a generative neural network and a discriminative neural network. It uses the NACM as its generator, which generates comments. Meanwhile, the discriminator evaluates them for authenticity. The goal of the generator network is to generate comments that are as similar to the samples in the training set as possible. The input of the discriminator network is the real sample or the output of the generator network. Its purpose is to distinguish the output of the generator network from training data as much as possible. Thus, GAN builds a sort of feedback loop where the generator is helping to train the discriminator, and the discriminator is helping to train the generator. They both get better together. With this structure, the generated comment is closer to the real evaluation. The structure of the model is shown in Figure 3.

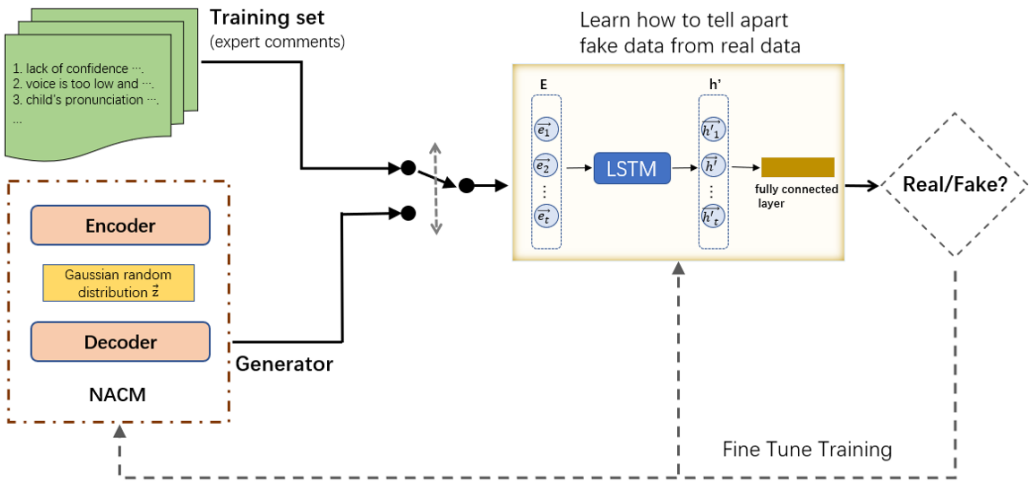

Figure 3. GAN-Based Neural Audio Caption Model. The model is divided into a generator and a discriminator. Among them, the generator follows the Neural Audio Caption Model. The discriminator uses an embedding layer, a deep LSTM layer and a fully connected layer.

3.3.1. Discriminator

The discriminator network of the model is composed of an embedding layer, a LSTM unit and a fully connected layer. The goal of training the discriminator is to maximize the probability of correctly classifying a given input as real or fake. The inputs of the discriminator are generated pseudo-expert comment and real comment. First, the fake/real comment is embed to be a matrix $\mathbf{E} = (\vec{e}_1, \vec{e}_2, .., \vec{e}_n)$ by a embedding layer, then a LSTM unit takes the embedding $\vec{e}_i$ of each word as input and outputs a hidden state $\vec{h}_t^{'}$ computed by Equations (14)–(19). Last, we feed the last hidden state vector into a fully connected layer with sigmoid activation, outputting a scalar probability that the input comment is real.

$$\vec{f}_t = \sigma(\mathbf{W}_f \cdot [\vec{h}_{t-1}, \vec{e}_t] + \vec{b}_f) \tag{15}$$

$$\vec{i}_t = \sigma(\mathbf{W}_i \cdot [\vec{h}_{t-1}, \vec{e}_t] + \vec{b}_i) \tag{16}$$

$$\vec{\tilde{c}}_t = tanh(\mathbf{W}_c \cdot [\vec{h}_{t-1}, \vec{e}_t] + \vec{b}_c) \tag{17}$$

$$\vec{c}_t = \vec{f}_t \circ \vec{c}_{t-1} + \vec{i}_t \circ \vec{\tilde{c}}_t \tag{18}$$

$$\vec{o}_t = \sigma(\mathbf{W}_o \cdot [\vec{h}_{t-1}, \vec{e}_t] + \vec{b}_o) \tag{19}$$

$$\vec{h}_t^{'} = \vec{o}_t \circ tanh(\vec{c}_t) \tag{20}$$

The loss function for the discriminator is used to judge the ability of the discriminator. The loss function is defined as:

$$L^D_{real/fake} = -(E(logD_{real/fake}(\vec{T})) + E(1 - log(D_{real/fake}(G(\mathbf{A}, \vec{z}))))) \tag{21}$$

where $G(\mathbf{A}, \vec{z})$ represents the comments generated by the generator. $D_{real/fake}(x)$ is the discriminator which outputs the scalar probability that x came from the truth sample rather than the generator, $\vec{T}$ represents the real sample.

3.3.2. Generator

The generator of the model inherits the structure from the Neural Audio Caption Model. The generator is divided into an encoder and a decoder. The encoder encodes the MFCC features into learnable feature vectors. Then, a Gaussian-distributed random vector $\vec{z}$ is connected to the encoded features. The decoder takes the encoded features as input and then generates comments.

The similarity between the real samples and the comments generated by the generator is the primary standard for measuring the generator compliance. Therefore, the comments generated by the generator must be similar to the real data to deceive the discriminator. The generator loss function is designed to evaluate this ability of the generator. The goal of training the generator is to maximize $D(G(\mathbf{A}, \vec{z}))$. The loss of the generator is defined as:

$$L^G_{real/fake} = E(-logD_{real/fake}(G(\mathbf{A}, \vec{z}))) \tag{22}$$

where $G(\mathbf{A}, \vec{z})$ represents the comments generated by the generator, and $D_{real/fake}(x)$ represents scalar probability that x came from the truth sample rather than the generator.

The results of supervision using only discriminators are uncertain. Therefore, we turn the learning problem into an optimization problem. Define a loss function to measure the distribution difference between the result and the actual sample to minimize the loss. The loss function is as follows:

$$L^G_{dis} = \propto \times (G(\mathbf{A}, \vec{z}) - \vec{T})^2 \tag{23}$$

where α is the balance factor and $\vec{T}$ represents the real sample.

Finally, we can calculate the loss via Equation (23) and update the parameters of generator network.

$$L^G = L^G_{real/fake} + L^G_{dis} \tag{24}$$

4. Case Study

This section describes the application of the Neural Audio Caption Model (NACM) and the Generative Adversarial Network-Based Neural Audio Caption Model (GNACM) to actual cases in detail. In Section 4.1, we describe the application scenario of the oral evaluation system. We explain the dataset required for model training and performance testing of the system in Sections 4.2 and 4.3. We introduce the application of the system in the entity enterprise, which fully shows the oral evaluation system's practical application and significance value in Section 4.4. The advantages and disadvantages of the model are thoroughly analyzed according to the experimental results in Section 4.5.

4.1. Scenario

At present, the education and training industry has maintained a high growth rate of nearly one trillion Chinese Yuan. According to statistics, the scale of China's training market in 2015 was about 882.1 billion Chinese Yuan. Language training institutions account for a large proportion in the training market, around 17.3%. The size of the education and training market in 2016 has exceeded one trillion Chinese Yuan, maintaining a growth rate of 13.1% per year. After investigation, the types of language training institutions are increasing in the market. The language training institutions mainly includes four categories: comprehensive curriculum education institutions, spoken language institutions, study abroad institutions, and minor language institutions. According to the survey of the language education market, we find that language training institution will be a long-standing industry in the future market.

The development prospects of artificial intelligence language training institutions are excellent [35–37]. Online language education directly hits the pain points of the industry. It solves many problems, such as time-consuming, labor-intensive, expensive, and one-to-one teaching. Although online language training systems overcome the disadvantages of manual education, it still has a lot of obstacles. For example, most current online language training systems can only score speech according to the principle of the Scoring Rubric or give a single evaluation according to the unified Scoring Rubric standards. Figure 4 shows the test process of the traditional language evaluation system.

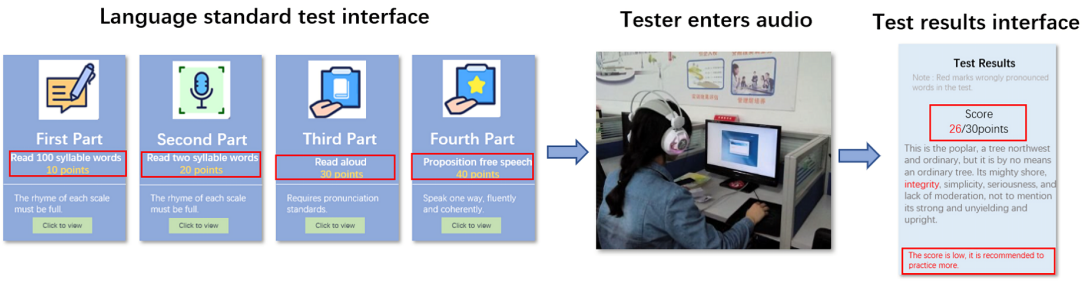

Figure 4. Test process of the traditional language evaluation system. The language standard test interface displays the four parts of the language test. The test contents marked by the red rectangle are made based on the gauge. The Test results interface displays the audio test results of "Read aloud".

With the continuous advancement of science and technology, artificial intelligence products have entered human life [38–40]. To solve the problems of the online education systems, we apply artificial intelligence technology to the online language evaluation systems. The development of an artificial intelligence online language evaluation system will solve the problems in the language education market. The Neural Audio Caption Model (NACM) and Generative Adversarial Network-Based

Neural Audio Caption Model (GNACM) in the paper can generate meaningful expert comments for the oral audio evaluation. The system based on NACM and the system based on GNACM meet the needs of the language training market. NACM and GNACM are applied in the oral evaluation system, which will realize the automation and intelligence of language evaluation.

4.2. Dataset

The first step in the development of the intelligent oral evaluation system is collecting relevant datasets, including audios and expert comments corresponding to audios. The data set affects the accuracy and professionalism of the comments generated by the final model. Therefore, the training set and testing set should include the objective phonemes of different people, different ages, and different environments. The data set was provided by children's language education institutions for this paper and is not publicly available. With the consent of the child's guardian, the oral audio of children 5–6 years old was manually collected. During the collection process, the children watch a cartoon video and say their thoughts. After the audio collection is complete, the relevant children's language experts are invited to comment accordingly. This dataset is called the children oral audio dataset.

4.3. Performance Testing

The NACM and GNACM proposed in this paper will be an essential model of the oral evaluation system. The oral evaluation system is used in enterprise software. Application software has particular evaluation indicators. Therefore, NACM and GNACM are tested using applied evaluation indicators. The evaluation indicators and evaluation results of the model are introduced in detail below in this paper.

4.3.1. Evaluation Metrics

The evaluation metrics of models include three aspects in this paper. The model evaluation metrics are the quality score of model generation comments, the average response time, and the scalability. The quality score of the model-generated reports will be manually reviewed and scored by language education experts. For the performance of the model, we assessment the model from the response time and scalability.

4.3.2. Evaluation Results

The experiments study two models NACM and GNACM. The dataset used is children oral audio dataset. The experimental process is shown in Figure 5. Children 5–6 years old watch cartoon video and answer questions. The child's oral audio is input into two models, and the models automatically generate comments. During the experiment, we assessment the model using the model evaluation metrics proposed in Section 4.3.1.

Relevant language experts evaluate the quality score of the comments generated by the models, as shown in Figure 6. The Score represents the score of the generated comment graded by human experts. The Expert number of the chart represents the expert number. The figure shows the comparison between the quality score of evaluation generated by the baseline model NACM and the quality score of evaluation generated by the GNACM model. The two systems output audio comments by inputting different audios. Multiple experts score the corresponding comments of the output multiple audios. The average score obtained is the system evaluation quality score corresponding to each expert. The process of collecting expert comments is that we first get the test results, then send the audio and generated comments to the anonymous experts, and then the experts send their score results to us, and finally sort out each expert score on the quality of the generated evaluation. Among them, we asked experts to rate the audio and comments given on a 100-point scale. As Figure 6 shown, we can find that the quality score of the comment generated by GNACM is better than NACM.

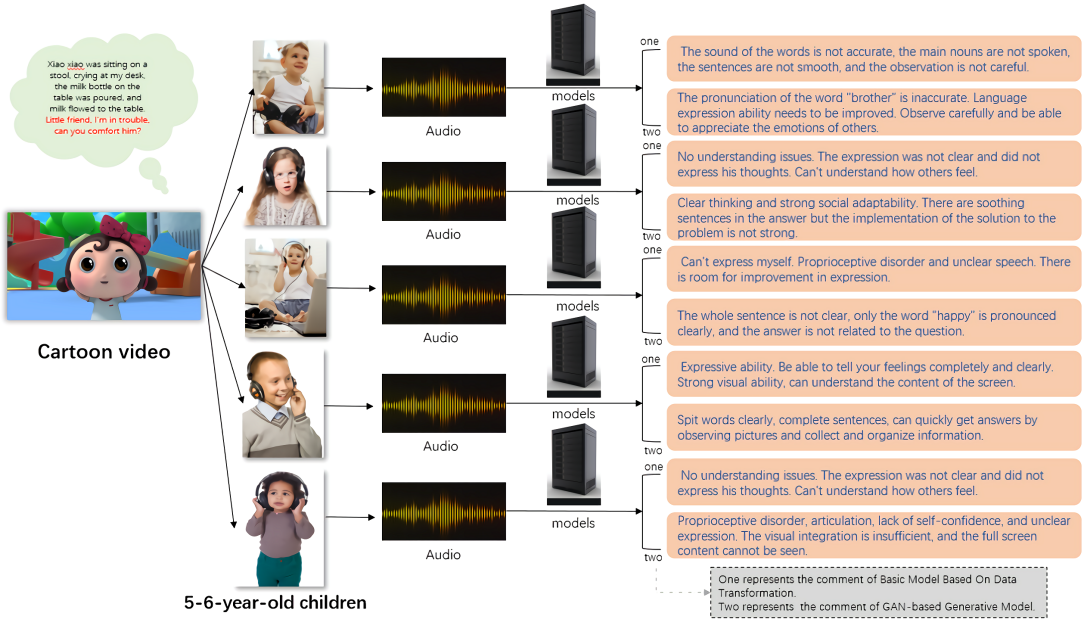

Figure 5. Experimental of an oral evaluation test for children. The participants are asked to watch a cartoon. In the cartoon, there are well designed oral questions. The participants answer the corresponding questions orally. Then, the system will automatically generate the expert comments for the participants.

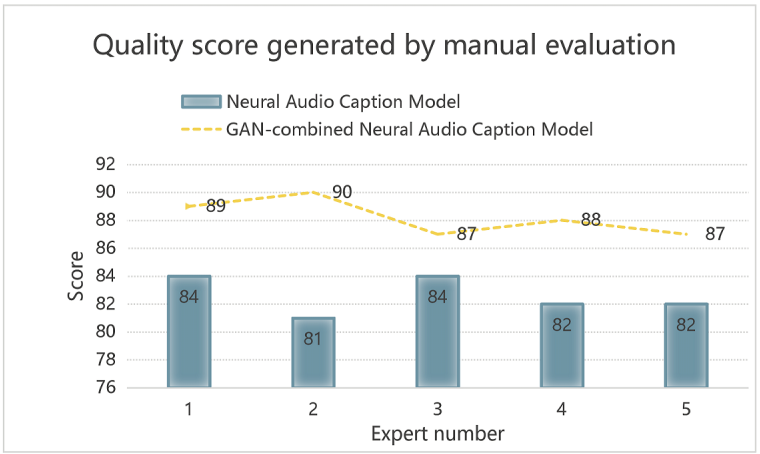

Figure 6. Quality score of evaluations generated by manual assessment models.

The average response time of the model intuitively reflects the performance in the application environment, as shown in Table 1. We tested all audio samples for NACM and GNACM and then calculated the average response time. The NACM model and the GNACM model run and tested on Windows10. The software environment is anaconda3, PyTorch V1.3.1, cuDNN V7.0, and CUDA V10.1. According to the data in Table 1, the model can meet the requirements for practical production.

Table 1. Average response time (ms) of the proposed models.

NACM	GNACM
129.75	125.50

The scalability of the model describes the system's ability to respond to load growth, as shown in Figure 7. The Response time of the chart represents the total time consumption. The Number of samples of the chart represents the number of test samples. The scalability of the model describes the system's ability to respond to load growth, as shown in Figure 7. The Response time of the chart

represents the total time consumption. The Number of samples of the chart represents the number of test samples. For scalability evaluation, samples are called increasingly. The evaluation is used to test whether the models suitable for large scale deployment. As the number of test samples increases, the total response time is rising linearly. It means, in the deployment, response time for each audio sample can always be stable by applying acceptable computing resources. The scalability of the model can fully meet the needs of the online language education system.

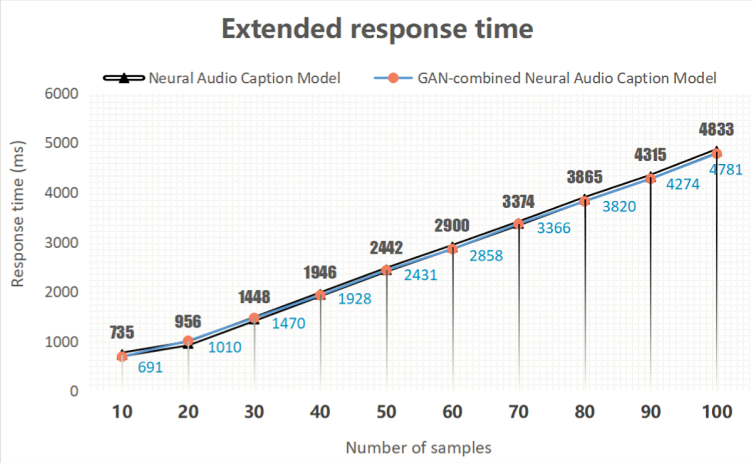

Figure 7. Scalability evaluation for NACM and GNACM.

4.4. Application

The artificial intelligence oral evaluation system is devoted to creating an online testing software that enhances human language proficiency. The software is convenient, fast, comprehensive, and inexpensive. With the development of computer technology, online education has become routine. Therefore, the intelligent oral evaluation system we develop has met the needs of human language education and reduced the cost of human and material resources. In this section, we will apply NACM and GNACM to the oral evaluation system on the market. The NACM is applied to the baseline system of the intelligence oral evaluation for children system. The example of GNACM is compared with NACM in application software.

4.4.1. Baseline System Based on NACM

Language education is an essential part of children's learning and life development. The design of the intelligence oral evaluation for children system will provide each child with a language-centric education platform with fluent language skills. The research and development of this artificial intelligence product is an important step to integrate intelligent technology in the field of children's education. The development of the intelligence oral evaluation system provides a solution to personalized services for children's home education and early childhood school education. Each child can use the software according to their own needs. Under the guidance of parents and teachers, they can test the development level of language skills. The system can also provide suggestions for children's language skills development. At the same time, parents and teachers can monitor the progress of children's language skills in real-time.

This paper proposes an NACM that automatically generates expert comments, which is first applied to the intelligence oral evaluation for children system. We apply the trained NACM to the "Speaking Practice" section of the intelligent oral evaluation for children system. The basic operating environment of the model is as follows. The software environment is anaconda3 (based on Python 3.7), PyTorch V1.3.1, cuDNN V7.0, and CUDA V10.1. The hardware environment is Intel (R) Core (TM) i7-9700 CPU @ 3.00 GHz (8 CPUs) 3.0 GHz, NVIDIA GeForce RTX 2070 SUPER 8 GB, 16 GB RAM. The system imitates a language expert to assess the user language ability and give personalized

comments. Through the experimental research, the NACM-based generated comment is incomplete. In the next section, we introduce the use of the GNACM model to solve this problem.

4.4.2. GNACM for Children Oral Evaluation

The intelligent oral evaluation system is currently equipped with GNACM on the market. The server hardware environment is E5-2680v2 CPU @ 20 cores and 40 threads, the main frequency is 2.80 GHz, 128 G memory, four GeForce RTX ™ 2080 Ti GPUs. To simplify the calling method, we deploy a web service on the server. Web service can quickly provide data transmission services to third parties due to the advantages of cross-platform and cross-language. The enterprise establishes a web service client and initiates a connection with our server, which transmits audio information. Then the web service program calls our deep learning module and returns the evaluation result to the client. We provide a web service interface to outside companies, and they can call the corresponding model and apply it to the corresponding function module of the system. Figure 8 shows a language education company that accesses our model by invoking the web APIs of the model. The enterprise uses the API we provided to call the model. The user can use the model with a mobile phone or PC and other devices, input oral audio, and transfer the data to the layers. The server and the model give comments, and the comments data is transmitted layer by layer and fed back to the user. Finally, the display interface displays comments. GNACM gives the audio comment for children oral evaluation is more accurate, similar to expert comment.

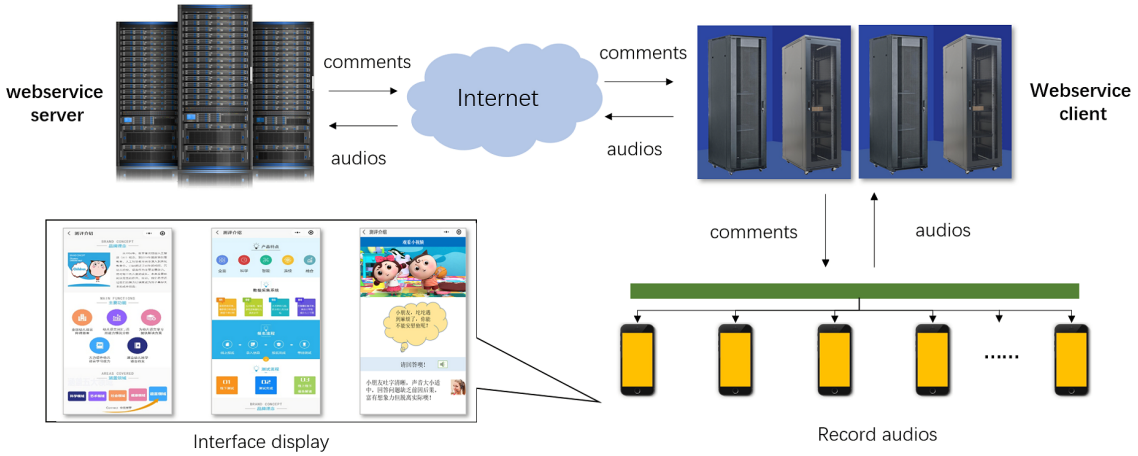

Figure 8. The way the third-party users to access the proposed model for an oral evaluation.

4.5. Lesson Learned

The study finds that people pay attention to the development of language education. There is a great demand for intelligent systems for language evaluation in the market. Therefore, this paper proposes two models of NACM and GNACM and applies them to the intelligent oral evaluation system. Through case analysis, we can find that the intelligent oral evaluation system has many advantages. (1) The development of the intelligent oral evaluation system will solve the problem of time-consuming manual participation of oral evaluation in the market. Compared with traditional online systems, the comments generated by the model are personalized and comprehensive. (2) The extraction and encoder of audio features are difficult. This paper uses the MFCC feature extraction to extract audio features and uses the Bi-GRU model to encoder audio features to solve these two problems. Moreover, in the audio preprocessing process, we use the method of setting the maximum frame and zero-padding to solve the problem of different lengths of audio feature sequences. (3) After NACM is completed, the GAN method is added to the NACM, which significantly improves the accuracy of the evaluation. The intelligent oral evaluation system is more applicable in the business model. (4) The performance test results of the model indicate that the proposed approach can generate

meaningful expert comments for the oral audio evaluation. It is suitable for language learning and testing market. By adjusting the parameters and training set, the approach can be also applied in industrial applications. (5) Oral audio evaluation is widely used in various domains, including education, security, finance, and industry. The proposed approach can be further applied to worker status evaluation through analyzing oral questions and answers in many industrial environments.

The development of the intelligent oral evaluation system solves the pain points of intelligent evaluation of language education. However, deficiencies are also found during the research of the model. (1) Audio features are extracted through MFCC feature extraction. In the encoder, the accuracy of the audio feature encoding needs to be further improved. Encoder of the model can still be optimized in the future. At the time of comment generation, as the audio feature layer deepens, the audio features gradually weaken. It may lead the generated comment, to some extent, deviates from the audio. (2) The generated comments have a high similarity with the trained expert comments, resulting in inflexible generated comments. We want the model to generate a more personalized comment for different participants. GAN-based model is hard to generate varieties beyond the training set. Based on the deficiencies of the model, we will continue to study the model to improve the accuracy of the generated evaluation.

5. Conclusions

This paper proposes a generative model for the oral evaluation. Compared with the traditional method, the proposed model is more effective and efficient in oral evaluation. It can generate meaningful comments according to the oral audio without manual works. The proposed approach consists of two parts. The first part is Neural Audio Caption Model (NACM). It applies a Gated Recurrent Unit (GRU) to encode the audio features into the neural network. It also applies a Long Short-Term Memory (LSTM) model to discover the mappings between audio features and text features. The second part of the approach is the Generative Adversarial Network-Based Neural Audio Caption Model (GNACM). It uses the output of NACM as its input to improve the quality of generated comments. The proposed approach is evaluated in a real-world dataset. It also is applied in a commercial application. The evaluation results and the lessons learned from the application show that the proposed approach is effective and efficient in oral evaluation. In the future, we plan to apply the knowledge graph to process the content and context of oral audio. Therefore, the model will consider not only the audio analysis but also the semantic analysis for the oral evaluation.

Author Contributions: Conceptualization, C.X. and L.Z.; methodology, C.S., H.Z.; software, C.S., H.Z.; validation, C.X., J.G., and Q.L.; formal analysis, C.X.; data curation, L.Z.; writing—original draft preparation, L.Z.; writing—review and editing, L.Z. and J.G.; visualization, L.Z.; supervision, C.X. and Q.L.; project administration, C.X. and Q.L.; funding acquisition, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This word was funded by Scientific Research Fund of Yunnan Provincial Department of Education.

Acknowledgments: The authors would like to thank to all the reviewers who helped us in the review process of our work. Moreover, special thanks to International Conference on e-Business Engineering (ICEBE 2019) recommend the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NACM	Neural Audio Caption Model
GNACM	Generative Adversarial Network-Based Neural Audio Caption Model
MDPI	Multidisciplinary Digital Publishing Institute
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory network
NIC	Neural Image Caption
NAC	Neural Audio Caption

CNN	Convolutional Neural Network
AM	Acoustic Model
CTC	Connectionist Temporal Classification
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network
MFCC	Mel Frequency Cepstral Coefficient
BP	Error Back Propagation
APP	Application
AI	Artificial Intelligence

References

1. Voice Evaluation. Available online: http://global.xfyun.cn/products/ise (accessed on 22 September 2019).
2. Smart Oral Evaluation-English. Available online: https://cloud.tencent.com/product/soe-e (accessed on 22 September 2019).
3. Computer Assisted Pronunciation Training. Available online: https://ai.youdao.com/product-assess.s (accessed on 22 September 2019).
4. Moskal, B.M.; Leydens, J.A. Scoring rubric development: Validity and reliability. *Pract. Assess. Res. Eval.* **2000**, *7*, 10.
5. Toshniwal, S.; Sainath, T.N.; Weiss, R.J.; Li, B.; Rao, K. Multilingual Speech Recognition with a Single End-to-End Model. *arXiv* **2018**, arXiv:1711.01694.
6. Gatt, A.; Krahmer, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Vestn. Oftalmol.* **2018**, *45*, 75–170. [CrossRef]
7. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
8. Kennedy, J.; Séverin, L.; Montassier, C.; Lavalade, P.; Irfan, B.; Papadopoulos, F. Child speech recognition in human-robot interaction: Evaluations and recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017; pp. 82–90.
9. Wang, J.; Kothalkar, P.V.; Kim, M.; Bandini, A.; Green, J.R. Automatic prediction of intelligible speaking rate for individuals with als from speech acoustic and articulatory samples. *Int. J. Speech Lang. Pathol.* **2018**, *20*, 669–679. [CrossRef] [PubMed]
10. Ma, Z.; Yu, H.; Chen, W.; Guo, J. Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features. *IEEE Trans. Veh. Technol.* **2018**, *68*, 121–128. [CrossRef]
11. Liu, Z.; Hanyi, Z.; Jin, G.; Detao, J.; Qing, L.; Cheng, X. Speech Evaluation based on Deep Learning Audio Caption. In Proceedings of the International Conference on e-Business Engineering, Bali, Indonesia, 21–23 December 2019; pp. 51–66.
12. Cho, K.; van Merriënboer, B.; Bahdanau, D.; Bengio, Y. *On the Properties of Neural Machine Translation: Encoder–Decoder Approaches*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 103–111.
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Bing, X.; Warde-Farley, D.; Sherjil, O.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
14. Deshpande, A.; Aneja, J.; Wang, L.; Schwing, A.G.; Forsyth, D. Fast, diverse and accurate image captioning guided by part-of-speech. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
15. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
16. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2008**, *2*, 3104–3112.
17. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
18. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.

19. Phil, K. Convolutional Neural Network. In *MATLAB Deep Learning*; Apress: Berkeley, CA, USA, 2017.
20. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
21. Jia, X.; Gavves, E.; Fernando, B.; Tuytelaars, T. Guiding the Long-Short Term Memory Model for Image Caption Generation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
22. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
23. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Proc. Mag.* **2012**, *29*, 82–97. [CrossRef]
24. Graves, A.; Navdeep, J. Towards end-to-end speech recognition with recurrent neural networks. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1764–1772.
25. Graves, A.; Santiago, F.; Gomez, F. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006.
26. Shi, Y.; Hwang, M.Y.; Lei, X. End-to-end speech recognition using a high rank lstm-ctc based model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 7080–7084.
27. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *39*, 664–676. [CrossRef] [PubMed]
28. Socher, R.; Karpathy, A.; Le, Q.V.; Manning, C.D.; Ng, A.Y. Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 207–218. [CrossRef]
29. Chen, X.; Zitnick, C.L. Learning a recurrent visual representation for image caption generation. *arXiv* **2014**, arXiv:1411.5654.
30. Fedus, W.; Goodfellow, I.; Dai, A.M. Maskgan: Better text generation via filling in the_. *arXiv* **2018**, arXiv:1801.07736.
31. Upadhya, S.; Cheeran, A.N.; Nirmal, J.H. Discriminating Parkinson diseased and healthy people using modified MFCC filter bank approach. *Int. J. Speech Technol.* **2019**, *224*, 1021–1029. [CrossRef]
32. Mingyi, C.; Xuanji, H.; Jing, Y.; Han, Z. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444.
33. Liu, F.; Zheng, J.; Zheng, L.; Chen, C. Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. *Neurocomputiong* **2019**, *371*, 39–54. [CrossRef]
34. Yan, S.; Xie, Y.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image captioning based on a hierarchical attention mechanism and policy gradient optimization. *arXiv* **2018**, arXiv:1811.05253.
35. Dalim, C.; Samihah, C.; Sunar, M.S.; Dey, A.; Billinghurst, M. Using augmented reality with speech input for non-native children's language learning. *Int. J. Hum. Comput. Stud.* **2019**, *134*, 44–64. [CrossRef]
36. Schepens, J.; van Hout, R.; Jaeger, T. Florian. Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition* **2019**, *194*, 104056. [CrossRef]
37. Cho, K.; van Merrienboer, B.; Bahadanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder aaaroaches. *arXiv* **2014**, arXiv:1409.1259.
38. Chen, C.; Gao, L.; Xie, X.W.; Wang, Z. Enjoy the most beautiful scene now: A memetic algorithm to solve two-fold time-dependent arc orienteering problem. *Front. Comput. Sci.* **2020**, *14*, 364–377. [CrossRef]
39. Manikandan, R.; Patan, R.; Gandomi, A.H.; Sivanesan, P.; Kalyanaraman, H. Hash polynomial two factor decision tree using IoT for smart health care scheduling. *Expert Syst. Appl.* **2019**, *141*, 112924. [CrossRef]
40. Pan, J.S.; Xi, T.; Jiang, R. Emotional Effects of Smart Aromatherapeutic Home Devices. In Proceedings of the International Conference on Applied Human Factors and Ergonomics, Washington, DC, USA, 24–28 July 2019; pp. 498–503.