





Article

Automatic Fluency Assessment Method for Spontaneous Speech without Reference Text

Jiajun Liu ^{1,2} , Aishan Wumaier ^{2,3,*} , Cong Fan ^{2,3}  and Shen Guo ^{2,3} 

- ¹ College of Software, Xinjiang University, Urumqi 830046, China; liujiajun@stu.xju.edu.cn
² Key Laboratory of Multilingual Information Technology in Xinjiang Uyghur Autonomous Region, Urumqi 830046, China
³ College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; fanc@stu.xju.edu.cn (C.F.); guoshen@stu.xju.edu.cn (S.G.)
* Correspondence: hasan1479@xju.edu.cn

Abstract: The automatic fluency assessment of spontaneous speech without reference text is a challenging task that heavily depends on the accuracy of automatic speech recognition (ASR). Considering this scenario, it is necessary to explore an assessment method that combines ASR. This is mainly due to the fact that in addition to acoustic features being essential for assessment, the text features output by ASR may also contain potentially fluency information. However, most existing studies on automatic fluency assessment of spontaneous speech are based solely on audio features, without utilizing textual information, which may lead to a limited understanding of fluency features. To address this, we propose a multimodal automatic speech fluency assessment method that combines ASR output. Specifically, we first explore the relevance of the fluency assessment task to the ASR task and fine-tune the Wav2Vec2.0 model using multi-task learning to jointly optimize the ASR task and fluency assessment task, resulting in both the fluency assessment results and the ASR output. Then, the text features and audio features obtained from the fine-tuned model are fed into the multimodal fluency assessment model, using attention mechanisms to obtain more reliable assessment results. Finally, experiments on the PSCPSF and Speechocean762 dataset suggest that our proposed method performs well in different assessment scenarios.



Citation: Liu, J.; Wumaier, A.; Fan, C.; Guo, S. Automatic Fluency Assessment Method for Spontaneous Speech without Reference Text. *Electronics* **2023**, *12*, 1775. <https://doi.org/10.3390/electronics12081775>

Academic Editor: Chiman Kwan

Received: 2 March 2023

Revised: 2 April 2023

Accepted: 7 April 2023

Published: 9 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: automatic fluency assessment; automatic speech recognition; spontaneous speech; multi-task learning; multimodal

1. Introduction

Speech fluency is an essential criterion of speakers' verbal ability and is one of the scoring requirements for many language proficiency tests. For example, the TOEFL iBT speaking test in English and the Mandarin Proficiency Test (Putonghua Shuiping Ceshi, PSC) in Chinese assess speaking skills in areas such as pronunciation, grammar, vocabulary, fluency and coherence. Both the language proficiency tests require the speakers to speak spontaneously within a specified time frame without reference text. Factors affecting speech fluency include speech rate, time to fill speech versus silence, hesitation, and repair phenomena [1]. The assessment of speech fluency is more subjective than the common criteria used in tests, such as pronunciation, vocabulary, and grammar. As spontaneous oral test requires a multi-dimensional examination of speakers, manual scoring is still used. However, manual scoring is susceptible to cost, time, and space, resulting in unreliable manual scoring results. Therefore, automatic fluency assessment models are needed to address the above mentioned issues. Unlike assessment models with reference texts, the two main challenges of spontaneous speech fluency assessment are as follows: (a) Pattern matching or forced alignment methods cannot be used as there is no reference text. (b) Spontaneous speech fluency assessment tasks need to be run on transcriptions of ASR, which can contain errors. Therefore, to solve the task of assessing speech fluency

without reference text, an automatic fluency assessment method precisely for spontaneous speech needs to be designed.

Automatic speech fluency assessment methods are generally studied for specific tasks and datasets [2]. Different modeling methods exist for specific tasks. Common methods for speech fluency assessment can be classified into two categories: traditional machine learning and deep learning approaches. Traditional research in automatic speech assessment methods has focused on speech signal processing to extract acoustic features and then using machine learning models to map the features to scores. The deep learning model eliminates the need for tedious feature extraction and obtains fluency features from the raw audio, resulting in improved performance of the speech assessment model.

However, there are some limitations to the existing studies, which tend to assess only audio-based aspects and do not combine textual information. However, we found that some disfluent textual features, such as “um”, “un” and repetition, are also frequently present in spontaneous speech. For assessment scenarios with no textual reference, textual features can be obtained using ASR and combined with acoustic features to assess the oral expression of the speakers. This paper proposes a multimodal automatic speech fluency assessment method combining ASR output. Specifically, multi-task learning is first applied to fine-tune the Wav2Vec2.0 model so that the fluency assessment task is jointly optimized with the ASR task. Then we combined the last hidden layer’s output of the fine-tuned Wav2Vec2.0 model with the ASR output and fed them into a fluency assessment model based on multimodality. After fully extracting the unimodal internal features, feature fusion and enhancement are successively achieved using cross-modal attention and self-attention to obtain the final fluency assessment results.

The main contributions of this paper can be summarized as follows:

- A 9.26-h PSC propositional speaking fluency assessment dataset PSCPSF is constructed to address the scarcity of spontaneous speech data resources in Chinese.
- To validate the relevance of the speech assessment task to the ASR task, the multi-task learning framework is used to fine-tune the Wav2Vec2.0 model on the specific dataset so that the model is jointly optimized by using the speech fluency assessment task as the main task and the ASR task as the auxiliary task, thus improving the accuracy of the model in speech fluency assessment.
- A multimodal automatic fluency assessment method combining ASR output is proposed, which combines the final layer of the hidden output of the fine-tuned Wav2Vec2.0 model with the decoded text output, followed by a cross-modal attention mechanism to fuse acoustic and textual features, resulting in more useful fluency features for achieving adequate speech fluency assessment.
- Experiments on the PSCPSF and Speechocean762 datasets for spontaneous speech and read-aloud scenarios demonstrate that our method performs well in the different assessment scenarios.

The structure of the paper is as follows. Section 2 presents work related to the method, and Section 3 describes the proposed model, including the fine-tuning Wav2Vec2.0 model, and the multimodal model. Section 4 then describes the data set and the experimental setup. Section 5 is devoted to the analysis of the experimental results. Section 6 concludes the paper.

2. Related Work

2.1. Automatic Fluency Assessment

Automatic speech fluency assessment models typically involve ASR to generate time-aligned word sequences for the input speech, as well as fluency feature extraction and scoring models. Fluency features generally include long silence, words per second, phone duration, etc. [2,3]. Scoring models are typically processed for classification or regression tasks. Researchers in [4–7] used support vector machines to classify speech into different fluency levels. In [8], the authors implemented fluency assessment through a Gaussian model. Authors in [9–12] used manual features as input and multiple linear or ordered

regressions to predict fluency scores. With the widespread use of deep learning, deep neural network (DNN)-based approaches were used to improve the performance of spoken English fluency scores by using DNN-based acoustic models and confidence features [13,14]. Chung et al. [15] proposed a convolutional neural network (CNN)-based approach to learn fluency features directly from a corpus of raw data. In addition, the authors in [16,17] used long and short-term memory (LSTM) and bidirectional long and short-term memory (BLSTM) to better capture the dynamic changes in phone-level fluency features.

For most studies on speech, fluency assessment is conducted in read-aloud scenarios. The scenarios of asking speakers to read a given text are relatively easy to assess. In the read-aloud scenario, speech interruptions, speech rate, articulatory quality features, and goodness of pronunciation (GOP) are frequently used in fluency assessment tasks [7,9,17,18]. In contrast to the assessment of speech fluency in the read-aloud scenario, spontaneous speech without reference text is more complex, as the speakers are asked to express themselves freely according to the topic. Unlike the read-aloud scenarios, spontaneous speech requires ASR to obtain transcriptions. Afterward, the transcriptions are forcibly aligned with the audio to get phone, word, and utterance boundaries, after which predefined linguistic rules are applied to extract prosodic and lexical features, and finally regression or classification models are used to obtain assessment results [2,3,5,6,19].

Currently, automatic fluency assessment tasks achieve satisfactory results in both read-aloud scenarios and spontaneous speech scenarios. However, there is still a need for more generalized speech fluency assessment methods. Traditional assessment methods rely heavily on the validity of hand-crafted fluency features that are only applicable to specific datasets and may not apply to other datasets. Furthermore, we find that text features transcribed by ASR are rarely used for fluency assessment. Few studies have combined fluency with text features for comprehensive assessment. Words containing fluency features are present in both English and Chinese, and these text-related fluency features can also be used for fluency assessment. In this work, we drew inspiration from ASR models, speech emotion classification models, music classification models, and multi-layer frameworks [20–27] to propose a multimodal model that focuses on learning useful features from raw speech data for fluency assessment.

2.2. Multi-Task Learning

Multi-task learning (MTL) refers to combining multiple single tasks with relevance to learn from each other, allowing these tasks to share information during the learning process and using the valuable information implicit in various related tasks to help improve performance and generalization on each task [28]. MTL is widely used in deep neural models in the speech domain. In end-to-end ASR tasks, researchers in [29] used a joint connectionist temporal classification (CTC) [30] attention model within the multi-task learning framework. For the speech emotion recognition task, the authors in [31–33] used a multi-task learning approach to improve emotion recognition accuracy through shared representation. This paper explores the correlation between speech fluency assessment and ASR tasks and uses MTL when fine-tuning the Wav2Vec2.0 model. Based on the idea of MTL, the speech fluency assessment task is used as the main task and the ASR task as a secondary task, making the speech fluency assessment task learn further useful information from the ASR task to compensate for the shortcomings of the single-task model that does not learn enough useful information.

2.3. Wav2Vec2.0

Wav2Vec2.0 [34] is a framework for the self-supervised learning of speech representations [35], which has been extensively utilized for multi-task speech emotion recognition [24] and various speech activated tasks, such as speaker verification, keyword spotting, voice activity detection, etc. [36]. This paper uses MTL to fine-tune the Wav2Vec2.0 model for different downstream tasks in the speech fluency assessment task.

3. Proposed Method

In the fluency assessment scenario without reference text, the lack of access to the exact content of the speaker's expression results in low accuracy of the ASR. We integrate the ASR and fluency assessment models to address this problem. This section proposes a multimodal automatic speech fluency assessment method combining ASR outputs while using the combined features from audio and text modalities to improve the stability of the speech fluency assessment model.

The structure of the proposed model is illustrated in Figure 1, which includes a fine-tuned Wav2Vec2.0 model and a multimodal fusion network. The implementation of the whole model can be done step by step. We first fine-tune the Wav2Vec2.0 model using multi-task learning so that the ASR task is jointly optimized with the fluency assessment task to obtain the fluency assessment results and the ASR output. Following this, we extract the hidden states from the fine-tuned Wav2Vec2.0 model as audio feature input and the text decoded by CTC as text feature input. The audio and text features are then processed through CNN to capture contextual features, which are then fused by cross-modal attention. Afterward, the fused text-audio and audio-text features are enhanced by the self-attentive mechanism, respectively. Finally, after connecting these two output vectors from the self-attentive mechanism, we obtain the predicted fluency levels through the fully connected layer. We will provide a detailed explanation of the two components of our model.

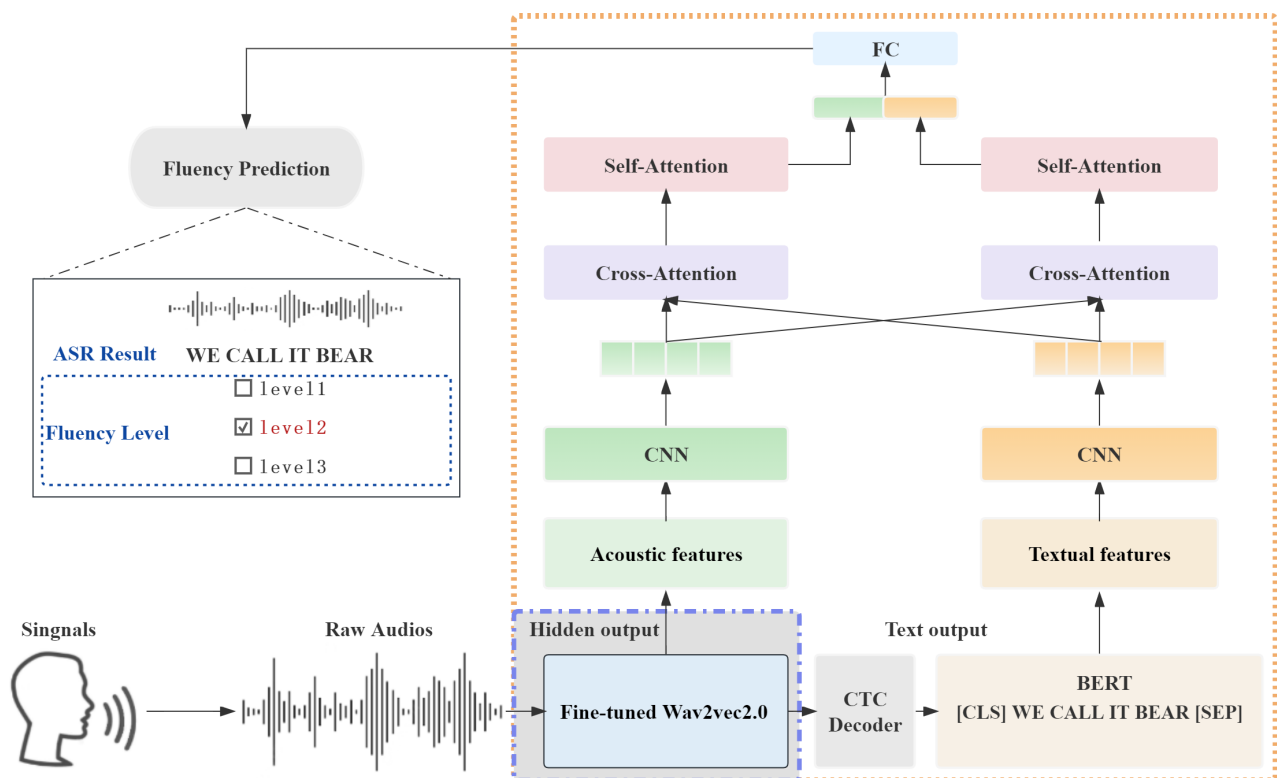


Figure 1. The architecture of the proposed model.

3.1. Fine-Tuning Wav2Vec2.0 Model

In recent years, the speech self-supervised pre-trained model has attracted much attention for its excellent performance. The combination of pre-trained and fine-tuning has been shown to meet the challenges posed by insufficient resources for annotated data, enabling models' performance benefits. Using speech self-supervised pre-trained models such as Wav2Vec2.0 instead of traditional audio feature extraction modules such as MFCC and Fbank can effectively extract disfluency features and build a more suitable speech fluency assessment model.

This section starts with a pre-trained Wav2Vec2.0 model that uses two different tasks to fine-tune specific downstream tasks. The architecture of the model is depicted in Figure 2 and can be divided into two components: a feature encoder based on CNN and a contextual encoder based on Transformer. To prevent interference and corruption of the CNN layers, we fixed all parameters of these CNN blocks and only fine-tuned the parameters of the Transformer.

We propose to fine-tune the Wav2Vec2.0 model with multi-task learning, input the original audio waveform, and produce two output paths (purple and green in Figure 2). We denote the input waveform as $X \in R^L$, where L is the length. We obtain the output features from the last hidden layer of Wav2Vec2.0:

$$Z = f_{Wav2Vec2.0}(X) \quad (1)$$

where $X \in R^L$ represents the original speech signal, L denotes the length, and $f_{Wav2Vec2.0}$ is the pre-trained Wav2Vec2.0 model.

As shown in Figure 2, the ASR and fluency assessment tasks use Z as input. The purple task is fluency assessment (FA): The FA model takes speech as input and produces fluency labels as output. The ASR task, represented by the green module, takes speech as input and outputs text. The purple task is the main task in our task, the green task plays a supporting role in the training, and the specific downstream tasks will be described in detail.

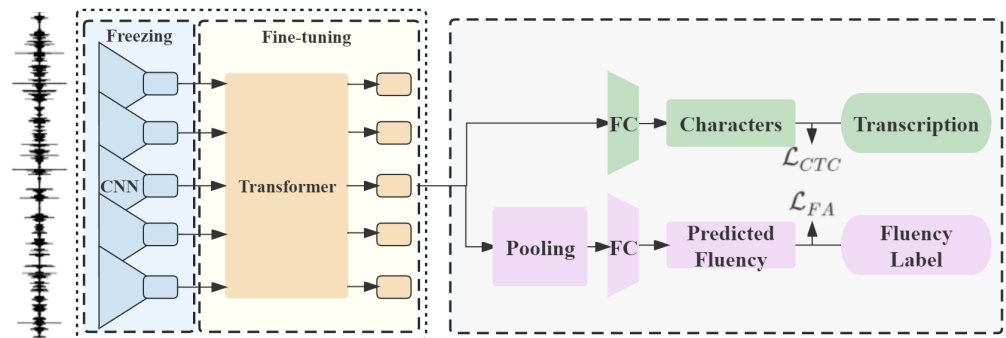


Figure 2. The proposed fine-tuning Wav2Vec2.0 model.

3.1.1. ASR Task

For the ASR task, after obtaining the features, we use a fully connected layer (the FC block in the figure) to map the features Z into Logits, after which we apply a softmax operator to convert the character predictions expressed in Logits to the probability vector y. The formula is as follows, where W and b represent the weight vector and bias:

$$y = \text{softmax}(WZ + b) \quad (2)$$

This task involves the use of connectionist temporal classification (CTC) loss to encode the provided transcription. CTC is a technique used to map input signals to output targets in situations where they have varying lengths and no alignment information is available. Given that the length of the speech signal is usually much longer than that of the transcription, CTC is employed as a loss function to quantify the degree of difference between the input sequence data and the actual output generated by the neural network. We compute the CTC loss as

$$\mathcal{L}_{CTC} = CTC(y, \text{transcription}) \quad (3)$$

3.1.2. FA Task

For the FA task, after obtaining the features, a sequence of length L is transformed into vectors by summarizing the length L through a pooling layer. Afterward, we also use a

fully connected layer to map the processed features \hat{Z} into Logits, after which we apply a softmax operator to convert the fluency predictions represented in Logits to the probability vector f :

$$f = \text{softmax}(W\hat{Z} + b) \quad (4)$$

In this task, we compute the cross-entropy loss between the probability vector and the fluency label:

$$\mathcal{L}_{\text{CrossEntropy}} = \text{CrossEntropy}(f, \text{label}) \quad (5)$$

3.1.3. Multi-Task

The model can be fine-tuned on the specific dataset to adjust to the new environment, and we adopt multi-task learning to optimize the joint loss:

$$\text{Min}\mathcal{L} = \mathcal{L}_{\text{CrossEntropy}} + \alpha\mathcal{L}_{\text{CTC}} \quad (6)$$

We use the hyperparameter α to combine the \mathcal{L}_{CTC} and $\mathcal{L}_{\text{CrossEntropy}}$ into one loss. When α is close to 0, the model will pay more attention to FA.

3.2. Modal Based on Multimodality

This section presents the network architecture in Figure 3. First, CNN is used to encode audio features and text features. Then one modality representation is fused into another using the cross-attention module respectively. Based on this, the feature representation is enhanced by utilizing self-attention modules, and the fluency assessment results are obtained from the fully connected layer. Then, we will describe our proposed model in detail.

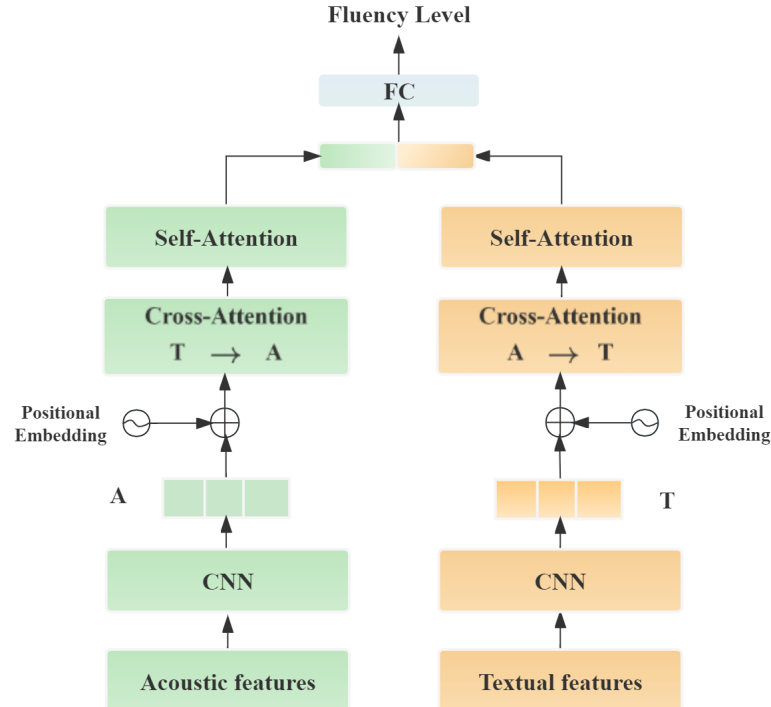


Figure 3. The architecture of the multimodal model.

3.2.1. Modality Encoding

The data consist of audio and text modalities, which are denoted by $X_a \in R^{T_a d_a}$ and $X_t \in R^{T_t \times d_t}$, where T_m and d_m denote sequence length and feature vector size of each modality. For text and audio modalities, convolution layers are used to extract feature representation. Specifically, we pass the input sequence through a 1D time convolution

layer. It is worth noting that here the temporal convolution projects features of different modalities into the same dimension:

$$Z_{\{a,t\}} = \text{Conv 1D}(X_{\{a,t\}}) \quad (7)$$

3.2.2. Multimodal Fusion

This subsection will focus on the fusion of different modalities inspired by Tsai [22]. Audio and text modalities play an important role in multimodal fluency assessment tasks. Attention is widely used in multimodal learning, which exploits the human mind to obtain more valuable information and ignores useless information. Cross-attention captures connections between modalities and enables modalities' fusion. Self-attention is utilized to capture the internal relevance of modality. Therefore, this module combines the strengths of cross-modal attention and self-attention to propose a multi-level cross-modal feature fusion method.

Figure 4 shows the architecture of the cross-attention module. The following formulas related to the cross-attention module are referenced from papers [22,37]. To pass audio (a) to text (t), denoted as " $a \rightarrow t$ ", we utilize attention modules that require three inputs: query matrix, key matrix, and value matrix. The modality fusion process can be expressed as follows, where W represents the weight:

$$\begin{aligned} Q_t &= Z_t W_{Q_t} \\ K_a &= Z_a W_{K_a} \\ V_a &= Z_a W_{V_a} \end{aligned} \quad (8)$$

The cross-attention Y_t , which represents the attention from audio (a) to text (t), can be expressed as follows:

$$\begin{aligned} Y_t &= CM_{a \rightarrow t}(Z_t, Z_a) \\ &= \text{softmax}\left(\frac{Q_t K_a^\top}{\sqrt{d_k}}\right) V_a \\ &= \text{softmax}\left(\frac{Z_t W_{Q_t} W_{K_a}^\top Z_a^\top}{\sqrt{d_k}}\right) Z_a W_{V_a} \end{aligned} \quad (9)$$

Following the cross-attention module, a residual connection and layer normalization are implemented to incorporate the original modality from the other modality:

$$z_{a \rightarrow t} = \text{LN}(Y_t + Z_t) \quad (10)$$

After applying layer normalization (LN), a feed-forward layer is utilized to combine the feature representations:

$$Z_{a \rightarrow t} = \text{LN}(z_{a \rightarrow t} + \text{FFN}(z_{a \rightarrow t})) \quad (11)$$

where FFN means fully connected feed-forward network. Similarly, we can get to the final embedding representation vector $Z_{t \rightarrow a}$ from t to a .

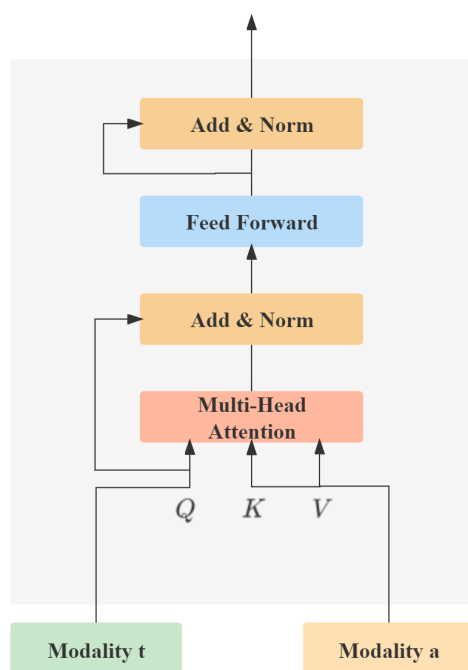


Figure 4. The architecture of the cross-attention module.

The self-attention is better at capturing the internal relevance of data. The difference between self-attention and cross-modal attention is that its query, key, and value are mapped from the same matrix. In the multimodal model, we put the $Z_{a \rightarrow t}$ and the $Z_{t \rightarrow a}$ through self-attention to enhance the information. Then, we concatenate the outputs and obtain the fluency assessment result by a layer of the fully connected network.

4. Experiments

4.1. Dataset

To validate the proposed method for spontaneous speech fluency assessment, we use the self-built PSCPSF dataset. Additionally, to demonstrate the applicability of the proposed method in other scenarios, such as read-aloud tasks, we conduct similar experiments using the publicly available Speechocean762 dataset. These experiments highlight the versatility of our proposed method for application in different assessment scenarios.

4.1.1. PSCPSF Dataset

PSCPSF is a self-built PSC propositional speaking fluency assessment dataset consisting of 3368 Chinese utterances and their corresponding scoring files. The PSC propositional speaking requires speakers to choose one of two given topics and speak freely for three minutes without reference text. All the topics in the dataset are selected from the “Topics for PSC” created by the National Putonghua Training and Testing Center. In the subsequent section, we will provide a detailed introduction to the PSCPSF dataset from three perspectives: dataset recording and processing, dataset scoring details, and dataset statistics.

During the data recording, the speakers are asked to simulate a real PSC exam scenario and complete a free speech on a relevant topic within three minutes. Some data are recorded in a silent office environment using a desktop computer and Walkers K815 headphones. Some are recorded in a designated App using a mobile phone as the recording device. The audio files are all saved in mono WAV format at 16,000 Hz. As there is no pre-existing reference text for PSC propositional speaking, we must rely on ASR transcriptions, which may contain errors. Therefore, Praat [38], an open source speech analysis software, is used for the alignment of audio and text files.

The manual scoring part of the dataset is evaluated using the National Putonghua Proficiency Test Propositional Speaking Scoring Criteria in seven dimensions: speech standardization, lexical and grammatical standardization, natural fluency, lack of time, relevance to the topic, reading aloud, and invalid discourse. This paper primarily focuses on the assessment of fluency, which employs three levels of fluency as presented in Table 1.

Table 1. Manual scoring metrics of fluency on the PSCPSF.

Fluency	Description
level1	Natural and fluent speech
level2	Generally fluent in language but poor in oral expression
level3	Incoherent speech with a stiff tone

We invited three experienced experts to rate our self-built dataset. As fluency assessment is highly subjective, the results among experts may differ. Therefore, we calculated the average score of the three experts as the primary basis for scoring, using the formula shown in Equation (12). In addition, we calculated the correlation coefficient between each expert's score and the expert average score, as shown in Table 2. The data in the table indicate that the expert average score is reliable and can be used as the label for training models.

$$\text{Expert average score} = \frac{\text{Expert1} + \text{Expert2} + \text{Expert3}}{3} \quad (12)$$

Table 2. Correlation coefficients for scoring the fluency between experts on the PSCPSF.

Pearson's Correlation Coefficient	Expert1	Expert2	Expert3
Expert average score	76.34%	86.25%	64.14%

Information on the specific distribution of the data is shown in Table 3 below. The test set consists of 1200 utterances from 70 propositional speaking files and 1777 unique words. The training set consists of 2168 utterances from 120 propositional speaking files and 2111 unique words. Figure 5 shows the distribution of fluency scores on the train and test sets. The figure shows that the amount of fluency is consistent across the data set.

Table 3. Distribution of the data on the PSCPSF.

Dateset	Amount	Total Unique Words	Duration (Hours)	Speakers	Gender Proportion
Train	2168	2111	5.95	63	2:1
Test	1200	1777	3.31	45	2:1

4.1.2. Speechocean762 Dataset

The Speechocean762 dataset is a collection of 5000 English utterances from 250 non-native speakers aged 6 to 43, aimed at assessing pronunciation. This dataset undergoes manual scoring by five experts at the word, phoneme, and sentence levels, covering multiple dimensions. The focus of this paper is on evaluating fluency at the sentence level, and the dataset provides a fluency score for each sentence with a scale of 10 points. However, this wealth of annotation needs to be fully exploited, and little research is dedicated to fluency assessment. Therefore, in the Speechocean762 dataset, the sentence-level fluency scores are extracted as research data.



Figure 5. The distribution of fluency scores on the training and test sets of PSCPSF.

Table 4 shows a detailed description of the scoring metrics for manual annotation regarding sentence fluency on the Speechocean762.

Table 4. Manual scoring metrics of sentence-level fluency on Speechocean762.

Fluency	Description
0–3	The speaker either cannot read the entire sentence, or there is no sound being produced
4–5	The speech exhibits incoherence, characterized by frequent pauses, repetitions, and stammering
6–7	Overall, the speech is coherent with some occasional pauses, repetitions, and stammering
8–10	The speech is coherent and does not exhibit noticeable pauses, repetition, or stammering

The sentence-level training set consisted of 2500 sentences and 15,849 words; the test set consisted of 2500 utterances and 15,967 words, and the sampling rate of all speech data is 16,000 Hz. Table 5 provides specific information on the distribution of the data, while Figure 6 illustrates the distribution of fluency scores in both the training and testing sets. We find that the range of fluency scores is mainly concentrated in the range of 7–9.

Table 5. Distribution of the data on Speechocean762.

Dateset	Amount	Word	Duration (Hours)	Speakers	Gender Proportion
Train	2500	15,849	2.88	125	1:1
Test	2500	15,967	2.69	125	1:1

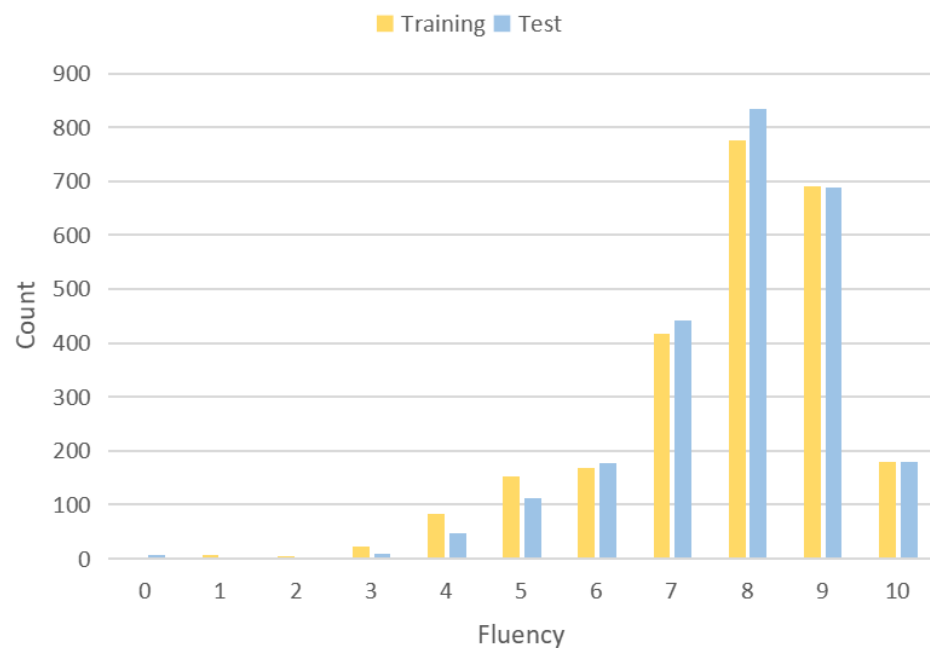


Figure 6. The distribution of fluency scores on the training and test sets of Speechocean762.

5. Results and Analysis

5.1. Experimental Settings

Our experiments use PyTorch [39], a flexible and efficient deep learning framework, on NVIDIA TESLA V100 GPU. For fine-tuning tasks, Pearson correlation (Corr), character error rate (Cer), and word error rate (Wer) are used in experiments to assess model performance. We conduct fine-tuned experiments for PSCPSF and Speechocean762 datasets using the pre-trained model: wav2vec-2.0-large and wav2vec-2.0-base, which are fine-tuned jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn and fine-tuned facebook/wav2vec2-base-960h. The implementations are based on the Huggingface transformers repository [40]. To extract features from the text modality, we utilize BERT [35], generating 1024-dimensional features on the PSCPSF dataset and 768-dimensional features on the Speechocean762 dataset. Tables 6 and 7 list the settings of hyperparameters that are used in fine-tuning and multimodal experiments. It is worth noting that the learning rates in the Table 6 are different when $\alpha = 0$. Since only fluency assessment tasks are performed at this time, it is necessary to reduce the learning rate to prevent significant oscillation after several epochs.

Table 6. The hyperparameters used for fine-tuning.

Hyperparameters	Setting
training epochs	100
optimizer	AdamW
α	0, 0.001, 0.01, 0.1, 1
learning rate	10^{-5} if $\alpha = 0$, 5×10^{-5} otherwise
batch size	8

Table 7. The hyperparameters used for multimodal model.

Hyperparameters	Setting
training epochs	20
optimizer	Adam
learning rate	10^{-3}
batch size	16

5.2. The Results of Fine-Tuned Wav2Vec2.0

To verify whether multi-task learning contributes to the FA task, we control the strength of the CTC loss by adjusting the hyper-parameter α to obtain different Corr. We adjust α from 0 to 1. When $\alpha = 0$, the model is trained exclusively on fluency labels. As the value of α increases, the importance of CTC loss also increases. To show the effect of CTC loss on the ASR component, we present the Cer and Wer obtained during the experiment for the PSCPSF and the Speechocean762 datasets. Tables 8 and 9 report Corr and Cer, and Wer for the five α choices on the two datasets. As seen in Tables 8 and 9, when α is set to 0, training only the FA task results in lower accuracy for ASR and a lower Corr compared to other multi-task training results. As we increase α , we observe promising results for ASR, as well as an increase in Corr for the FA task. This validates the effectiveness of the multi-task learning mechanism, where the auxiliary task ASR can effectively improve the performance of speech fluency assessment. By considering the performance of Corr along with Cer and Wer on both datasets, we select $\alpha = 1$ for subsequent experiments.

Table 8. The impact of the CTC loss on PSCPSF dataset.

The Value of Hyperparameter	Corr	Cer
$\alpha = 0$	0.765	3.68
$\alpha = 0.001$	0.783	0.176
$\alpha = 0.01$	0.805	0.160
$\alpha = 0.1$	0.818	0.151
$\alpha = 1$	0.824	0.150

Table 9. The impact of the CTC loss on Speechocean762 dataset.

The Value of Hyperparameter	Corr	Wer
$\alpha = 0$	0.750	0.999
$\alpha = 0.001$	0.755	0.274
$\alpha = 0.01$	0.763	0.273
$\alpha = 0.1$	0.763	0.301
$\alpha = 1$	0.765	0.273

5.3. The Results of Multimodal Model

This section analyzes the results of the multimodal model experiments from three perspectives. First, we compare the audio-based and text-based models to ascertain the significance of audio and text modality in fluency assessment. Second, we compare the effectiveness of different multimodal models in the evaluation task to verify the effectiveness of the model proposed in this paper. Furthermore, we compare the effects of text generated by ASR output and manually proofread text in multimodal experiments. Lastly, through ablation experiments, we verify the impact of different components of the model on the experimental results.

Our proposed model demonstrates superior performance on the PSCPSF and Speechocean762 datasets, as evidenced by the results presented in both Tables 10 and 11. Firstly, we conduct experiments on audio-based and text-based models. The experimental results show that the audio-based model outperforms the text-based model in speech fluency assessment, indicating that speech features are more important for fluency assessment. On the other hand, the text-based model does not perform as well as the audio-based model in the fluency assessment task. In the PSCPSF dataset, the text features after ASR assessed in the text-based model resulted in the Corr that is 0.103 higher than that of the manual translation. The text features in the Speechocean762 dataset after ASR only showed the Corr of 0.261 by the text-based model, while the manually translated text features showed the Corr of 0.585 by the text-based model. This indicates that text features influence fluency assessment, and the text results after different ASR have different effects. The text

features in Chinese may contain more disfluent features, resulting in the text-based model outperforming the PSCPSF on the Chinese dataset than the English dataset.

For the fusion of text features with audio features, other classical multimodal analysis models are chosen as comparison models in this paper. The later fusion DNN (LF-DNN) combines unimodal features by concatenating them prior to classification. The lowrank multimodal fusion (LMF) [41] employs modality-specific low-rank factors for effective multimodal fusion. The memory fusion network (MFN) proposes a multi-view continuous learning approach that uses delta-memory attention and multi-view gated memory to capture the interactions between time and patterns [42]. To investigate the performance of text features in different multimodal models, this study input fine-tuned Wav2Vec2.0 decoded text features and manually translated text features into various multimodal modalities. Figures 7 and 8 present a more intuitive display of the differences in fluency assessment results for different text features in various multimodal models. These results show that our proposed model outperforms most multimodal models mentioned in this paper. The text features obtained from human translation are better in the multimodal model than after ASR transcription. Experiments show that more accurate ASR is better in multimodal speech fluency assessment tasks. It is worth noting that since the Speechocean762 dataset has been widely used in pronunciation assessment models, we compare our fluency model with the four pronunciation assessment models on fluency assessment, including LSTM, GOPT [43], an SSL-based method proposed by Kim [44], and a multi-task learning method proposed by Wong [45]. The results are shown in Figure 9. Compared to the pronunciation detection models mentioned in this paper, which have excellent performance in the Speechocean762 dataset, our model performs well in terms of the fluency assessment.

To investigate the impact of distinct components on model performance, we conduct separate ablation experiments on both the PSCPSF and Speechocean762 datasets. Tables 12 and 13 present the findings of these experiments. First, we assess the impact of fine-tuning the Wav2Vec2.0 model on the experiments. We generate audio, and text features using the Wav2Vec2.0 model without fine-tuning and obtain the fluency results. The results show that the Corr of the features without fine-tuning decreased by 0.079 and 0.051 for the two datasets, respectively, indicating the effectiveness of the fine-tuning of the Wav2Vec2.0 model. Secondly, self-attention and cross-attention are removed separately. Self-attention had a more significant impact on the results than cross-attention, indicating that applying self-attention to encode the modal representation is necessary. The impact of the cross-attention module on fluency assessment results has not met expectations, possibly due to the influence of textual modality. However, the module's effectiveness may improve as the dataset expands and experiments are conducted to evaluate longer texts with more non-fluent markers. In addition, we removed the CNN to demonstrate the importance of feature extraction. The ablation study demonstrates that every component is essential and cannot be omitted.

Table 10. The performance on the PSCPSF dataset.

Model	Corr
Audio-based	0.878
Text-based	0.784
Text-based (Manual transcripts)	0.681
LF-DNN	0.868
LF-DNN (Manual transcripts)	0.863
LMF	0.869
LMF(Manual transcripts)	0.872
MFN	0.872
MFN (Manual transcripts)	0.874
Ours	0.879
Ours (Manual transcripts)	0.879

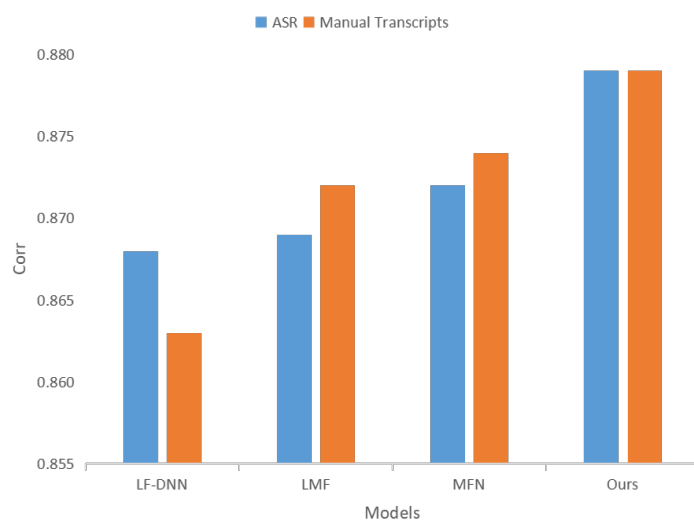


Figure 7. Effect of different text features in PSCPSF on experimental results.

Table 11. The performance on Speechoccean762.

Model	Corr
Audio-based	0.781
Text-based	0.261
Text-based (Manual transcripts)	0.585
LF-DNN	0.776
LF-DNN (Manual transcripts)	0.773
LMF	0.770
LMF (Manual transcripts)	0.774
MFN	0.770
MFN (Manual transcripts)	0.771
Ours	0.789
Ours (Manual transcripts)	0.790

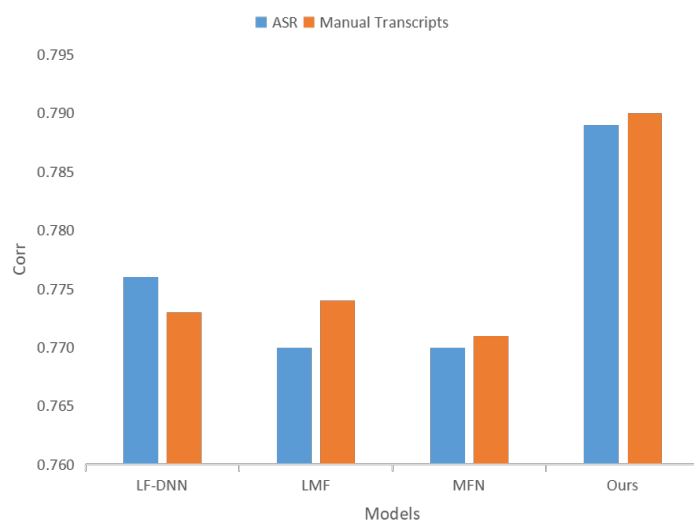


Figure 8. Effect of different text features in Speechoccean762 on experimental results.

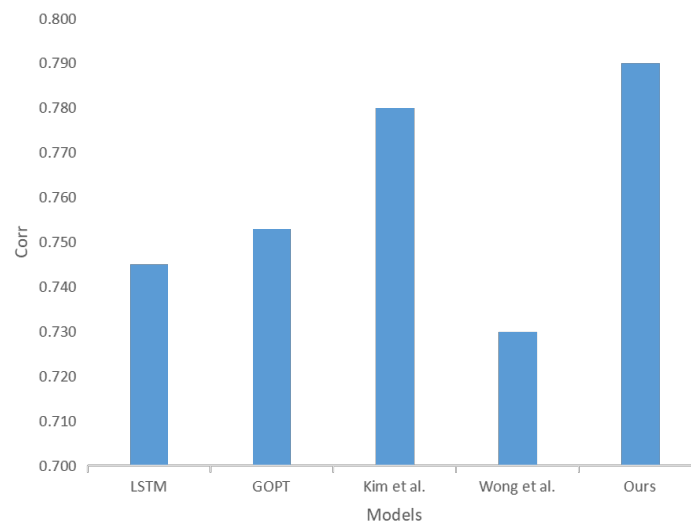


Figure 9. Comparison of different models on Speechocean762 of fluency scores.

Table 12. Ablation study on PSCPSF.

Model	Corr
Wav2Vec2.0 (Without fine-tuning)	0.800
self-attention (w/o)	0.874
cross-attention (w/o)	0.879
CNN(w/o)	0.876
Ours	0.879

Table 13. Ablation study on Speechocean762.

Model	Corr
Wav2Vec2.0 (Without fine-tuning)	0.738
self-attention (w/o)	0.784
cross-attention (w/o)	0.788
CNN(w/o)	0.782
Ours	0.789

6. Conclusions

Assessing speech fluency is a critical evaluation criterion for spontaneous speech. However, automatic fluency assessment of spontaneous speech is a challenging task that heavily relies on the accuracy of ASR. In this paper, we propose a multimodal method for the automatic speech fluency assessment of spontaneous speech that combines ASR output. We first create a dataset called PSCPSF for propositional speaking fluency assessment, consisting of 9.26 h of speech. To validate the relevance of the speech assessment task to the ASR task, the multi-task learning framework is used to fine-tune the Wav2Vec2.0 model on the specific datasets. By jointly optimizing the model using the speech fluency assessment task as the main task and the ASR task as the auxiliary task, we are able to improve the accuracy of the model in speech fluency assessment. We then propose a multimodal automatic fluency assessment method that combines ASR output. Specifically, we combine the final layer of the fine-tuned Wav2Vec2.0 model's hidden output with the decoded text output, followed by a cross-modal attention mechanism to fuse audio and text features. This results in more effective fluency features, allowing for more accurate speech fluency assessment. Experiments on the PSCPSF and Speechocean762 datasets for spontaneous speech and read-aloud scenarios demonstrate that our method performs well in the different assessment scenarios. In the future, our focus will be on three main areas. First, we plan to expand the size of our self-built dataset, which will enable us

to obtain more reliable results for the fine-tuning of Wav2Vec2.0 and the evaluation of multimodal methods using larger datasets. Second, we will address the issue that the impact of text features on the experimental results is not as significant as that of speech features. To address this, we will annotate non-fluent text in spontaneous speech, such as affective words and repetitions that affect fluency. Finally, we aim to explore more efficient multimodal evaluation methods to continuously improve the accuracy of fluency assessment and ASR.

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

Funding: This research was funded by the National Science Foundation of China, under grant U1903213, and the Basic Research Program of Tianshan Talent Plan of Xinjiang, China, under grant 2022TSYCJU0005.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Speechocean762 dataset during the current study is available at <https://www.openslr.org/101>, accessed on 1 March 2023. The self-built PSCPSF dataset is in the process of being expanded and has not yet been publicly published.

Acknowledgments: The authors gratefully acknowledge all anonymous reviewers and editors for their constructive suggestions for the improvement of this paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
PSC	Putonghua Shuiping Ceshi
DNN	Deep Neural Network
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GOP	Goodness of Pronunciation
MTL	Multi-task Learning
CTC	Connectionist Temporal Classification
Corr	Pearson Correlation
Cer	Character Error Rate
Wer	Word Error Rate

References

1. Riggensbach, H. Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Process*. **1991**, *14*, 423–441. [CrossRef]
2. Zechner, K.; Higgins, D.; Xi, X.; Williamson, D.M. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Commun.* **2009**, *51*, 883–895. [CrossRef]
3. Bhat, S.; Hasegawa-Johnson, M.; Sproat, R. Automatic fluency assessment by signal-level measurement of spontaneous speech. In Proceedings of the Second Language Studies: Acquisition, Learning, Education and Technology, Tokyo, Japan, 22–24 September 2010.
4. Hirabayashi, K.; Nakagawa, S. Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.
5. Deng, H.; Lin, Y.; Utsuro, T.; Kobayashi, A.; Nishizaki, H.; Hoshino, J. Automatic fluency evaluation of spontaneous speech using disfluency-based features. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 9239–9243.

6. Deshmukh, O.D.; Kandhway, K.; Verma, A.; Audhkhasi, K. Automatic evaluation of spoken English fluency. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 4829–4832.
7. Tong, R.; Lim, B.P.; Chen, N.F.; Ma, B.; Li, H. Subspace Gaussian mixture model for computer-assisted language learning. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 5347–5351.
8. van Dalen, R.C.; Knill, K.M.; Gales, M.J. *Automatically Grading Learners' English Using a Gaussian Process*; ISCA: Narrabeen, Australia, 2015.
9. Mao, S.; Wu, Z.; Jiang, J.; Liu, P.; Soong, F.K. NN-based Ordinal Regression for Assessing Fluency of ESL Speech. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7420–7424.
10. Fontan, L.; Coz, M.L.; Detey, S. Automatically Measuring L2 Speech Fluency without the Need of ASR: A Proof-of-concept Study with Japanese Learners of French. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, 2–6 September 2018; pp. 2544–2548.
11. Hassanali, K.N.; Yoon, S.Y.; Chen, L. Automatic scoring of non-native children's spoken language proficiency. In Proceedings of the SLaTE, Leipzig, Germany, 4–5 September 2015; pp. 13–18.
12. Loukina, A.; Zechner, K.; Bruno, J.; Klebanov, B.B. Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1–12.
13. Cheng, J.; Chen, X.; Metallinou, A. Deep neural network acoustic models for spoken assessment applications. *Speech Commun.* **2015**, *73*, 14–27. [[CrossRef](#)]
14. Metallinou, A.; Cheng, J. Using deep neural networks to improve proficiency assessment for children English language learners. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
15. Chung, H.; Lee, Y.K.; Lee, S.J.; Park, J.G. Spoken english fluency scoring using convolutional neural networks. In Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, Republic of Korea, 1–3 November 2017; pp. 1–6.
16. Deng, H.; Utsuro, T.; Kobayashi, A.; Nishizaki, H. Comparison of Static and Time-Sequential Features in Automatic Fluency Detection of Spontaneous Speech. In Proceedings of the 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Singapore, 18–20 November 2021; pp. 158–163.
17. Zhang, H.; Shi, K.; Chen, N.F. Multilingual Speech Evaluation: Case Studies on English, Malay and Tamil. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czech Republic, 30 August–3 September 2021; pp. 4443–4447.
18. Fu, K.; Gao, S.; Tian, X.; Li, W.; Ma, Z.; Bytedance, A. Using Fluency Representation Learned from Sequential Raw Features for Improving Non-native Fluency Scoring. *Proc. Interspeech* **2022**, 4337–4341.
19. Yoon, S.Y.; Bhat, S. A comparison of grammatical proficiency measures in the automated assessment of spontaneous speech. *Speech Commun.* **2018**, *99*, 221–230. [[CrossRef](#)]
20. Zeghidour, N.; Usunier, N.; Synnaeve, G.; Collobert, R.; Dupoux, E. End-to-End Speech Recognition From the Raw Waveform. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, 2–6 September 2018; pp. 781–785.
21. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [[CrossRef](#)]
22. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 2019, p. 6558.
23. Feng, H.; Ueno, S.; Kawahara, T. End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 501–505.
24. Cai, X.; Yuan, J.; Zheng, R.; Huang, L.; Church, K. Speech Emotion Recognition with Multi-Task Learning. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czechia, 30 August–3 September 2021; pp. 4508–4512.
25. Ashraf, M.; Abid, F.; Din, I.U.; Rasheed, J.; Yesiltepe, M.; Yeo, S.F.; Ersoy, M.T. A Hybrid CNN and RNN Variant Model for Music Classification. *Appl. Sci.* **2023**, *13*, 1476. [[CrossRef](#)]
26. Lane, I.R.; Kawahara, T.; Matsui, T. Language model switching based on topic detection for dialog speech recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hongkong, China, 6–10 April 2003; Volume 1, pp. 616–619.
27. Farooq, M.S.; Khalid, H.; Arooj, A.; Umer, T.; Asghar, A.B.; Rasheed, J.; Shubair, R.M.; Yahyaoui, A. A Conceptual Multi-Layer Framework for the Detection of Nighttime Pedestrian in Autonomous Vehicles Using Deep Reinforcement Learning. *Entropy* **2023**, *25*, 135. [[CrossRef](#)] [[PubMed](#)]

28. Caruana, R. *Multitask Learning*; Springer: Berlin, Germany, 1998.
29. Kim, S.; Hori, T.; Watanabe, S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4835–4839.
30. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. 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; pp. 369–376.
31. Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 15–19 September 2019; pp. 2803–2807.
32. Atmaja, B.T.; Sasou, A.; Akagi, M. Speech emotion and naturalness recognitions with multitask and single-task learnings. *IEEE Access* **2022**, *10*, 72381–72387. [[CrossRef](#)]
33. Li, Y.; Bell, P.; Lai, C. Fusing asr outputs in joint training for speech emotion recognition. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7362–7366.
34. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
35. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
36. Hussain, S.; Nguyen, V.; Zhang, S.; Visser, E. Multi-task voice activated framework using self-supervised learning. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6137–6141.
37. Xie, B.; Sidulova, M.; Park, C.H. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors* **2021**, *21*, 4913. [[CrossRef](#)] [[PubMed](#)]
38. Boersma, P. Praat, a system for doing phonetics by computer. *Glott. Int.* **2001**, *5*, 341–345.
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
40. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
41. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.B.; Morency, L.P. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 5–20 July 2018; pp. 2247–2256.
42. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
43. Gong, Y.; Chen, Z.; Chu, I.H.; Chang, P.; Glass, J. Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7262–7266.
44. Kim, E.; Jeon, J.J.; Seo, H.; Kim, H. Automatic pronunciation assessment using self-supervised speech representation learning. *arXiv* **2022**, arXiv:2204.03863.
45. Wong, J.H.; Zhang, H.; Chen, N.F. Variations of multi-task learning for spoken language assessment. *Proc. Interspeech* **2022**, 4456–4460. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.