

## Article

# A Pronunciation Prior Assisted Vowel Reduction Detection Framework with Multi-Stream Attention Method

Zongming Liu <sup>1,2</sup>, Zhihua Huang <sup>2,3</sup>, Li Wang <sup>1,\*</sup> and Pengyuan Zhang <sup>1,2</sup>

<sup>1</sup> Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China; liuzongming@hcl.ioa.ac.cn (Z.L.); zhangpengyuan@hcl.ioa.ac.cn (P.Z.)

<sup>2</sup> University of Chinese Academy of Sciences, Beijing 100049, China; echohzh@163.com

<sup>3</sup> The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China

\* Correspondence: wangli@hcl.ioa.ac.cn

**Abstract:** Vowel reduction is a common pronunciation phenomenon in stress-timed languages like English. Native speakers tend to weaken unstressed vowels into a schwa-like sound. It is an essential factor that makes the accent of language learners sound unnatural. To improve vowel reduction detection in a phoneme recognition framework, we propose an end-to-end vowel reduction detection method that introduces pronunciation prior knowledge as auxiliary information. In particular, we have designed two methods for automatically generating pronunciation prior sequences from reference texts and have implemented a main and auxiliary encoder structure that uses hierarchical attention mechanisms to utilize the pronunciation prior information and acoustic information dynamically. In addition, we also propose a method to realize the feature enhancement after encoding by using the attention mechanism between different streams to obtain expanded multi-streams. Compared with the HMM-DNN hybrid method and the general end-to-end method, the average F1 score of our approach for the two types of vowel reduction detection increased by 8.8% and 6.9%, respectively. The overall phoneme recognition rate increased by 5.8% and 5.0%, respectively. The experimental part further analyzes why the pronunciation prior knowledge auxiliary input is effective and the impact of different pronunciation prior knowledge types on performance.

**Keywords:** CALL; CAPT; second language acquisition; vowel reduction detection; end-to-end ASR



**Citation:** Liu, Z.; Huang, Z.; Wang, L.; Zhang, P. A Pronunciation Prior Assisted Vowel Reduction Detection Framework with Multi-Stream Attention Method. *Appl. Sci.* **2021**, *11*, 8321. <https://doi.org/10.3390/app11188321>

Academic Editors: Changchun Bao and Yoshinobu Kajikawa

Received: 14 August 2021

Accepted: 5 September 2021

Published: 8 September 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 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/>).

## 1. Introduction

With the widespread popularity of second language acquisition applications on smart devices, more powerful computer-assisted pronunciation training (CAPT) systems are needed. In addition to focusing on each phoneme's pronunciation in a sentence, CAPT systems are expected to help learners practice more native-like accents. Prosody and suprasegmental related pronunciation phenomena are important for learning native-like speaking skills. Some research has paid attention to language learners' lexical stress [1,2], intonation [3], and vowel reduction [4].

Vowel reduction is a common language phenomenon in stress-timed languages. In British as well as American varieties of English, there is a tendency for most vowels in weakly stressed syllables to approach schwa in quality [5]. Some researchers concluded that the inability of second language (L2) learners to perform appropriate vowel reduction contributed to their non-native-like production of English [6].

In previous studies, acoustic characteristics of vowel reduction are concerned [7–11]. However, few studies have contributed to the area of automatically detecting vowel reduction. Vowel reduction detection is expected to recognize the reduced vowel phoneme from an utterance of speech. The results are compared with standard or automatically predicted pronunciation templates [4] to analyze whether the vowel reduction is correct. The method can be used as a function of CAPT systems. Learners can master the pronunciation of

vowel reduction through pronunciation practice so that their pronunciation sounds natural. When they communicate with others in a second language, they will be easily understood by others.

Among related works, the most relevant research for the vowel reduction detection task is [4]. They developed a feedback system that includes two main parts: vowel reduction prediction and vowel reduction detection. Each part contains multiple sub-modules for analyzing text or speech, extracting several handcrafted features, and training prediction models or detection models. A final feedback decision model integrates the output of the prediction and detection models and gives results on whether the vowel reduction is correct. This paper is committed to proposing a vowel reduction detection method based on automatic speech recognition (ASR) rather than a complex system designed for the vowel reduction feedback function. Our method is easy to integrate with the widely used ASR-based CAPT systems.

ASR-based automatic detection of vowel reduction faces two major challenges. One is that this kind of data annotation requires professional phonetics knowledge, leading to a lack of annotated data. The other is that there is no fixed rule for vowel reduction. Its location and the articulation manner are highly influenced by context [5,12]. The free phone recognition methods used in many other CAPT tasks may not learn this weak pattern through pure data driving methods due to insufficient data. A solution is to introduce information that affects the vowel reduction as an auxiliary. For example, sentence stress was adopted to predict and detect vowel reductions in [4].

Inspired by the idea of adding auxiliary information, we studied more efficient methods for using it. Due to conditional independence assumptions, a conventional HMM-based acoustic model has poor ability to modeling the influence of long-term context. Many end-to-end (E2E) methods have been proposed to alleviate this problem [13–15]. The attention-based end-to-end method solves the ASR problem as a sequence mapping from speech feature sequences to text using an encoder–decoder architecture [16]. It utilizes long-term contextual information more directly. Moreover, the reference text is usually available in CAPT tasks. It is a sentence that L2 learners are asked to read aloud during pronunciation training. Some researchers have tried to incorporate this kind of prior knowledge into an E2E model for mispronunciation detection and diagnose (MDD) [17–19]. However, these methods are not designed for vowel reduction detection. Since vowel reduction is not included in the pronunciation prior knowledge, the model needs to adjust the degree of dependence on auxiliary information dynamically. In other words, it is expected to use auxiliary information to determine possible vowel reduction positions, and at the same time, to detect whether there is vowel reduction in acoustic information. There are some methods for fusing multi-stream input, such as the direct merging of high-dimensional feature vectors and the hierarchical attention dynamic fusion method used by [20,21]. However, these methods treat different inputs as independent streams. In our scenario, the prior articulation input and the acoustic signal input are related. Specifically, they are related in content but not completely corresponding in specific pronunciation. For example, there is no vowel reduction information in the pronunciation prior.

This paper proposes a novel vowel reduction detection framework that uses prior pronunciation knowledge as auxiliary input. The innovations of our proposed work are as follows:

- Among the methods for this task, it is the first to perform an E2E framework that is easy to integrate into widely used ASR-based CAPT systems;
- A method of adopting the auxiliary encoder to utilize the prior information of pronunciation is proposed and several prior types of pronunciation are designed;
- A multi-stream mechanism is proposed. It uses the attention mechanism to process the association between the speech signal and the prior knowledge of pronunciation and generates a fusion information stream, which is sent to the back-end with the original coded information streams.

Taking advantage of the hierarchical attention mechanism, encoding streams are dynamically weighted and fused to highlight the more informative streams. In the experimental part, we compare our proposed method with the two baselines and the CNN-RNN-CTC method [22]. Then, the effects of the auxiliary encoder and the multi-stream expansion method are analyzed. We also studied the impact of different auxiliary input types on performance.

The structure of our paper is designed as follows: Section 2 describes the methods, Section 3 presents the experimental setup, Section 4 discusses the experimental results, Section 5 concludes.

## 2. Method

We performed vowel reduction detection on a multi-input fusion E2E framework trained by the Connectionist Temporal Classification (CTC)/attention-based multi-task learning scheme. It outputs a phoneme sequence containing vowel reduction labels. In the fusion framework, pronunciation prior knowledge is adopted as an auxiliary input for the hierarchical attention mechanism. A specially designed attention mechanism can fuse acoustic signal input and pronunciation prior knowledge input to generate a fusion stream, which is provided to the hierarchical attention part together with the above inputs. Two methods of automatically generating auxiliary input through reference text are implemented.

### 2.1. CTC-Attention Based Multi-Task Learning

Taking advantage of explicitly using the history of the target character without any conditional independence assumptions, the attention-based encoder–decoder framework has often been shown to improve the performance over many sequence-to-sequence tasks. To mitigate misalignment between the speech and labels, a joint CTC-attention model within the multi-task learning framework was put forward in [23].

The E2E model maps  $T$ -length feature sequence  $X = \{x_t \in \mathbb{R}^D \mid t = 1, 2, \dots, T\}$  to an  $L$ -length label sequence  $C = \{c_l \in \mathcal{U} \mid l = 1, 2, \dots, L\}$ , where  $D$  is the dimension of feature and  $\mathcal{U}$  is a set of distinct labels. The encoder is typically Bidirectional Long Short-Term Recurrent (BLSTM) layers shared by both attention and CTC networks. The hidden vector  $h_t$  at input feature index  $t$  is derived by an encoder when modeling temporal dependencies of the input sequence:

$$h_t = \text{Encoder}(X). \quad (1)$$

In the attention-based encoder–decoder method, a weighted summation is performed on  $h_t$  as shown in Equation (2):

$$r_l = \sum_{t=1}^T a_{lt} h_t, \quad (2)$$

where  $a_{lt}$  is obtained from a content-based attention mechanism using previous decoder state  $q_{l-1}$ :

$$a_{lt} = \text{ContentAttention}(q_{l-1}, h_t). \quad (3)$$

An LSTM-based decoder network predicts the next label based on  $r_l$  and the previous prediction. The whole architecture is optimized by a multi-task objective function with joint CTC and attention loss as follows:

$$\mathcal{L} = \lambda \log p_{ctc}(C \mid X) + (1 - \lambda) \log p_{att}^+(C \mid X), \quad (4)$$

where  $0 \leq \lambda \leq 1$  is a trade-off parameter set in advance. During the decoding phase, the joint model uses a label-synchronous beam search scheme which jointly predicts the next character. Given the input feature sequence  $X$ , the most probable label sequence is computed as:

$$\hat{C} = \arg \max_{C \in \mathcal{U}^*} \{\lambda \log p_{ctc}(C \mid X) + (1 - \lambda) \log p_{att}(C \mid X)\} \quad (5)$$

## 2.2. Automatic Auxiliary Input Sequence Generation

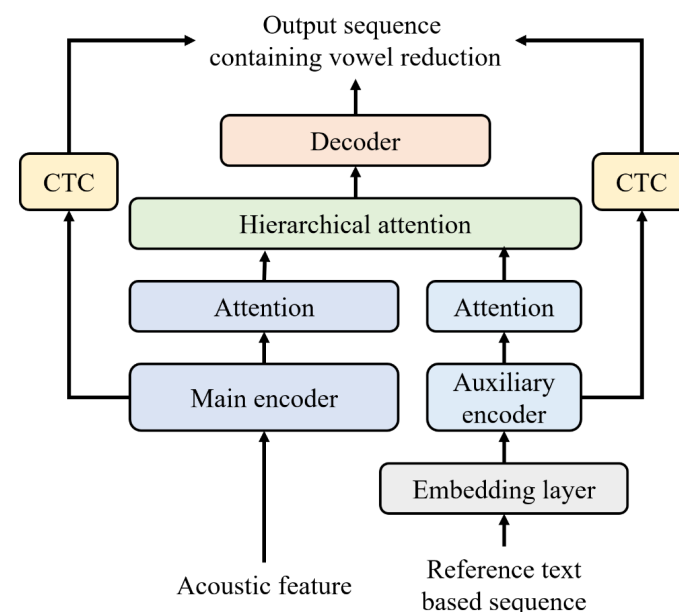
As described above, pronunciation prior knowledge based on the reference text is considered to be introduced into the E2E model. However, due to the mispronunciation of L2 learners, this prior only has a strong correlation with the actual pronunciation, they are not the same. Considering that, in our approach, the reference text is used to generate an auxiliary input to help the model obtain canonical transcripts. In detail, we propose two methods to acquire reference-text-derived sequences. The first method is a G2P (Grapheme-to-Phoneme) conversion method which uses CMU pronunciation dictionary [24] to obtain the phoneme sequence with lexical stress marks. The second method is to use the letter sequence of each word directly. Table 1 shows the auxiliary input sequences generated by the two methods. All these methods can be performed automatically without manual labeling.

**Table 1.** Input sequences generated by automatic methods.

Text	The Bedroom Wall
G2P	DH AH0 B EH1 D R UW2 M W AO1 L
Letter	t h e b e d r o o m w a l l

## 2.3. Pronunciation Prior Knowledge Assisted Multi-Encoder Structure

The design of the pronunciation prior knowledge assisted approach is inspired by [20], which used multi-encoder architecture to process speech information from a microphone-array. This structure is only designed for the same types of inputs. As the reference text has a strong correspondence with the recognition target, the decoder will over-rely on the text encoding. It will cause the audio encoding to be ignored, making the model insensitive to incorrect pronunciation. Meanwhile, the model cannot detect vowel reduction through a reference text sequence that does not contain vowel reduction information. We propose a main and auxiliary multi-encoder method to alleviate this problem, limiting the single encoder's over-reliance through model complexity. As shown in Figure 1, a network with a larger parameter amount is used as the main encoder to perform strong encoding on the filter bank acoustic feature input, and another network with a smaller parameter amount is used as the auxiliary encoder to perform weak encoding on the pronunciation prior knowledge input. Before the auxiliary encoder, we added a word embedding layer to map the generated text sequence to a semantic vector space.



**Figure 1.** The E2E model with pronunciation prior input.

## 2.4. Multi-Stream Information Fusion

Varieties of fusion techniques, such as the encoder's output concatenation or addition, are available for incorporating information from multi-encoders. However, these methods are not necessarily effective for the pronunciation prior knowledge assisted multi-encoder structure because of the different lengths of input sequences. In addition, as discussed in paragraph 4 of Section 1, the vowel reduction detection task requires dynamic adjustment of different inputs' weights. We take advantage of the hierarchical attention mechanism. In detail, context vectors,  $r_l^1$  and  $r_l^2$  from the main and auxiliary encoders are computed in a similar way to Equation (2). The fusion vector containing multi-stream information is obtained as a combination of  $r_l^1$  and  $r_l^2$  as follows:

$$r_l = \beta_{l1}r_l^1 + \beta_{l2}r_l^2, \quad (6)$$

$$\beta_{li} = \text{ContentAttention}(q_{l-1}, r_l^i), i = 1, 2. \quad (7)$$

The hierarchical attention weights  $\beta_{l1}$  and  $\beta_{l2}$  are estimated by the previous decoder state  $q_{l-1}$  and  $r_l^1, r_l^2$ .

## 2.5. Multi-Stream Expansion Based on Location-Aware Attention Mechanism

As shown in Figure 1, in the above structure, the encoded features of the acoustic information from the main encoder and the encoded features of the auxiliary encoder's pronunciation prior knowledge are calculated separately. In this process, the correlation between pronunciation prior and phonetic-specific is not considered. We propose a multi-stream expansion method by introducing the attention mechanism to calculate the correlation between the main and auxiliary inputs, obtain enhanced information streams, and provide richer sources for subsequent information fusion. The location-aware attention mechanism [25] introduces the attention weight calculated in the previous step as location awareness in the general attention calculation. In this paper, we use the attention weight  $a_o$ , used to calculate the origin stream output  $r$ , as the location awareness. It serves as a constraint to stabilize the process of enhancing features.

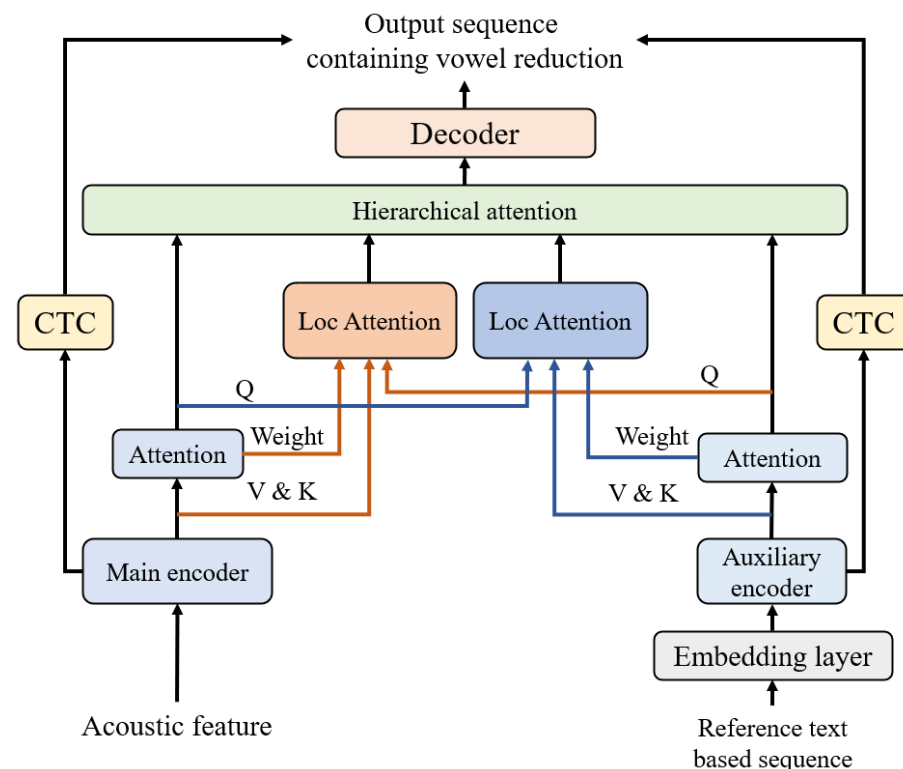
Specifically, the output  $h_t$  of the acoustic signal encoder is used as the *Key* and the *Value*. The output vector  $r_l$  obtained by Equation (2) after the prior encoder.  $r_l$  is used as the *Query* to calculate the attention weight as in Equation (8).

$$a_{lt} = \text{LocAttention}(r_l, h_t, a_o). \quad (8)$$

By using this weight to perform a weighted summation on the encoder output in the original stream, we obtain the enhanced acoustic vector  $r_e$  as follows:

$$r_e = \sum_{t=1}^T a_{lt}h_t. \quad (9)$$

In the same way, the enhanced pronunciation prior vector can be obtained. This process can be more intuitively understood from Figure 2.



**Figure 2.** Multi-stream expansion based on location-aware attention mechanism.

### 3. Experimental Setup

#### 3.1. Dataset Preprocessing and Acoustic Features

We performed experiments on the TIMIT corpus [26]. It is a reading speech corpus developed to provide speech data for acoustic–phonetic research studies and the evaluation of automatic speech recognition systems. It was produced by people from the eight major dialect regions of the United States. All data are manually divided and annotated at the phoneme level.

The phoneme sequences in the data annotation are preprocessed. As described in the TIMIT corpus documentation, /ix/ /ax/ are used to represent schwas produced by vowel reduction. /ix/ is usually used for schwas between two alveolars (“roses” /r ow z ix z/), otherwise /ax/ is used (“ahead” /ax hh eh d/). Therefore, in the experiments, we still distinguish these two kinds of schwa. /ih/ and /ah/ are the phonemes that most similar to /ix/ /ax/. Generally speaking, when there is no stress on /ih/ and /ah/, they are tend to be reduced to schwa. In our experiments, /ih/ /ah/ /ix/ /ax/ are marked as  $ih_{normal}$ ,  $ah_{normal}$ ,  $ih_{schwa}$ ,  $ah_{schwa}$ .

Data set division and acoustic feature extraction are performed in the usual way. We have maintained the TIMIT suggested training/test subdivision, which follows criteria containing restrictions on speaker, gender, text overlap, and occurrence times of phonemes. The development set is divided from the training set with reference to the TIMIT recipe in the ESPNET toolkit [27]. The division of the data set and the number of phonemes related to vowel reduction in each subset are shown in Table 2. We also use the TIMIT recipe in the ESPNET toolkit to extract 23-dimensional filter bank features and 3-dimensional fundamental frequency features to form a total of 26-dimensional acoustic features for each experiment.



**Table 2.** Number of sentences and phonemes in each subset.

Subset	Sentences	<i>ah</i> <i>Schwa</i>	<i>ah</i> <i>Normal</i>	<i>ih</i> <i>Schwa</i>	<i>ih</i> <i>Normal</i>
Train	3696	3892	2266	7370	4248
Dev	400	403	248	757	444
Test	192	186	135	377	203

### 3.2. HMM-DNN Hybrid Baseline

We build a basic HMM-DNN hybrid system using the Kaldi toolkit [28] to compare with the E2E framework. It follows the training method and configuration in the Kaldi toolkit. After the training of mono phone and triphone, the Nnet3 script is used to train a time-delay neural network (TDNN), consisting of five layers, each with 650 nodes. The time delay is configured as  $(-2, -1, 0, 1, 2)$   $(-1, 0, 1)$   $(-1, 0, 1)$   $(-3, 0, 3)$   $(-6, -3, 0)$ . The decoding process uses the default beam search method with the beam size set to 10.

### 3.3. Settings in Our Proposed Method

The E2E framework in this paper is implemented by the ESPnet toolkit. The main encoder, namely the acoustic feature encoder, contains four BLSTM layers, in which each layer has 320 cells in both directions followed by a linear projection layer. The auxiliary encoder, namely the pronunciation prior knowledge encoder, consists of a text embedding layer and two BLSTM layers. The text embedding layer is used to map each element in the text sequence to a 15-dimensional vector space, and each BLSTM layer has 80 cells in both directions, followed by a linear projection layer. Similarly, the frame-level attention mechanism's content-based attention mechanism has 320 attention units for the main encoder and 80 attention units for the auxiliary encoder. The content-based attention mechanism used in the encoder-level has 320 attention units. The decoder is a one layer BLSTM-based network with 300 cells in both directions. For the multi-stream expanded attention mechanism, all attention modules adopt the same structure, and each attention module contains 80 units.

### 3.4. Evaluation Metrics

Our research focused on the model's ability to distinguish between reduced and unreduced phonemes as well as the overall phoneme recognition rate. We employed the F1 score, which is a harmonic mean of precision and recall, for comprehensive performance evaluation of vowel reduction detection. It is defined as follows:

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (10)$$

$$\text{Precision} = C_{D \cap H} / C_D, \quad (11)$$

$$\text{Recall} = C_{D \cap H} / C_H, \quad (12)$$

where  $C_D$  is the total number of  $ih_{schwa}$  or  $ah_{schwa}$  that output from the model.  $C_{D \cap H}$  is the number of  $ih_{schwa}$  or  $ah_{schwa}$  that are correctly detected after a dynamic time warping (DTW) among outputs and human transcripts.  $C_H$  is the total number of  $ih_{schwa}$  or  $ah_{schwa}$  in the test set labeled by a human.

Considering that our proposed vowel reduction detection method can be flexibly integrated into ASR-based CAPT systems, we also counted the phoneme correct rate (PCR), which is commonly used in ASR research.

## 4. Experimental Results and Discussion

### 4.1. Comparison with HMM-DNN Hybrid Baseline System and Related Work

The first set of experiments compares the end-to-end method with the traditional HMM-DNN hybrid method and the CNN-RNN-CTC method [22] in terms of vowel reduction detection and phoneme recognition. The CNN-RNN-CTC method was originally

designed for the mispronunciation detection and diagnosis (MDD) task. Recently, it has been cited by some CAPT studies for comparison [18,29–31]. Since it is an E2E ASR-based method, as is ours, we implement it on the vowel reduction detection task as a related work for comparison. Table 3 lists the results of HMM-DNN hybrid method, the CNN-RNN-CTC method, the general E2E method, and our proposed E2E method with pronunciation prior knowledge assistance. It shows that the general E2E method has a slight advantage in  $ih_{schwa}$  phoneme detection and overall phoneme recognition performance compared with the HMM-DNN method. Furthermore, our proposed E2E method has more improvements in various indicators compared with the traditional HMM-DNN method. Since the vowel reduction phenomenon will be affected by the context, E2E methods may benefit from directly modeling the whole sequence. The CNN-RNN-CTC method is superior to the HMM-DNN method and the general E2E method in the vowel reduction detection performance, but it has obvious defects in phoneme recognition performance. Furthermore, it is inferior to our proposed method in both performances. The reason the CNN-RNN-CTC method exceeds the general E2E method in the detection of vowel reduction may be that the convolutional layers can obtain the frequency characteristics related to this phenomenon.

**Table 3.** Performance of proposed method, CNN-RNN-CTC method and baselines.

Model	$ih_{schwa}$		$ah_{schwa}$		PCR (%)
	Pre/Rec	F1	Pre/Rec	F1	
Hybrid	0.58/0.64	0.61	0.54/0.53	0.53	77.3
CNN-RNN-CTC [22]	0.64/0.67	0.65	0.57/0.56	0.56	74.7
E2E	0.59/0.68	0.63	0.53/0.54	0.53	77.9
E2E + prior	0.64/0.67	0.66	0.62/0.55	0.58	81.8

#### 4.2. Analysis of Auxiliary Input

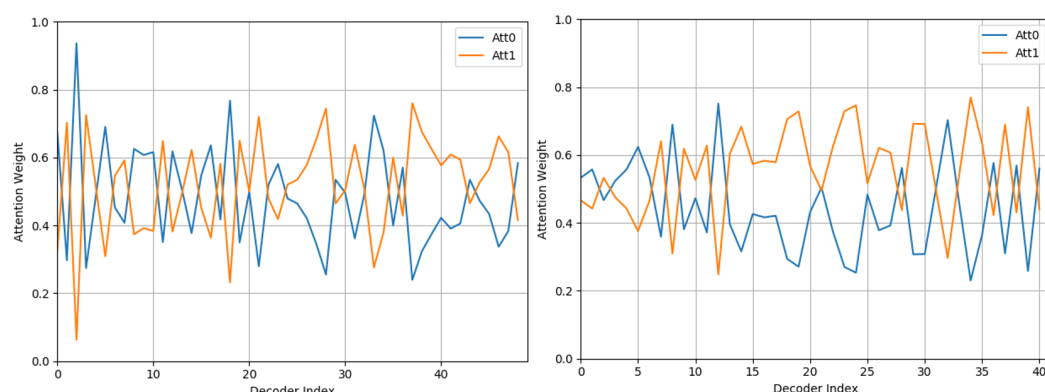
This section aims to evaluate the impact of the prior pronunciation auxiliary input on performance and to analyze in detail why this method is effective.

The last two rows of Table 3 present the performance of vowel reduction detection and overall phoneme recognition in the case of two different input types. One is to use the acoustic feature input alone as in the usual E2E speech recognition method. The other is to add pronunciation prior knowledge as an auxiliary input. Experiments show that, when the auxiliary input is added, vowel reduction detection and overall phoneme recognition have been improved. This improvement is more significant than with simply using an E2E method.

We randomly selected two sentences for decoding and analysis. Figure 3 shows the main and auxiliary encoders' weight change in the hierarchical attention mechanism during the decoding process. The horizontal axis is the decoding step, the vertical axis is the weight value, the yellow line (Att1) represents the auxiliary encoder, and the blue line (Att0) represents the main encoder. It can be seen that, when decoding phonemes contained in the prior articulation knowledge, the auxiliary encoder has a higher weight. When decoding phonemes that the prior articulation knowledge cannot provide enough information about, the main encoder that uses the acoustic feature input has a higher weight. We argue that the complementarity of these two types of information leads to the improvement. The input of the prior articulation knowledge improves the recognition rate while reducing the model's burden to recognize phonemes from acoustic features. It results in an improvement of the vowel reduction detection's performance.

We also observed the convergence of the frame-level attention mechanism during the training process. The experiment is consistent with the analysis in Section 2.3. If two encoders with the same complexity are designed, over-reliance on the pronunciation prior encoder will lead to poor convergence of the acoustic encoder's attention mechanism.





**Figure 3.** The changing weight of encoder-level attention.

#### 4.3. Analysis of Multi-Stream Expansion Method

We compared the model's performance after multi-stream expansion with the model's performance with the original main and auxiliary encoder structure through experiments, as shown in Table 4. In the performance of vowel reduction detection, the multi-stream model does not exceed the original stream model. When using G2P as the pronunciation prior, it does not even surpass the situation without adding the pronunciation prior input. The phone recognition performance at this time is not as good as the original streaming model. When using the letter sequence as the pronunciation prior and expanding the multi-stream, its performance at detecting weak vowels exceeds that without adding the pronunciation prior. However, we have observed that the most considerable significance of this method is that when the multi-stream expansion model is combined with a priori composed letter sequences, the phoneme recognition performance is improved a lot, which is the best phoneme recognition performance in our models. The reason this method is relatively poor when combined with the G2P input may be that G2P is obtained through a conversion of the letter sequence, and this process introduces errors. These errors make the relationship between it and the acoustic signal no longer direct, which is not conducive to the function of the attention mechanism.

**Table 4.** Performance of original streams and expanded multi-streams.

Method	$ih_{schwa}$		$ah_{schwa}$		PCR (%)	Time (s)
	Pre/Rec	F1	Pre/Rec	F1		
original-streams + G2P	0.64/0.67	0.66	0.62/0.55	0.58	81.8	244
original-streams + char	0.58/0.72	0.64	0.60/0.57	0.59	82.9	232
multi-streams + G2P	0.55/0.63	0.59	0.55/0.52	0.54	80.6	248
multi-streams + char	0.58/0.68	0.63	0.58/0.53	0.56	86.3	241

We also evaluated the time consumption of the multi-stream algorithm. We used 20 threads for CPU decoding under the same experimental configuration and test set. The average time consumption of each thread was calculated, as shown in the right column of Table 4. However, the execution time of the algorithm was affected by many factors such as calculation optimization, computer performance, system status, I/O load, and so on. We mainly focus on the relative significance of the execution time. The results show that the multi-stream algorithm increases time consumption. We suggest that the increased execution time comes from the calculation of the attention mechanism between different inputs. In view of the fact that our method does not add operations with significantly increased complexity, the increased time consumption is relatively small.

#### 4.4. Comparison of Different Auxiliary Input Types

Table 5 compares the influence of different input auxiliary sequence generation methods. Human trans means to directly use the training target sequence that does not contain

schwa labels after preprocessing. Compared with Table 3 in Section 4.1, it suggests that both automatic methods are effective. Moreover, the method that uses the letter sequence and the G2P method achieve a comparable performance. As for the overall phoneme recognition performance (PCR), the former method (82.9%) is higher than the latter. The reason could be that the reference text letter sequence is closer to the phoneme recognition target than the G2P generated sequence. The G2P method also did not show a great performance improvement after the introduction of stress information. A possible explanation for this is that the gap between the automatically generated sequence and the actual pronunciation introduces interference to the model.

**Table 5.** Performance of different auxiliary inputs.

Method	$ih_{schwa}$		$ah_{schwa}$	
	Pre/Rec	F1	Pre/Rec	F1
Human trans	0.80/0.82	0.810	0.82/0.82	0.820
G2P	0.64/0.67	0.655	0.62/0.55	0.583
Letter	0.58/0.72	0.642	0.60/0.57	0.585

Experiments using human transcription input are expected to explore this issue further. As shown in Table 5, a significant performance improvement can be observed when using human transcripts as an auxiliary. In other words, accurate context is constructive for vowel reduction detection. This conclusion is consistent with previous research conclusions in phonetics that the phenomenon of vowel reduction is affected by context.

## 5. Conclusions

This work proposes a pronunciation prior knowledge-assisted end-to-end vowel reduction detection method. Based on the CTC/Attention joint framework, a novel main and auxiliary encoder structure is designed to introduce auxiliary information generated from the reference text for vowel reduction detection. The end-to-end model adopts a hierarchical attention mechanism to fuse pronunciation prior information and the acoustic information dynamically. Furthermore, we study a multi-stream expansion method based on the attention mechanism, which uses the correlation between different inputs to enhance the encoded vector. Experiments show that the proposed method is better than the HMM-DNN hybrid method and the general end-to-end method in vowel reduction detection and overall phoneme recognition performance. Moreover, the multi-stream expansion method can effectively improve the phoneme recognition performance when combined with a priori composed letter sequences. The mechanism by which the auxiliary encoder information works and the impact of different automatically generated auxiliary inputs on performance have been further analyzed. In order to ensure the pronunciation prior contains less interference and provides more reliable information, a model that automatically predicts the stress and generates the pronunciation prior can be added in the future. In addition, acoustic features or front-end filters for the vowel reduction phenomenon are also worth considering.

**Author Contributions:** Methodology, validation and writing—original draft, Z.L. Review and formal analysis, Z.H. Review, editing, and formal analysis, L.W. Review and supervision, P.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is partially supported by the National Key Research and Development Program (No. 2020YFC2004100).

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

## References

- Li, K.; Mao, S.; Li, X.; Wu, Z.; Meng, H. Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. *Speech Commun.* **2018**, *96*, 28–36. [\[CrossRef\]](#)
- Lee, G.G.; Lee, H.Y.; Song, J.; Kim, B.; Kang, S.; Lee, J.; Hwang, H. Automatic sentence stress feedback for non-native English learners. *Comput. Speech Lang.* **2017**, *41*, 29–42. [\[CrossRef\]](#)
- Li, K.; Wu, X.; Meng, H. Intonation classification for L2 English speech using multi-distribution deep neural networks. *Comput. Speech Lang.* **2017**, *43*, 18–33. [\[CrossRef\]](#)
- Bang, J.; Lee, K.; Ryu, S.; Lee, G.G. Vowel-reduction feedback system for non-native learners of English. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 935–939.
- Lindblom, B. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.* **1963**, *35*, 1773–1781. [\[CrossRef\]](#)
- Zhang, Y.; Nissen, S.L.; Francis, A.L. Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *J. Acoust. Soc. Am.* **2008**, *123*, 4498–4513. [\[CrossRef\]](#) [\[PubMed\]](#)
- van Bergem, D.R. Acoustic and lexical vowel reduction. In Proceedings of the Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication, Barcelona, Spain, 30 September–2 October 1991.
- Flemming, E. *A Phonetically-Based Model of Phonological Vowel Reduction*; MIT Press: Cambridge, MA, USA, 2005.
- Burzio, L. Phonology and phonetics of English stress and vowel reduction. *Lang. Sci.* **2007**, *29*, 154–176. [\[CrossRef\]](#)
- Kuo, C.; Weismer, G. Vowel reduction across tasks for male speakers of American English. *J. Acoust. Soc. Am.* **2016**, *140*, 369–383. [\[CrossRef\]](#) [\[PubMed\]](#)
- Byers, E.; Yavas, M. Vowel reduction in word-final position by early and late Spanish-English bilinguals. *PLoS ONE* **2017**, *12*, e0175226. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fourakis, M. Tempo, stress, and vowel reduction in American English. *J. Acoust. Soc. Am.* **1991**, *90*, 1816–1827. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gonzalez-Dominguez, J.; Eustis, D.; Lopez-Moreno, I.; Senior, A.; Beaufays, F.; Moreno, P.J. A real-time end-to-end multilingual speech recognition architecture. *IEEE J. Sel. Top. Signal Process.* **2014**, *9*, 749–759. [\[CrossRef\]](#)
- Dhakal, P.; Damacharla, P.; Javaid, A.Y.; Devabhaktuni, V. A near real-time automatic speaker recognition architecture for voice-based user interface. *Mach. Learn. Knowl. Extract.* **2019**, *1*, 504–520. [\[CrossRef\]](#)
- Yang, C.H.H.; Qi, J.; Chen, S.Y.C.; Chen, P.Y.; Siniscalchi, S.M.; Ma, X.; Lee, C.H. Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6523–6527.
- Watanabe, S.; Hori, T.; Kim, S.; Hershey, J.R.; Hayashi, T. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1240–1253. [\[CrossRef\]](#)
- Lo, T.H.; Weng, S.Y.; Chang, H.J.; Chen, B. An Effective End-to-End Modeling Approach for Mispronunciation Detection. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Shanghai, China, 25–29 October 2020.
- Feng, Y.; Fu, G.; Chen, Q.; Chen, K. SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis. In Proceedings of the ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3492–3496.
- Zhang, Z.; Wang, Y.; Yang, J. Text-conditioned Transformer for automatic pronunciation error detection. *Speech Commun.* **2021**, *130*, 55–63. [\[CrossRef\]](#)
- Wang, X.; Li, R.; Mallidi, S.H.; Hori, T.; Watanabe, S.; Hermansky, H. Stream attention-based multi-array end-to-end speech recognition. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7105–7109.
- Li, R.; Wang, X.; Mallidi, S.H.; Watanabe, S.; Hori, T.; Hermansky, H. Multi-stream end-to-end speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *28*, 646–655. [\[CrossRef\]](#)
- Leung, W.K.; Liu, X.; Meng, H. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8132–8136.
- 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.
- Weide, R. *The CMU Pronunciation Dictionary*; Release 0.6; Carnegie Mellon University: Pittsburgh, PA, USA, 1998.
- Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-Based Models for Speech Recognition. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15), Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 577–585.
- Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1*; NASA STI/Recon Technical Report; NASA: Washington, DC, USA, 1993; Volume 93, p. 27403.

27. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Yalta Soplin, N.E.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 2207–2211.
28. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Big Island, HI, USA, 11–15 December 2011; pp. 1–4.
29. Yan, B.C.; Chen, B. End-to-End Mispronunciation Detection and Diagnosis From Raw Waveforms. *arXiv* **2021**, arXiv:2103.03023.
30. Fu, K.; Lin, J.; Ke, D.; Xie, Y.; Zhang, J.; Lin, B. A Full Text-Dependent End to End Mispronunciation Detection and Diagnosis with Easy Data Augmentation Techniques. *arXiv* **2021**, arXiv:2104.08428.
31. Jiang, S.W.F.; Yan, B.C.; Lo, T.H.; Chao, F.A.; Chen, B. Towards Robust Mispronunciation Detection and Diagnosis for L2 English Learners with Accent-Modulating Methods. *arXiv* **2021**, arXiv:2108.11627.