International Journal of Network Dynamics and Intelligence Scilight

Article

CCE-Net: Causal Convolution Embedding Network for Streaming Automatic Speech Recognition

Feiteng Deng, Yue Ming*, and Boyang Lyu

Beijing Key Laboratory of Work Safety and Intelligent Monitoring, School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
* Correspondence: yming@bupt.edu.cn

Received: 11 March 2024
Accepted: 15 August 2024
Published: 27 September 2024

Abstract: Streaming Automatic Speech Recognition (ASR) has gained significant attention across various application scenarios, including video conferencing, live sports events, and intelligent terminals. However, chunk division for current streaming speech recognition results in insufficient contextual information, thus weakening the ability of attention modeling and leading to a decrease in recognition accuracy. For Mandarin speech recognition, there is also a risk of splitting Chinese character phonemes into different chunks, which may lead to incorrect recognition of Chinese characters at chunk boundaries due to incomplete phonemes. To alleviate these problems, we propose a novel front-end network - Causal Convolution Embedding Network (CCE-Net). The network introduces a causal convolution embedding module to obtain richer historical context information, while capturing Chinese character phoneme information at chunk boundaries and feeding it to the current chunk. We conducted experiments on Aishell-1 and Aidatatang. The results showed that our method achieves a character error rate (CER) of 5.07% and 4.90%, respectively, without introducing any additional latency, showing competitive performances.

Keywords: streaming speech recognition; causal convolution; contextual information; phoneme segmentation

1. Introduction

With its real-time capabilities, streaming speech recognition has supplanted non-streaming speech recognition in numerous scenarios, including spanning video conferencing, live sports or game broadcasts, speech input methods, and terminal smart voice assistants. Non-streaming speech recognition requires waiting for the entire speech input before decoding and outputting, whereas streaming speech recognition can gradually output recognition results as the speech input progresses, without the need to wait for the entire speech input. Therefore, for tasks requiring real-time interaction with speech input, streaming speech recognition can offer a better user experience.

End-to-end models have become the mainstream in the field of speech recognition, including connectionist temporal classification (CTC) [1], recurrent neural network transducer (RNN-T) [2], and attention-based encoder-decoder (AED) [3, 4]. Among them, CTC and RNN-T models are natural streaming speech recognition models, but they exhibit significantly lower performance compared to attention-based models. Therefore, CTC and RNN-T models often introduce attention mechanisms to improve recognition performance, such as using Transformer [3] or Conformer [4] as encoders. However, traditional attention mechanisms belong to global attention mechanisms, requiring the entire speech input to be processed and lacking real-time capability. As a result, researchers have begun studying streaming attention mechanisms. Currently proposed streaming attention mechanisms mainly include monotonic attention [5−10], trigger attention [11, 12], and chunk-based attention mechanism [13−17]. The chunk-based attention exhibits the best recognition accuracy and latency performance.

The chunk-based attention mechanism divides the input into several fixed-size chunks and calculates attention independently within each chunk. One drawback of this method is that it imposes limitations on the coverage range of attention, leading to a potential lack of contextual information. To take into account historical contextual informa-

tion in attention computation, enhanced memory block methods [13−15] introduce a memory bank to store the embedding information of all previous chunks. Transformer XL [16] and U2++ [17] introduce cache mechanism to cache the output of previous chunks. These methods can effectively enhance the long-range history context, but may not model the local information well at the chunk boundaries.

Furthermore, chunk-based attention also encounters the issue of phoneme segmentation. When the phonemes corresponding to Chinese characters are split into different chunks at the chunk boundaries, incomplete phoneme information may lead to incorrect recognition of the Chinese characters, as indicated in Table 1. MiniStreamer [18] is designed to calculate attention while focusing on the current chunk and part of the historical chunks. Although it can alleviate the issue of phoneme segmentation at chunk boundaries, it introduces significant additional attention computation, resulting in increased latency. Tsunoo et al. [33] employed overlapping context chunks as input to the encoder, providing ample contextual information for the boundaries of the central chunk. However, this approach introduces a dependency on future context, leading to increased latency.

Table 1 Phoneme representation of the recognition results. The phoneme combination "jie" is segmented into "ji" and "ie" during chunk processing. The segmented phonemes are assigned to the left chunk and right chunk, respectively. The model recognized and outputted the former. The phoneme combination "dao" is segmented into "d" and "ao" during chunk processing. The segmented phonemes are assigned to the left chunk and right chunk, respectively. The cache mechanism is capable of capturing historical contextual information, but its limited local modeling capacity ultimately led to an erroneous recognition of the right chunk

chunk size	Recognition result (U2++)
Annotated text	ju wei jia an jie ti gong de shu ju xian shi
4	ju wei jia an ji ti gong de shu ju xian shi
16	ju wei jia an jie ti gong de shu ju xian shi
full	ju wei jia an jie ti gong de shu ju xian shi
Annotated text	zeng jia hua she cheng shi gong gong jiao tong you xian che dao
4	zeng jia hua she cheng shi gong gong jiao tong you xian che tao
16	zeng jia hua she cheng shi gong gong jiao tong you xian che dao
full	zeng jia hua she cheng shi gong gong jiao tong you xian che dao

In this paper, we propose a Causal Convolution Embedding Network (CCE-Net) for streaming speech recognition. Firstly, we introduce a causal convolution embedding module in CCE-Net, which utilizes causal convolution operations to extract complete phonetic speech features at chunk boundaries and introduces richer local historical context information. We inject the obtained information into each encoding chunk through embedding, avoiding information redundancy caused by high feature correlations. By employing causal convolution embedding, we effectively inject historical context information and phonetic features into each encoding chunk, enabling the streaming speech recognition model to utilize richer context-related information for acoustic modeling and thus improve recognition accuracy. Secondly, to meet the low-latency requirements of streaming speech recognition, we use depthwise separable convolution for downsampling to reduce computational costs. Our contributions can be summarized as follows:

1. We propose CCE-Net, a front-end network designed to supplement historical local contextual information for encoding chunks in streaming speech recognition, enhancing recognition accuracy without introducing additional waiting delay.

2. We introduce a novel causal convolutional embedding module that utilizes causal convolution to extract complete phonetic features at the boundaries of Chinese characters and embeds them to the left boundary of the current chunk, reducing recognition errors caused by phonetic segmentation.

3. Experiments on the AISHELL-1 and Aidatatang datasets demonstrate that our method achieves character error rates of 5.07% and 4.90%, respectively, without introducing any additional latency, showcasing competitive performance.

2. Related Work

In this section, we first presented the current research status of streaming speech recognition and the streaming attention mechanism. We then introduced the streaming chunk strategy and related optimization for mainstream streaming attention mechanisms. Finally, to address the issue of insufficient contextual information, we provided a detailed overview of the causal convolutional neural network.

2.1. Streaming ASR

In end-to-end speech recognition models, CTC and RNN-T are naturally suited for streaming speech recogni-

tion, while attention-based models are known for their superior performance in non-streaming speech recognition. To improve recognition accuracy, CTC and RNN-T are combined with attention mechanisms, resulting in hybrid CTC/attention and Conformer-Transducer models [19–21] that can perform streaming recognition and bridge the gap between streaming and non-streaming speech recognition.

In order to ensure real-time processing and meet low-latency requirements, the attention mechanism needs to be improved for streaming. The streaming attention mechanism is mainly divided into three categories: monotonic attention, trigger attention, and chunk-based attention mechanism. Monotonic attention mechanisms, including Monotonic Chunkwise Attention (MoChA) [6] and improved versions [7, 8], and Monotonic truncated attention (MTA) [9, 10] , achieve linear complexity and real-time decoding. However, there is a problem of weak model generalization ability due to the large difference between training and inference. Trigger attention [11, 12] uses a CTC-trained neural network to dynamically partition the input sequence. The disadvantage of this method is that it cannot guarantee the accuracy of CTC boundary partitioning and has high computational complexity. The chunk-based attention [13–17] is currently the mainstream mechanism, which achieves real-time processing through a streaming chunk strategy, offering lower computational complexity and latency.

2.2. Streaming Chunk Strategy

The streaming chunk strategy [13–17] is at the core of chunk-based attention, dividing the input speech stream into several chunks and performing attention computation within each chunk to achieve streaming transcription. However, it limits attention within the chunk, which may lead to a lack of contextual information and possible phoneme segmentation at the chunk boundary, making it difficult to model local information at the chunk boundaries. To enhance the local modeling ability of attention within the chunk, it is usually necessary to increase the chunk length and add context chunks. However, this increases attention computation and introduces additional waiting delay. To avoid introducing actual future context into the model and introducing additional waiting delay, Chunking, Simulating Future Context and Decoding (CUSIDE) [24] introduces a simulation module that recursively simulates future context frames, thereby injecting virtual future context information into the current chunk. Zhao [25] proposes a multi-delay speech recognition with zeor lookahead that achieves recognition accuracy close to that with lookahead. Strimel proposes an adaptive non-causal attention transducer (ANCAT) [26], which is trained to adaptively acquire future context while considering the impact of accuracy and delay. The drawback of these methods is that they require additional attention computation, leading to a significant increase in computational costs. Additionally, ANCAT still introduces some future context, and none of these methods specifically address the issue of phoneme segmentation in Chinese characters at the boundaries of the chunks.

2.3. Causal Convolutional Neural Network

Convolutional neural network (CNN) have been widely used in speech recognition models to model local information. With the rise of streaming speech recognition, ordinary CNN are no longer applicable because they cannot guarantee the causality of input data in the time dimension. Causal convolution is a type of convolution operation that only considers historical information. It applies the convolution kernel only to the current frame and a fixed length of historical frames, without introducing any dependence on future frames in streaming speech recognition. Causal convolution has been widely used in speech recognition tasks [27, 28] as a replacement for traditional CNN. In this paper, we leverage causal convolution embedding to introduce a more extensive range of historical context information for speech chunks. This augmentation significantly enhances the modeling capability of local dependency relations at the boundaries of these chunks. Furthermore, this approach enables us to effectively capture the phonetic feature information that may be missing at the left boundary of the chunks associated with Chinese characters.

3. Method

In this Section, we will first introduce the overall model architecture of the streaming speech recognition in Section 1. Then, we will introduce the proposed front-end network architecture and explain the introduced causal convolution embedding.

3.1. Model Architecture

The proposed Causal Convolution Embedding Network (CCE-Net) is situated at the front-end of the streaming speech recognition, as illustrated in Figure 1. CCE-Net takes an 80-dimensional FBank feature concatenated spectrogram as input, encompassing a 2D convolutional downsampling module, a causal convolution embedding module for extracting historical chunk context information and phoneme features, and a linear layer for feature mapping. The feature spectrogram processed by CCE-Net is fed into the shared encoder to complete acoustic modeling. Finally,

The CTC decoder outputs the streaming recognition results, while the attention decoders rescore the CTC decoding results to generate more accurate recognition text.

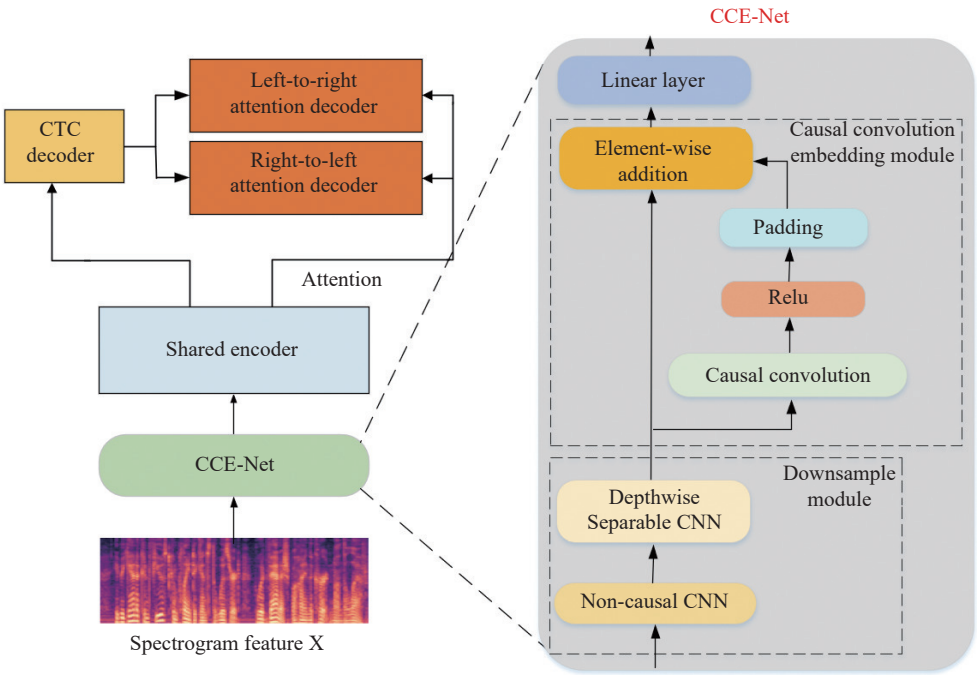

Figure 1. Overall framework diagram.

3.2. CCE-Net

This section will introduce the specific processing flow of CCE-Net, as shown in Figure 2.

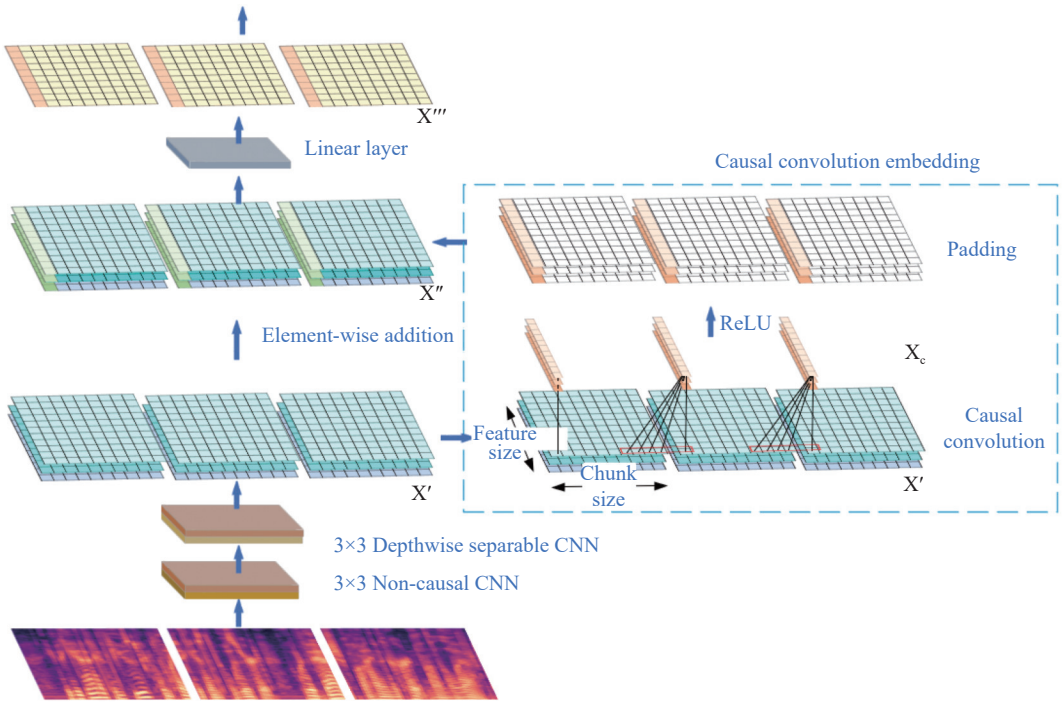

Figure 2. Acoustic features processed by the CCE-Net.

For the front-end network that applies the method proposed in this paper, the spectrogram feature $\mathbf{X} = (x_1, x_2, ..., x_L)$ is first fed into the downsampling layer, and is then processed by two convolutional neural networks with downsampling factors of 2, resulting in:

$$\mathbf{X}' = \text{Downsampling}(\mathbf{X}) \tag{1}$$

Then, it is processed by a causal convolutional neural network:

$$\mathbf{X}_c = \text{CausalConv}(\mathbf{X}') \tag{2}$$

since the baseline model U2++ [17] uses the dynamic chunk training method, a specific chunk size can be selected according to actual needs, and the smaller the chunk size, the lower the delay. We use causal convolution to extract historical local features as embedding, avoiding dependence on future frames.

The output of the causal convolution is split into $l/chunksize$ feature vectors, and then added to the first feature vector of each chunk in $\mathbf{X}'$, resulting in $\mathbf{X}'' = (x_1'', x_2'', \ldots, x_l'')$. Then, it is processed by a linear layer, resulting in:

$$\mathbf{X}'' = \text{Linear}(\mathbf{X}'') \tag{3}$$

Finally, the feature sequence $\mathbf{X}''$ is fed into the encoder layer by chunks.

3.2.1. Downsampling Module

As a front-end network for speech recognition, CCE-Net will downsample the network to reduce the size of the spectrogram feature, while also reducing computation and memory usage, and extracting local speech features. In the downsampling module, we use a 3 × 3 ordinary convolutional neural network as the first layer, with a stride of 2 and 512 channels to capture advanced speech features at different levels. Given the substantial number of input and output channels, we employ a 3 × 3 depthwise separable convolutional neural network as the second layer to amalgamate and integrate features from diverse channels via pointwise convolution. This enables the acquisition of spectral and phonetic characteristics within the speech signal. Then, we employ two one-dimensional depthwise separable convolutional neural networks with a kernel size of 3 to extract deeper features, enhancing the model's representational capacity.

3.2.2. Causal Convolution Embedding Module

During the encoding stage, we segment the speech feature sequence into equally-sized, non-overlapping chunks, and the streaming speech recognition delivers recognition results for each chunk. To address the issue of lacking contextual information due to chunking, as well as the incomplete phoneme information at chunk boundaries, we propose a causal convolution embedding module. This module comprises a one-dimensional causal convolution branch, as depicted in Figure 2, where we set the stride of the convolution operation to the chunk size. Each chunk undergoes a causal convolution operation to extract the historical context information and complete phoneme features at the left boundary of the chunk. The duration of the pronunciation of a Chinese character typically ranges from 150 milliseconds to 400 milliseconds. In order to cover all phonemes at chunk boundaries, this module is designed to perform convolution operations on the first frame of the chunk and the last 8 frames of the previous chunk, corresponding to a 385-millisecond duration of the speech. The output after causal convolution operation can be represented as:

$$x_c[t] = \sum_{m=0}^{8} w[m] \cdot x'[t-m] + b \tag{4}$$

where $t = \text{chunksize} \times m, m = 0, 1, 2, \ldots, T/chunksize$, $w[m]$ is the weight value of the convolutional kernel, b is the bias value, $x'[t-m]$ is the output of the downsampling module at frame t-m. Subsequently, the model's nonlinear fitting capability is enhanced through activation functions while retaining key feature information in the speech signal. Then, zero padding is applied in the time domain to ensure consistency with the size of the original feature map, resulting in the final causal convolutional embedding tensor:

$$\mathbf{X}_{\text{cce}} = Padding\left(\mathbf{Zero}, \mathbf{X}_c'\right) \tag{5}$$

where **Zero** represents a zero tensor of length $chunksize - 1$, $\mathbf{X}_c'$ is the output of $\mathbf{X}_c$ after passing through the activation function. Finally, the causal convolutional embeddings and the first frame of each chunk are weighted and summed to inject historical information and boundary phonetic feature information. The first feature of each chunk can be represented as:

$$x_i'' = x_i' + k x_{cj}' \tag{6}$$

where $j = i/\text{chunksize}$, k is the weight coefficient of the causal convolution embedding.

4. Experiment

In this section, we will introduce the experimental configurations and settings, followed by presenting our abla-

tion experiments, model analysis experiments, and comparative experimental results, and analyze them in detail.

4.1. Experiment settings

The experiments were conducted on a Linux server with the Linux Ubuntu (16.04) operating system. The CPU used was Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz, and the GPU was NVIDIA GeForce GTX 3090. The GPU driver version was 460.84, and the CUDA version was 11.2. PyTorch was used as the deep learning framework for the end-to-end streaming speech recognition model. We conducted experiments using the wenet speech recognition toolkit and used its U2++ model as our baseline model. The ASR input is 80-dimensional filterbank feature (FBank [26]). To augment the data, we used three-speed perturbation (0.9, 1.0 or 1.1), with the same configuration of SpecAugment [30] and SpecSub [17] as U2++. For the encoder, we used Conformer provided by the wenet toolkit as the encoder. The number of layers in the Conformer encoder is 16, and the number of attention heads is 4. The attention decoder contains 3 left-to-right and 3 right-to-left bitransformer chunks. We evaluated the performance of the proposed method under two decoding methods, ctc prefix beam search and attention rescoring. The weight of CTC is 0.5 and the weight of reverse is 0.3 when using attention rescoring decoding. We used the Adam optimizer, and the learning rate was warmed up for 25000 steps. For the causal convolutional neural network, we set the kernel size to 9 and the stride to 16, consistent with the default chunk length. We employed a weighted averaging method to reduce model variance, enhance model robustness, and decrease prediction uncertainty.

We evaluated our proposed method on two Chinese Mandarin speech corpora, AISHELL-1 and Aidatatang. We employed a chunk-based streaming speech recognition model, allowing us to use the duration of each chunk as the metric for measuring latency. When the chunk size is 16, the latency is 640 milliseconds, and when the chunk size is 4, the latency is 160 milliseconds.

4.1.1. Ablation study

Table 2 shows the parameters, character error rate (CER) and Real-Time Factor (RTF) of two-pass decoding at different delays before and after the incorporation of CCE-Net. RTF represents the ratio of the system's processing time for a speech signal to the actual duration of the speech. A smaller value indicates better real-time performance. The difference between our method and the baseline model lies in the addition of CCE-Net at the front end of the baseline model's encoder to optimize the streaming chunk strategy. This network can provide richer historical context information for the encoding chunks, enhance the local modeling capability at chunk edges. Therefore, compared to the baseline model, recognition accuracy has been further improved.

Table 2 Number of parameters and CER compared to the baseline model. The number in parentheses represents the chunk size, which represents the delay. The CER results are obtained using CTC prefix beam search decoding and attention-based rescoring, respectively

Model	params	delay(ms)	CER		RTF	
			attention rescore	CTC prefix beam search	attention rescore	CTC prefix beam search
U2++	48.3 M	640	5.11	6.04	0.036	0.033
		160	5.44	6.57	0.040	0.036
U2++ & CCE-Net	48.5 M	640	5.07	5.85	0.037	0.033
		160	5.29	6.33	0.042	0.037

Table 2 presents the results of the proposed method and the baseline model under CTC prefix beam search decoding and attention-based rescoring decoding at different chunk sizes. CTC prefix beam search decoding enables streaming recognition, while attention-based rescoring decoding requires more accurate results based on global speech output after the first round of decoding. It is evident from the Table 2 that using CCE-Net as the front-end network results in a negligible increase in parameter count, not exceeding 2%. Under chunk delays of 640ms and 160ms, our method achieves streaming recognition accuracies of 5.85% and 6.33%, representing reductions in CER of 3.15% and 3.65%, respectively, compared to the U2++ model. This indicates that the proposed CCE-Net effectively extracts historical context information for the current chunk, improving the accuracy of acoustic modeling and hence enhancing streaming recognition accuracy. Due to the improved performance of CTC decoding, the second round of rescoring decoding also demonstrates enhanced accuracy, with CER reaching 5.07% and 5.29% under chunk delays of 640ms and 160ms, respectively, both lower than the U2++ [17] model's CER. RTF represents the ratio of the system's processing time for a speech signal to the actual duration of the speech. A smaller value indicates better real-time performance. As shown in Table 2, our method's RTF is very close to the baseline, indicating that under similar latency and real-time conditions, our method achieves higher recognition accuracy.

Injecting contextual information into encoding chunks effectively is a key issue. We once directly added a causal convolutional neural network after the downsampling layer, but the performance improvement was not significant, especially when the chunk size was small. Considering that high inter-frame correlation may lead to redundant

feature information, reducing feature quality and resulting in insignificant performance improvement, we adopted an embedding approach to inject the feature information extracted by causal convolution.

Table 3 shows the comparison of the models' word error rates between directly adding a causal convolutional layer and using causal convolutional embedding. In CCE-Net v0, we directly added causal convolution after the downsampling network to obtain historical contextual information, which slightly improved the accuracy of online recognition. However, as the chunk size decreased, the improvement effect became worse. When we used the embedding method to add the causal convolutional network, the model's online recognition accuracy was higher than that of directly adding causal convolution, and the relative CER reduction was more significant as the chunk size decreased. This indicates that using the embedding method to obtain historical contextual information can more effectively model acoustic features.

Table 3 Comparison of CER between CCE-Net v0, the current version, and the baseline model with different chunk sizes

Model	CER(16)	CER(4)
U2++	6.04	6.57
U2++ & CCE-Net v0	5.94	6.52
U2++ & CCE-Net	5.85	6.33

To explore the optimal weight parameters for causal convolution embedding, we conducted experiments with different weight values *k*. We conducted experiments by setting the weight parameter *k* to 2.0, 1.5, 1.2, 1.0 and 0.8, and observed the results as shown in Figure 3. When the weight parameter is 1.2, the recognition performance of small chunks (chunk size = 4) is slightly better. Conversely, with a weight parameter of 0.8, the recognition performance of large chunks (chunk size = 16) shows a slight improvement. Larger weight parameters can lead to decreased performance, even worse than the baseline model. Considering the trade-offs, setting the weight parameter to 0.8 seems to be the optimal choice.

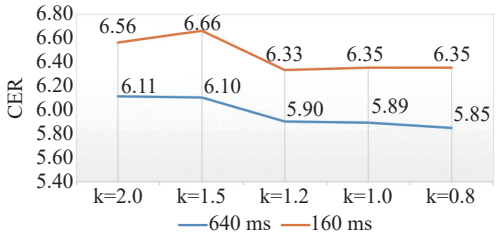

Figure 3. CER of the model with different weight parameter *k*.

4.1.2. Comparative Experimental Results

Table 4 and Table 5 shows the delay and Character Error Rate (CER) of various representative models on the AISHELL-1 and Aidatatang. The comparison was made between the results of streaming output without rescoring and language modeling, and the second output results that included the addition of rescoring or language modeling. The chunk size was set to 16, which is equivalent to a delay of 640ms. It can be observed that our method achieves lower CER results under similar chunk delay conditions compared to other representative models. Existing representative models often fail to provide sufficient and effective contextual information for encoding chunks. Although they can calculate the correlation between the current frame and historical frames during attention computation, this method lacks the ability to model local dependencies. In contrast, our approach utilizes causal convolutional neural networks to model local correlations at chunk boundaries, injecting historical context information for each chunk through embedding. This enables the capture of missing phonetic features and leads to higher recognition accuracy.

Table 4 The CER of various representative models on AISHELL-1

Model	Delay(ms)	CER(16)
Streaming Transformer[32]	640	12.2
Streaming Conformer[33]	640	6.8
RNN-Transducer	640	8.7
MMA[10]	640	6.60
U2[29]	640	6.30
U2++[17]	640	6.04
U2++ & rescore[17]	640 + Δ	5.11
WNARS rescoring[34]	640 + Δ	5.22
U2++ & CCE-Net	640	5.85
U2++ rescore & CCE-Net	640 + Δ	**5.07**

Table 5 The CER of various representative models on Aidatatang

Model	Delay(ms)	CER(16)
Streaming Transformer[32]	640	7.8
Streaming Conformer[33]	640	6.4
U2 [29]	640 + Δ	5.29
U2++[17]	640	6.17
U2++ & rescore[17]	640 + Δ	4.99
U2++ & CCE-Net	640	6.01
U2++ rescore & CCE-Net	640 + Δ	**4.90**

4.1.3. Improvement in Chinese character phoneme segmentation

Table 6 shows the pinyin of three recognition error examples caused by chunk processing and the pinyin of the recognition results after applying causal convolutional embedding.

Table 6 Different phoneme recognition results of two speech segments through the U2++ model [17] and the model using causal convolution embedding method under different chunk size settings

chunk size	Recognition result (U2++)	Recognition result (U2++ & CCE-Net)
text	ju wei jia an jie ti gong de shu ju xian shi	ju wei jia an jie ti gong de shu ju xian shi
4	ju wei jia an ji ti gong de shu ju xian shi	ju wei jia an jie ti gong de shu ju xian shi
16	ju wei jia an jie ti gong de shu ju xian shi	ju wei jia an jie ti gong de shu ju xian shi
full	ju wei jia an jie ti gong de shu ju xian shi	ju wei jia an jie ti gong de shu ju xian shi
text	zeng jia hua she cheng shi gong gong jiao tong you xian che dao	zeng jia hua she cheng shi gong gong jiao tong you xian che dao
4	zeng jia hua she cheng shi gong gong jiao tong you xian che tao	zeng jia hua she cheng shi gong gong jiao tong you xian che dao
16	zeng jia hua she cheng shi gong gong jiao tong you xian che dao	zeng jia hua she cheng shi gong gong jiao tong you xian che dao
full	zeng jia hua she cheng shi gong gong jiao tong you xian che dao	zeng jia hua she cheng shi gong gong jiao tong you xian che dao
text	zhu ban fang shi tu ji lu mei yi ge pao you hui sa ji qing de mei yi ge shun jian	zhu ban fang shi tu ji lu mei yi ge pao you hui sa ji qing de mei yi ge shun jian
4	zhu ban fang shi tu ji lu mei yi ge pao you hui cao ji qi de mei yi ge shun jian	zhu ban fang shi tu ji lu mei yi ge pao you hui sao ji qing de mei yi ge shun jian
16	zhu ban fang shi tu ji lu mei yi ge pao you hui cao ji qi de mei yi ge shun jian	zhu ban fang shi tu ji lu mei yi ge pao you hui sao ji qing de mei yi ge shun jian
full	zhu ban fang shi tu ji lu mei yi ge pao you hui sao ji qing de mei yi ge shun jian	zhu ban fang shi tu ji lu mei yi ge pao you hui sao ji qing de mei yi ge shun jian

For the first example, when the chunk size is 4, the chunk processing separates the phoneme combination "jie" corresponding to a Chinese character into "ji" and "ie" and assigns them to two different chunks, and even with the cache mechanism of U2++ [17], it is not recognized correctly. Through analysis of the annotated text corresponding to the speech, it was found that the phoneme combination "jie" was segmented into the 7th and 8th chunks. However, when the chunk size is 16, the phoneme combination "jie" is fully assigned to the 2nd chunk and recognized correctly without phoneme segmentation. For the second example, the phoneme combination "dao" is also segmented into "da" and "ao" and assigned to different chunks, resulting in recognition error due to incomplete phoneme information. When the chunk size is 16 or when no chunking is performed, there is no phoneme segmentation and the character is recognized correctly. Our method adds causal convolutional embedding to obtain more contextual information for the chunks, enabling the model to recognize them correctly when the chunk size is 4. For the third example, when the chunk size is 4 and 16, the phoneme combination "qing" is segmented into "qi" and "ing" and assigned to different chunks, with the phonemes in the left chunk being recognized and outputted. However, after obtaining complete phonemes through causal convolution, they are correctly recognized in the right chunk.

5. Conclusions

This paper introduces a front-end network based on causal convolution embedding - CCE-Net. The experimental results show that using CCE-Net as the front-end network can improve the accuracy of the streaming speech recognition model without introducing additional delay, and the parameter overhead can be neglected. The experiments on the AISHELL-1 and Aidatatang datasets showed that the character error rate of streaming decoding decreased by 3.3% and 2.6% respectively after the application of CCE-Net, with better improvement as the chunk size become smaller. Benefiting from the improved accuracy in streaming decoding, there has been a corresponding reduction in the character error rate of re-scoring decoding.We used the U2++ framework [17] as the baseline for

almost all experiments, but theoretically, our method can be applied to all chunk-based streaming speech recognition models. In the future, we will apply CCE-Net to more models to verify the generalization of our method. As our method only introduces richer historical context information and future context information is equally important, in the future, we can combine simulated future context methods to introduce future context information for encoding chunks. Additionally, we can also introduce a fast attention mechanism to reduce the memory and time overhead of attention calculation. Through these methods, we believe that we can achieve improvements in both accuracy and speed.

Author Contributions: **Deng Feiteng and Ming Yue:** Conceptualization, Methodology; **Deng Feiteng and Ming Yue:** Data curation, Experimental verification, Writing-Original draft preparation; **Ming Yue and Boyang Lyu:** Supervision, Writing-Reviewing and Editing.

Funding: Natural Science Foundation of China (Grant No. 62076030), Beijing Natural Science Foundation (Grant No. L241011), and basic research fees of Beijing University of Posts and Telecommunications (Grant No. 2023ZCJH08).

Data Availability Statement: The data used in this paper are all from public datasets.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Graves, A.; Fernández, S.; Gomez, F.; *et al*. 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*; ACM: New York, 2006; pp 369–376. doi:10.1145/1143844.1143891
2. Cui, X.D.; Saon, G.; Kingsbury, B. Improving RNN transducer acoustic models for English conversational speech recognition. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association, Dublin, Ireland, 20–24 August* 2023; ISCA: Dublin, Ireland, 2023; pp. 1299–1303.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; *et al*. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, 2017; pp. 6000–6010.
4. Gulati, A.; Qin, J.; Chiu, C.C.; *et al*. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020*; ISCA: Shanghai China, 2020; pp. 5036–5040.
5. Zeyer, A.; Schmitt, R.; Zhou, W.; *et al*. Monotonic segmental attention for automatic speech recognition. In *Proceedings of 2022 IEEE Spoken Language Technology Workshop* (*SLT), Doha, Qatar, 9–12 January 2023*, IEEE: New York, 2023; pp. 229–236. doi:10.1109/SLT54892.2023.10022818
6. Chiu, C.C.; Raffel, C. Monotonic chunkwise attention. In *Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018*; ICLR: Vancouver, Canada,2018.
7. Tsunoo, E.; Kashiwagi, Y.; Kumakura, T.; *et al*. Towards online end-to-end transformer automatic speech recognition. arXiv: 1910.11871, 2019. doi:10.48550/arXiv.1910.11871
8. Inaguma, H.; Mimura, M.; Kawahara, T. Enhancing monotonic multihead attention for streaming ASR. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020*; ISCA: Shanghai China, 2020; pp. 2137–2141.
9. Miao, H.R.; Cheng, G.F.; Zhang, P.Y.; *et al*. Online hybrid CTC/attention end-to-end automatic speech recognition architecture. IEEE/ACM Trans. Audio Speech Lang. Process., **2020**, *28*: 1452−1465. doi: 10.1109/TASLP.2020.2987752
10. Miao, H.R.; Cheng, G.F.; Gao, C.F.; *et al*. Transformer-based online CTC/attention end-to-end speech recognition architecture. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Barcelona, Spain, 4–8 May 2020*; IEEE: New York, 2020; pp. 6084–6088. doi:10.1109/ICASSP40776.2020.9053165
11. Moritz, N.; Hori, T.; Le, J. Streaming automatic speech recognition with the transformer model. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Barcelona, Spain, 4–8 May 2020*; IEEE: New York, 2020; pp. 6074–6078. doi:10.1109/ICASSP40776.2020.9054476
12. Zhao, H.B.; Higuchi, Y.; Ogawa, T.; *et al*. An investigation of enhancing CTC model for triggered attention-based streaming ASR. In *Proceedings of 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC), Tokyo, Japan, 14–17 December 2021*; IEEE: New York, 2021; pp. 477–483.
13. Wu, C.Y.; Wang, Y.Q.; Shi, Y.Y.; *et al*. Streaming transformer-based acoustic models using self-attention with augmented memory. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020*; ISCA: Shanghai China, 2020; pp. 2132–2136.
14. Shi, Y.Y.; Wang, Y.Q.; Wu, C.Y.; *et al*. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. In *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Toronto, ON, Canada, 6–11 June 2021*; IEEE: New York, 2021; pp. 6783–6787. doi:10.1109/ICASSP39728.2021.9414560
15. Wang, F.Y.; Xu, B. Shifted chunk encoder for transformer based streaming end-to-end ASR. In *Proceedings of the 29th International Conference on Neural Information Processing, IIT Indore, India, 22–26 November 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 39–50. doi:10.1007/978-981-99-1642-9_4
16. Dai, Z.H.; Yang, Z.L.; Yang, Y.M.; *et al*. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceed-*

ings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; ACL: Stroudsburg, USA, 2019; pp. 2978–2988. doi:10.18653/v1/P19-1285

17. Zhang, B.B.; Wu, D.; Peng, Z.D.; *et al*. WeNet 2.0: More productive end-to-end speech recognition toolkit. In *Proceedings of 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18–22 September 2022*; ISCA: Incheon, Korea, 2022; pp. 1661–1665.
18. Gulzar, H.; Busto, M.R.; Eda, T.; *et al*. miniStreamer: Enhancing small conformer with chunked-context masking for streaming ASR applications on the edge. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association, Dublin, Ireland, 20–24 August 2023*; ISCA: Dublin, Ireland, 2023; pp. 3277–3281.
19. Zhang, Q.; Lu, H.; Sak, H.; *et al*. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Barcelona, Spain, 4–8 May 2020*; IEEE: New York, 2020; pp. 7829–7833. doi:10.1109/ICASSP40776.2020.9053896
20. Shi, Y.Y.; Wu, C.Y.; Wang, D.L.; *et al*. Streaming transformer transducer based speech recognition using non-causal convolution. In *Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Singapore, Singapore, 23–27 May 2022*; IEEE: New York, 2022; pp. 8277–8281. doi:10.1109/ICASSP43922.2022.9747706
21. Swietojanski, P.; Braun, S.; Can, D.; *et al*. Variable attention masking for configurable transformer transducer speech recognition. In *Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Rhodes Island, Greece, 4–10 June 2023*; IEEE: New York, 2023; pp. 1–5. doi:10.1109/ICASSP49357.2023.10094588
22. Hu, K.; Sainath, T.N.; Pang, R.M.; *et al*. Deliberation model based two-pass end-to-end speech recognition. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Barcelona, Spain, 4–8 May 2020*; IEEE: New York, 2020; pp. 7799–7803. doi:10.1109/ICASSP40776.2020.9053606
23. Hu, K.; Pang, R.M.; Sainath, T.N.; *et al*. Transformer based deliberation for two-pass speech recognition. In *Proceedings of 2021 IEEE Spoken Language Technology Workshop* (*SLT), Shenzhen, China, 19–22 January 2021*; IEEE: New York, 2021; pp. 68–74. doi:10.1109/SLT48900.2021.9383497
24. An, K.Y.; Zheng, H.H.; Ou, Z.J.; *et al*. CUSIDE: Chunking, simulating future context and decoding for streaming ASR. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18–22 September 2022*; ISCA: Incheon, Korea, 2022; pp. 2103–2107.
25. Zhao, H.B.; Fujie, S.; Ogawa, T.; *et al*. Conversation-oriented ASR with multi-look-ahead CBS architecture. In *Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Rhodes Island, Greece, 4–10 June 2023*; IEEE: New York, 2023; pp. 1–5. doi:10.1109/ICASSP49357.2023.10094614
26. Strimel, G.; Xie, Y.; King, B.J.; *et al*. Lookahead when it matters: Adaptive non-causal transformers for streaming neural transducers. In *Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023*; PMLR: New York, USA, 2023; pp. 32654–32676.
27. Audhkhasi, K.; Farris, B.; Ramabhadran, B.; *et al*. Modular conformer training for flexible End-to-End ASR. In *Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Rhodes Island, Greece, 4–10 June 2023*; IEEE: New York, USA, 2023; pp. 1–5. doi:10.1109/ICASSP49357.2023.10095966
28. Boyer, F.; Shinohara, Y.; Ishii, T.; *et al*. A study of transducer based end-to-end ASR with ESPnet: Architecture, auxiliary loss and decoding strategies. In *Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU), Cartagena, Colombia, 13–17 December 2021*; IEEE: New York, USA, 2021; pp. 16–23. doi:10.1109/ASRU51503.2021.9688251
29. Yao, Z.Y.; Wu, D.; Wang, X.; *et al*. WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August–3 September 2021*; ISCA: Brno, Czechia, 2021; pp. 4054–4058.
30. Park, D.S.; Chan, W.; Zhang, Y.; *et al*. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September, 2019*; ISCA: Graz, Austria, 2019; pp. 2613–2617.
31. Burchi, M.; Vielzeuf, V. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. In *Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU), Cartagena, Colombia, 13–17 December 2021*; IEEE: New York, USA, 2021; pp. 8–15. doi:10.1109/ASRU51503.2021.9687874
32. Guo, P.C.; Boyer, F.; Chang, X.K.; *et al*. Recent developments on ESPnet toolkit boosted by conformer. In *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP), Toronto, ON, Canada, 6–11 June 2021*; IEEE: New York, USA, 2021; pp. 5874–5878. doi:10.1109/ICASSP39728.2021.9414858
33. Tsunoo, E.; Kashiwagi, Y.; Watanabe, S. Streaming transformer Asr with blockwise synchronous beam search. In *Proceedings of 2021 IEEE Spoken Language Technology Workshop* (*SLT), Shenzhen, China, 19–22 January 2021*; IEEE: New York, USA, 2021; pp. 22–29. doi:10.1109/SLT48900.2021.9383517
34. Wang, Z.C.; Yang, W.W.; Zhou, P.; *et al*. WNARS: WFST based non-autoregressive streaming end-to-end speech recognition. arXiv: 2104.03587, 2021. doi:10.48550/arXiv.2104.03587

Citation: Deng, F; Ming, Y; Lyu, B. CCE-Net: Causal Convolution Embedding Network for Streaming Automatic Speech Recognition. *International Journal of Network Dynamics and Intelligence*. 2024, 3(3), 100019. doi: 10.53941/ijndi.2024.100019