applied sciences

MDPI

Article

Improving Hybrid CTC/Attention Architecture with Time-Restricted Self-Attention CTC for End-to-End Speech Recognition

Long Wu [1,2], Ta Li [1,*], Li Wang [1] and Yonghong Yan [1,2,3]

[1] Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China; wulong@hccl.ioa.ac.cn (L.W.); wangli@hccl.ioa.ac.cn (L.W.); yanyonghong@hccl.ioa.ac.cn (Y.Y.)

[2] University of Chinese Academy of Sciences, Beijing 100190, China

[3] Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumchi 830001, China

* Correspondence: lita@hccl.ioa.ac.cn

Received: 13 September 2019; Accepted: 22 October 2019; Published: 31 October 2019

check for updates

Abstract: As demonstrated in hybrid connectionist temporal classification (CTC)/Attention architecture, joint training with a CTC objective is very effective to solve the misalignment problem existing in the attention-based end-to-end automatic speech recognition (ASR) framework. However, the CTC output relies only on the current input, which leads to the hard alignment issue. To address this problem, this paper proposes the time-restricted attention CTC/Attention architecture, which integrates an attention mechanism with the CTC branch. "Time-restricted" means that the attention mechanism is conducted on a limited window of frames to the left and right. In this study, we first explore time-restricted location-aware attention CTC/Attention, establishing the proper time-restricted attention window size. Inspired by the success of self-attention in machine translation, we further introduce the time-restricted self-attention CTC/Attention that can better model the long-range dependencies among the frames. Experiments with wall street journal (WSJ), augmented multiparty interaction (AMI), and switchboard (SWBD) tasks demonstrate the effectiveness of the proposed time-restricted self-attention CTC/Attention. Finally, to explore the robustness of this method to noise and reverberation, we join a train neural beamformer frontend with the time-restricted attention CTC/Attention ASR backend in the CHIME-4 dataset. The reduction of word error rate (WER) and the increase of perceptual evaluation of speech quality (PESQ) approve the effectiveness of this framework.

Keywords: automatic speech recognition; end-to-end; CTC; self-attention; hybrid CTC/attention

1. Introduction

Automatic speech recognition is an essential technology for realizing natural human–machine interfaces. A typical ASR system is factorized into several modules including acoustic, lexicon, and language models based on a probabilistic noisy channel model. However, the current algorithm leans heavily on the scaffolding of complicated legacy architectures that grew up around traditional techniques. In the last few years, an emerging trend in ASR is the study of end-to-end (E2E) systems [1–8]. An E2E ASR system directly transduces an input sequence of acoustic features x to an output sequence of probabilities of tokens (phonemes, characters, words, etc.) y. Three widely used contemporary E2E approaches are: (a) CTC [9,10], (b) Attention-based Encoder–Decoder (Attention ED) [11,12]. (c) Recurrent neural network (RNN) Transducer (RNN-T) [13,14].

Among the three aforementioned E2E methods, CTC enjoys its training simplicity and is one of the most popular methods. However, it has two modeling issues. First, CTC relies only on the hidden feature vector at the current time to make predictions, causing the hard alignment problem. Second, CTC imposes the conditional independence constraint that output predictions are independent, which is not true for ASR [15]. In order to remove the conditional independence assumption in the CTC model, RNN-T introduces a prediction network to learn context information, which functions as a language model. A joint network is subsequently used to combine the acoustic representation and the context representation to compute the posterior probability.

By contrast, the Attention ED does not require any conditional independence assumption. In Attention ED, the decoder network uses an attention mechanism to find an alignment between each element of the output sequence and the hidden states generated by the encoder network. At each output position, the decoder network computes a matching score between its hidden state and the states of the encoder network at each input time, to form a temporal alignment distribution, which is then used to extract an average of the corresponding encoder states. The Attention ED simplifies the ASR system and removes the conditional independence presumption. However, two issues exist in the system: first, the length variation of the input and output sequences in ASR makes it more difficult to track the alignment. Second, the basic temporal attention mechanism is too flexible in the scene in which it allows extremely non-sequential alignments. In speech recognition, the feature inputs and corresponding letter outputs generally proceed in the same order with only small within-word deviations.

To address the alignment problems of attention-based mechanisms in Attention ED, the hybrid CTC/Attention E2E architecture is proposed in [16,17]. During training, a CTC objective is attached to the attention-based encoder network as a regularization. The constrained CTC alignments provide rigorous constraints to guide the encoder–decoder training. This greatly reduces the number of irregularly aligned utterances without any heuristic search techniques. When employing decoding, both attention-based scores and CTC scores are combined in a rescoring/one-pass beam search algorithm to eliminate the irregular alignments. Although the hybrid CTC/Attention method has demonstrated the effectiveness with English, spontaneous Japanese, and Mandarin Chinese ASR tasks, it still has some disparities with conventional systems. As mentioned above, the CTC branch is very helpful to solve the misalignment issues existing in the ordinary Attention ED ASR. However, the hard alignment issue of CTC still exists. This motivates us to strengthen the modeling ability of CTC to boost the hybrid CTC/Attention system.

To solve the CTC hard alignment problem, we first investigate the combination of location-aware attention and CTC, establishing the proper restricted attention window size. Inspired by the success of a transformer on neural machine translation (NMT) tasks [18], we further introduce the time-restricted self-attention [19] CTC/Attention that can better model the long-range dependencies among the frames. Experiments with four datasets all verify the effectiveness of this method.

This paper is organized as follows. Section 2 introduces CTC, Attention-based Encoder–Decoder, and the hybrid CTC/Attention Encoder–Decoder models. Section 3 details the proposed methods including the time-restricted location-aware attention CTC/Attention and the time-restricted self-attention CTC/Attention. Section 4 presents the experimental setup, the experiments are conducted in WSJ [20], AMI [21], and SWBD [22] datasets firstly. Moreover, the algorithm is verified on the multichannel end-to-end distant speech recognition task using the CHIME-4 [23] dataset. Section 5 concludes the paper.

2. End-to-End Speech Recognition

2.1. Connectionist Temporal Classification (CTC)

A CTC network uses a recurrent neural network (RNN) and CTC error criterion to directly optimize the prediction of a transcription sequence. To deal with the issue that output length is shorter than input speech frames, CTC adds a blank symbol as an additional label to the label set and allows repetition of labels or blank across frames. The CTC model predicts the conditional probability of the whole label sequence as:

$$L_{CTC} = -lnp(\boldsymbol{y}|\boldsymbol{x}) = -ln \sum_{\boldsymbol{\pi} \in B^{-1}(\boldsymbol{y})} p(\boldsymbol{\pi}|\boldsymbol{x}), \tag{1}$$

With the conditional independence assumption, $p(\boldsymbol{\pi}|\boldsymbol{x})$ can be decomposed into a product of posteriors of each frame. Thus, Equation (1) can be written as:

$$L_{CTC} = -ln \sum_{\pi:\pi \in \boldsymbol{L}', B(\pi_{1:T}) = \boldsymbol{y}} \prod_{t=1}^{T} p(\pi_t|\boldsymbol{x}). \tag{2}$$

where $\boldsymbol{y}$ denotes the output label sequence. $\boldsymbol{y} \in \boldsymbol{L}$ and $\boldsymbol{L}$ is the label set for ASR. $\boldsymbol{x} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_T})$ is the corresponding feature sequence, t is the index of frame, and T is the total number frames. $\boldsymbol{\pi_{1:T}} = (\pi_1, ..., \pi_T)$ is the frame-wise CTC output symbol path from 1 to T. Each output symbol $\pi_t \in \boldsymbol{L}'$ and $\boldsymbol{L}' = \boldsymbol{L} \cup blank$. $p(\pi_t|\boldsymbol{x})$ is the probability of output symbol of CTC network at time t. A many-to-one mapping B is defined as $B : \boldsymbol{L}' \rightarrow \boldsymbol{L}$ to determine the correspondence between a set of paths and the output label sequence.

2.2. Attention-Based Encoder–Decoder

Compared with the CTC approach above, the Attention-based Encoder–Decoder model does not make any conditional independence assumptions and directly estimates the posterior. It employs two distinct networks: an RNN encoder network that transforms the input feature $\boldsymbol{x}$ into hidden vectors $\boldsymbol{h}$ and an RNN decoder network that transforms $\boldsymbol{h}$ into output labels $\boldsymbol{y}$. Using these, the posterior probability is:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{u=1}^{\boldsymbol{U}} p(\boldsymbol{y_u}|\boldsymbol{y_{1:u-1}}, \boldsymbol{c_u}), \tag{3}$$

where $\boldsymbol{c_u}$ is the context vector that is a function of $\boldsymbol{x}$ at time u. $\boldsymbol{U}$ is the length of output sequence that is allowed to differ from the input length T. The $p(\boldsymbol{y_u}|\boldsymbol{y_{1:u-1}}, \boldsymbol{c_u})$ is obtained as :

$$\boldsymbol{h_t} = Encoder(\boldsymbol{x}), \tag{4}$$

$$a_{ut} = Attend(\boldsymbol{s_{u-1}}, \boldsymbol{a_{u-1}}, \boldsymbol{h_t}), \tag{5}$$

$$\boldsymbol{c_u} = \sum_{t=1}^{T} a_{ut}\boldsymbol{h_t}, \tag{6}$$

$$p(\boldsymbol{y_u}|\boldsymbol{y_{1:u-1}}, \boldsymbol{c_u}) = Decoder(\boldsymbol{y_{u-1}}, \boldsymbol{s_{u-1}}, \boldsymbol{c_u}), \tag{7}$$

where $Encoder(\cdot)$ and $Decoder(\cdot)$ are RNN networks. $\boldsymbol{s}$ is the hidden state of Decoder. $\boldsymbol{h}$ is the hidden vector generated by an Encoder. a_{ut} is an attention weight. $Attend(\cdot)$ computes the attention weight a_{ut} using a single layer feedforward network:

$$e_{ut} = Score(\boldsymbol{s_{u-1}}, \boldsymbol{a_{u-1}}, \boldsymbol{h_t}), \tag{8}$$

$$a_{ut} = \frac{exp(e_{ut})}{\sum_{t'=1}^{T} exp(e_{ut'})}, \tag{9}$$

where $Score(\cdot)$ can either be content-based attention or location-aware attention.

Content-based attention mechanism is represented as follows:

$$e_{ut} = \boldsymbol{v^T} tanh(\boldsymbol{K}\boldsymbol{s_{u-1}} + \boldsymbol{W}\boldsymbol{h_t}), \tag{10}$$

where $\boldsymbol{v}$ is a learnable vector parameter. $tanh(\cdot)$ is a hyperbolic tangent activation function. $\boldsymbol{K}$ and $\boldsymbol{W}$ are learnable matrix parameters of the linear layers.

Corresponding to the content-based attention, location-aware attention encodes both content factor $\boldsymbol{s_{u-1}}$ and location information $\boldsymbol{a_{u-1}}$. Thus, $Score(\cdot)$ is computed as follows:

$$e_{ut} = \boldsymbol{v}^{\boldsymbol{T}} tanh(\boldsymbol{K}\boldsymbol{s_{u-1}} + \boldsymbol{Q}(\boldsymbol{F} * \boldsymbol{a_{u-1}}) + \boldsymbol{W}\boldsymbol{h_t}). \tag{11}$$

where attention parameters $\boldsymbol{W}, \boldsymbol{v}$ have the same meaning as above. The operation $*$ denotes convolution. Meanwhile, $\boldsymbol{F}$ and $\boldsymbol{Q}$ are the convolution and linear layer parameters, respectively.

2.3. Hybrid CTC/Attention Encoder–Decoder

To address the irregular alignments problem in attention mechanism, the authors in [17] propose the hybrid CTC/Attention architecture. Unlike the attention encoder–decoder model, it utilizes a CTC objective function as an auxiliary task to train the encoder network within a multitask learning (MTL) framework. During training, the forward-backward algorithm of CTC can enforce monotonic alignment between speech and label sequences. The objective to be maximized is a logarithmic linear combination of the CTC and attention objectives:

$$L_{MTL} = \lambda log p_{ctc}(\boldsymbol{y}|\boldsymbol{x}) + \lambda log p_{att}(\boldsymbol{y}|\boldsymbol{x}) \tag{12}$$

with a tunable parameter $\lambda : 0 \leq \lambda \leq 1$.

3. Time-Restricted Attention CTC/Attention Encoder–Decoder

As mentioned above, the CTC branch in hybrid CTC/Attention is effective to solve the irregular alignments problem. Inspired by this, the modeling ability of CTC is strengthened to boost the hybrid CTC/Attention system. In the theory of CTC, the conditional independence assumption is adopted to decompose the posterior probability of the frame sequences. Since CTC relies on the hidden feature vector at the current time to make predictions, it does not directly attend to feature vectors of the neighboring frames. This is the hard alignment problem that makes CTC's output independent assumption worse.

In this section, we propose the time-restricted attention CTC/Attention Encoder–Decoder shown in Figure 1. An additional attention layer is placed before the final projection layer in the CTC branch. The attention layer generates new hidden features that carry attention weighted context information. Moreover, inspired by the temporally selective mechanism in speech perception, we employ a time-restricted window in the attention layer. Finally, two representative attention CTC mechanisms, location-aware attention CTC and self-attention CTC, are investigated. In this study, the proposed attention CTC model is different from the existed CTC or attention modeling approaches since we use attention mechanism to improve the hidden layer representations with more context information without changing the CTC objective function and the training process. Our primary motivation is to address the hard alignment problem of CTC by modeling attention directly within the CTC branch in the hybrid CTC/Attention architecture. The location-aware attention CTC/Attention (LA CTC/Attention) is introduced in Section 3.1 firstly. Then, we improve our modeling ability further by proposing self-attention CTC/Attention (SA CTC/Attention) in Section 3.2.

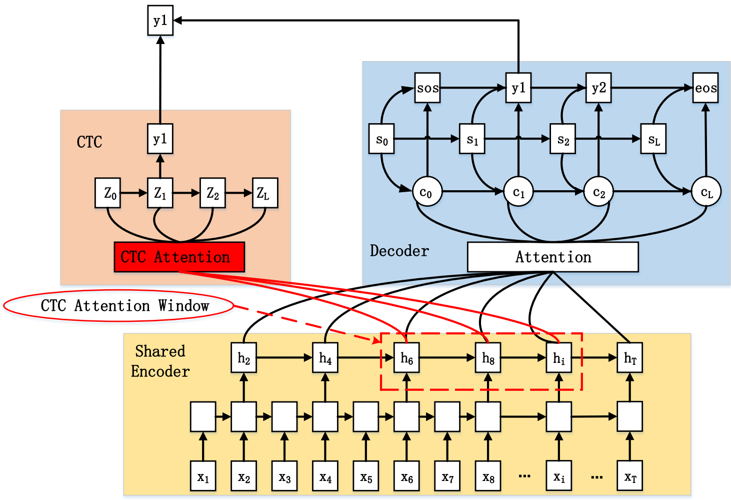

Figure 1. The time-restricted attention CTC/Attention architecture.

3.1. LA CTC/Attention

In the hybrid CTC/Attention architecture, the CTC objective function is utilized as an auxiliary task with an Attention ED cost function in a multitask learning framework. Denoting $\boldsymbol{h}$ as the output sequence of the encoder network, a projection layer employs it as input and transforms it to a particular dim representing the number of CTC output labels. Then, the projected output is optimized with the CTC criterion discussed in Section 2.1:

$$\boldsymbol{ph} = \boldsymbol{W_{proj}h} + \boldsymbol{b}, \tag{13}$$

$$L_{CTC} = -lnp(\boldsymbol{y}|\boldsymbol{ph}) = -ln \sum_{\pi \in B^{-1}(\boldsymbol{y})} p(\pi|\boldsymbol{ph}), \tag{14}$$

where $\boldsymbol{W_{proj}}$ and $\boldsymbol{b}$ are the weight matrix and bias of the CTC projection layer. $\boldsymbol{y}$ denotes the output label sequence. $\boldsymbol{ph}$ represents the output of the CTC projection layer. The many-to-one mapping B is defined in Section 2.1.

In order to address the conditional independence assumption in CTC, an attention layer is placed before the CTC projection layer. Then, the attention layer output that carries context information is served as the input of CTC projection layer at the current time u:

$$a_{ut} = Attend(\boldsymbol{ph_{u-1}}, \boldsymbol{a_{u-1}}, \boldsymbol{h_t}), \tag{15}$$

$$\boldsymbol{c_u} = \sum_{t=1}^{T} a_{ut}\boldsymbol{h_t}, \tag{16}$$

$$\boldsymbol{ph_u} = \boldsymbol{W_{proj}c_u} + \boldsymbol{b}, \tag{17}$$

where $\boldsymbol{ph_u}$ is the output of CTC projection layer at time u. a_{ut} is the attention weight. $\boldsymbol{c_u}$ is the weighted hidden features. $Attend(\cdot)$ computes the attention weight a_{ut} using a single layer feedforward network:

$$e_{ut} = \boldsymbol{v^T} tanh(\boldsymbol{Kph_{u-1}} + \boldsymbol{Q}(\boldsymbol{F} * \boldsymbol{a_{u-1}}) + \boldsymbol{Wh_t}), \tag{18}$$

$$a_{ut} = \frac{exp(e_{ut})}{\sum_{t'=1}^{T} exp(e_{ut'})}, \tag{19}$$

where $\boldsymbol{v}$ is a learnable vector parameter. $tanh(\cdot)$ is a hyperbolic tangent activation function. $\boldsymbol{K}$ and $\boldsymbol{W}$ are learnable matrix parameters of the linear layers. The operation $*$ denotes convolution. Meanwhile, $\boldsymbol{F}$ and $\boldsymbol{Q}$ are the convolution and linear layer parameters, respectively.

In practice, our attention model considers a small subsequence of $\boldsymbol{h}$ rather than the entire sequence. This subsequence, $(\boldsymbol{h}_{u-\tau}, \cdots, \boldsymbol{h}_u, \cdots, \boldsymbol{h}_{u+\tau})$, is referred to as the *attentionwindow*. It is centered around the current time u. Let τ represent the length of the *attentionwindow* on either side of u. Thus, the resulting vector $\boldsymbol{c_u}$ in Equation (16) is replaced by:

$$\boldsymbol{c_u} = \sum_{t=u-\tau}^{u+\tau} \boldsymbol{a_{ut}h_t}. \tag{20}$$

3.2. SA CTC/Attention

In this section, we investigate another attention-based paradigm known as self-attention. In self-attention, the weights are computed from the hidden features only. It does not use any past output predictions, which is more parallelizable than location-aware attention.

First, the hidden features are converted into input embeddings using the embedding matrix $\boldsymbol{W_{embd}}$ as:

$$\boldsymbol{b_t} = \boldsymbol{W_{embd}h_t}, t = u-\tau, \cdots, u+\tau. \tag{21}$$

Second, use linear projections of the embedding vector $\boldsymbol{b_t}$ to compute three kinds of vectors, keys, values and query, in the self-attention block:

$$\begin{aligned} \boldsymbol{q_t} &= \boldsymbol{Qb_t}, t = u, \\ \boldsymbol{k_t} &= \boldsymbol{Kb_t}, t = u-\tau, \cdots, u+\tau, \\ \boldsymbol{v_t} &= \boldsymbol{Vb_t}, t = u-\tau, \cdots, u+\tau, \end{aligned} \tag{22}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are the query, key and value matrices, respectively.

Third, the attention weight $\boldsymbol{a_u}$ and attention result $\boldsymbol{c_u}$ are derived by:

$$\begin{aligned} e_{ut} &= \frac{\boldsymbol{q_u^T k_t}}{\sqrt{d_k}}, \\ a_{ut} &= \frac{exp(e_{ut})}{\sum_{t'=1}^{T} exp(e_{ut'})}, \\ \boldsymbol{c_u} &= \sum_{t=u-\tau}^{u+\tau} \boldsymbol{a_{ut}h_t}. \end{aligned} \tag{23}$$

Finally, the weighted hidden feature $\boldsymbol{c_u}$ is projected by the CTC projection layer as described in Equation (17). Then, the projected information is optimized by Equation (14).

As noted above, a self-attention layer connects all positions with a constant number of sequentially executed operations, whereas the location-aware attention requires $O(n)$ sequential operations. Learning long-range dependencies is a key challenge in many sequence transduction tasks. One key factor affecting the ability to learn such dependencies is the length of the paths forward and backward signals have to traverse in the network. The shorter these paths between any combination of positions in the input and output sequences, the easier it is to learn long-range dependencies. Therefore, self-attention can better model the long-range dependencies than the location-aware attention.

4. Experiments

4.1. Experimental Setup

We demonstrate the effectiveness of proposed hybrid attention CTC/Attention framework firstly in three different ASR datasets, WSJ [20], AMI [21], and SWBD [22]. All experiments are implemented by ESPnet [17] with the default configurations. In particular, for AMI, ESPnet only provides the individual headset microphone (IHM) data training scripts. To ensure fairness, we just exhibit the AMI-IHM results. Moreover, for SWBD, ESPnet only shows the decoding results without CTC and language model. However, we decode SWBD with both CTC constraint and recurrent neural network language model (RNNLM). In addition, we also experiment decoding without language model for all the above datasets. The detail experimental configuration is exhibited in Table 1.

Since the above-mentioned experimental datasets mainly focus on ASR in clean environments, we intend to explore whether the proposed method is robust to noise and reverberation. In particular, CHIME-4 [23], an ASR task for public noisy environments, consists of speech recorded using a tablet device with 6-channel microphones in four environments: cafe (CAF), street junction (STR), public transportation (BUS), and pedestrian area (PED). Therefore, we conduct the multichannel distant ASR experiments in CHIME-4. In addition, et05simu and dt05simu are the simulated evaluate and development datasets. Meanwhile, et05real and dt05real represent the real recorded evaluate and development datasets, respectively.

In this paper, we follow the unified architecture proposed in [24], which jointly optimizes the multichannel enhancement and the ASR components. The experiment details are exhibited in Section 4.3. "blstmp" means that encoder is the projected bidirectional long short-term memory neural network. "Vggblstmp" means that an encoder is composed of vaccination guidelines group (VGG) [25] layer and blstmp layer.

Table 1. Experimental configuration.

Model	WSJ	AMI	SWBD
Encoder type	Vggblstmp	blstmp	blstmp
Encoder layers	6	8	6
Subsampling	4	4	4
Attention	Location-aware	Location-aware	Location-aware
CTCWeight train	0.5	0.5	0.5
CTCWeight decode	0.3	0.3	0.3
RNNLM-unit	Word	Word	Word

4.2. *WSJ, AMI, and SWBD*

4.2.1. Baseline Results

The baseline results of the above three sets are listed in Table 2. It is worth noting that all the character error rate (CER) and WER results are derived with the official scripts in ESPnet. Anyone can easily repeat the baseline results by following the ESPnet [17] scripts. However, one of the issues of ESPnet results is that their performances do not reach those of the state-of-the-art hybrid HMM/DNN systems. According to the explanation of ESPnet, applying this technique to these English datasets will encounter the issue of long sequence lengths, which requires a large computational cost and makes it difficult to train a decoder network. Furthermore, end-to-end speech recognition usually requires more data than the traditional model to achieve the best performance.

For a fair comparison, we will try our best to list the comparable systems referred to in the literature. Thus, for WSJ, consistent with the discussion in [17], the comparable end-to-end method is described in [2]. This method utilizes the CTC as the training criterion and decodes based on the weighted finite-state transducers (WFSTs). It achieves 7.3% WER in the eval set, while the ESPnet reaches 5.4% WER in this data set.

For AMI-IHM, we haven't found clear results of end-to-end method in the references. Therefore, we employ the result in Kaldi [26]. The best method uses the TDNN-LSTM acoustic model, lattice-free version of the maximum mutual information criterion [27] and the per-frame dropout techniques. Finally, it reaches 19.8% WER in dev set and 19.2% WER in eval set.

For SWBD, the state-of-the-art ASR system is proposed in [28], which utilizes a blstmp acoustic model, spatial smoothing, and speaker adaptive modeling techniques. It achieves 12.0% WER in the eval set, training with 300h SWBD data. In the meantime, paying attention to the end-to-end methods, the best system [29] acquires 15.5% WER in the eval set with the Attention ED framework. When applying spectrogram augmentation methods, it reaches 10.6% WER, surpassing the performance of hybrid HMM/DNN systems.

For CHIME-4, a traditional HMM/DNN method [30] reaches 5.42% WER in et05real and 3.9% WER in et05simu. However, in consideration of the end-to-end method, the authors in [31] acquire 22.7% CER in et05real and 22.5% CER in et05simu. Although the ESPnet results fall behind the HMM/DNN system, it reaches comparable results to the other end-to-end methods.

In this paper, although the performances do not reach those of the start-of-the-art, we believe it has no relation with our key improvements. We pay more attention to the relative decrease of WER compared with the original hybrid CTC/Attention system.

4.2.2. LA CTC/Attention

Before applying location-aware attention CTC/Attention, the *attentionwindow* length C should be determined first. In this section, considering the size of three data sets, we choose the smaller sets, WSJ and AMI, to find the proper C. The results are listed in Table 2.

Table 2. The performances of the LA CTC/Attention for WSJ, AMI, and SWBD with different *attentionwindow* lengths. Additionally, the baseline is the results of ESPnet.

WSJ	**LM Weight = 1**				**LM Weight = 0**			
	CER		**WER**		**CER**		**WER**	
Model	**dev**	**test**	**dev**	**test**	**dev**	**test**	**dev**	**test**
Baseline	3.9	2.4	8.6	5.4	7.5	5.7	22.2	17.7
$C = 3$	**3.8**	2.4	**8.4**	5.3	7.8	5.9	22.8	17.9
$C = 5$	3.9	**2.4**	8.6	**5.1**	**7.5**	**5.5**	22	**16.7**
$C = 7$	3.9	2.6	8.6	5.6	7.3	5.4	**21.5**	16.7
$C = 9$	3.9	2.5	8.6	5.5	7.7	6.0	22.6	18.2
AMI	**LM Weight = 1**				**LM Weight = 0**			
	CER		**WER**		**CER**		**WER**	
Model	**dev**	**test**	**dev**	**test**	**dev**	**test**	**dev**	**test**
Baseline	**22.3**	23.2	**35.1**	37.4	**23.4**	**24.7**	**39.6**	**42.2**
$C = 3$	22.3	23.6	35.3	38.1	23.5	25.1	40.1	43.1
$C = 5$	24.3	**23.0**	39.2	**36.3**	24.2	25.8	41.5	44.5
$C = 7$	26.1	24.3	41.6	38.8	25.4	27.5	43.7	47.1
$C = 9$	24.9	26.2	39.3	42.0	27.6	25.9	47.2	44.3
SWBD	**LM Weight = 1**				**LM Weight = 0**			
	CER		**WER**		**CER**		**WER**	
Model	**eval**	**rt03**	**eval**	**rt03**	**eval**	**rt03**	**eval**	**rt03**
Baseline	30.2	31.9	47.8	50.6	30.8	32.5	51.0	54.0
$C = 5$	**29.4**	**31.5**	**47.3**	**50.3**	**30.1**	**32.4**	**50.5**	**53.8**

For WSJ, LA CTC/Attention obtains overall the best CER and WER performances in both dev and test sets when *attentionwindow* size equals 5. For AMI-IHM, the LA CTC/Attention performs worse than baseline results. However, focusing only on the results of LA CTC/Attention, the best results are achieved in the case that *attentionwindow* size is 3. Moreover, with *attentionwindow* length increasing, the performances degrade rapidly and fall far behind the original hybrid CTC/Attention framework. We suppose the reason is that the speech duration of AMI is too short. Revisiting the AMI-IHM and WSJ data, we figure out that the average frames of WSJ and AMI training data are 782 and 257, respectively. Moreover, the encoder network subsamples four times. Thus, the phenomenon that proper *attentionwindow* size in AMI is shorter than that in WSJ is reasonable.

For SWBD, since the average speech duration is 432 frames, which is twice that in AMI, we choose 5 as the SWBD *attentionwindow* size. Compared with the baseline, LA CTC/Attention obtains 2.6% and 1.3% relative CER reduction and 1.0% and 0.6% relative WER reduction for eval and rt03 with RNNLM. When decoding without RNNLM, it also achieves 2.3% and 0.3% relative CER decrease and 1.0% and 0.4% relative WER decrease.

4.2.3. SA CTC/Attention

In this section, the SA CTC/Attention architecture is conducted on the above three datasets. Based on the experiments and analyses in Section 4.2.2, we choose 5 as the *attentionwindow* size. In addition, the self-attention key, query, and value dimensions are all set 1024. To reduce computing complexity, we only experiment with single attention and fix the self-attention head quantity as 1. The results are shown in Table 3.

Table 3. The performances of the SA CTC/Attention for WSJ, AMI, and SWBD compared with the LA CTC/Attention and baseline architectures.

WSJ	LM Weight = 1				LM Weight = 0			
	CER		WER		CER		WER	
Model	dev	test	dev	test	dev	test	dev	test
Baseline	3.9	2.4	8.6	5.4	7.5	5.7	22.2	17.7
LA CTC/Attention	3.9	2.4	8.6	5.1	7.5	5.5	22.0	**16.7**
SA CTC/Attention	**3.7**	**2.2**	**8.1**	**4.9**	**7.4**	**5.5**	**21.5**	16.8
AMI	LM Weight = 1				LM Weight = 0			
	CER		WER		CER		WER	
Model	dev	test	dev	test	dev	test	dev	test
Baseline	22.3	23.2	35.1	37.4	23.4	24.7	39.6	42.2
LA CTC/Attention	22.3	23.6	35.3	38.1	23.5	25.1	40.1	43.1
SA CTC/Attention	**21.0**	**22.1**	**33.4**	**35.8**	**22.1**	**23.7**	**37.7**	**40.8**
SWBD	LM Weight = 1				LM Weight = 0			
	CER		WER		CER		WER	
Model	eval	rt03	eval	rt03	eval	rt03	eval	rt03
Baseline	30.2	31.9	47.8	50.6	30.8	32.5	51.0	54.0
LA CTC/Attention	29.4	31.5	47.3	50.3	30.1	32.4	50.5	53.8
SA CTC/Attention	**26.9**	**29.0**	**43.7**	**47.0**	**27.8**	**30.0**	**46.7**	**50.2**

For WSJ, although applying locate-aware attention with CTC contributes to improving the performance, the WER or CER remain constant in some cases. For example, in Table 3, the CER and WER of dev set have not been improved when decoding with RNNLM. However, the WER and CER of SA CTC/Attention achieve a consistent decline in all cases. In particular, it obtains 5.8% and 9.3% relative WER reduction for dev and test, when decoding with RNNLM.

For AMI-IHM, LA CTC/Attention performs worse than the baseline as described in the previous section. However, compared with the baseline, the SA CTC/Attention achieves more than 3% performance improvement in all sets whether decoding with or without RNNLM. Particularly, it achieves 4.8% and 4.3% relative WER reduction for dev and eval with RNNLM. At the same time, it also acquires 4.8% and 3.3% relative WER reduction without RNNLM.

For SWBD, the LA CTC/Attention already achieves overall performance improvement in all situations. Nevertheless, the SA CTC/Attention gets further improvement than the LA CTC/Attention. While decoding with RNNLM, the SA CTC/Attention acquires 8.6% and 7.1% relative WER reduction compared with the baseline. Furthermore, it achieves 8.4% and 7.0% relative WER decrease without RNNLM.

Apart from the CER and WER improvement, self-attention is very helpful in accelerating the convergence during training. Figure 2 shows the train and valid losses of the baseline, LA CTC/Attention and SA CTC/Attention architectures for SWBD. It can be observed that the SA CTC/Attention converges most quickly and has the lowest loss curves among the three systems.

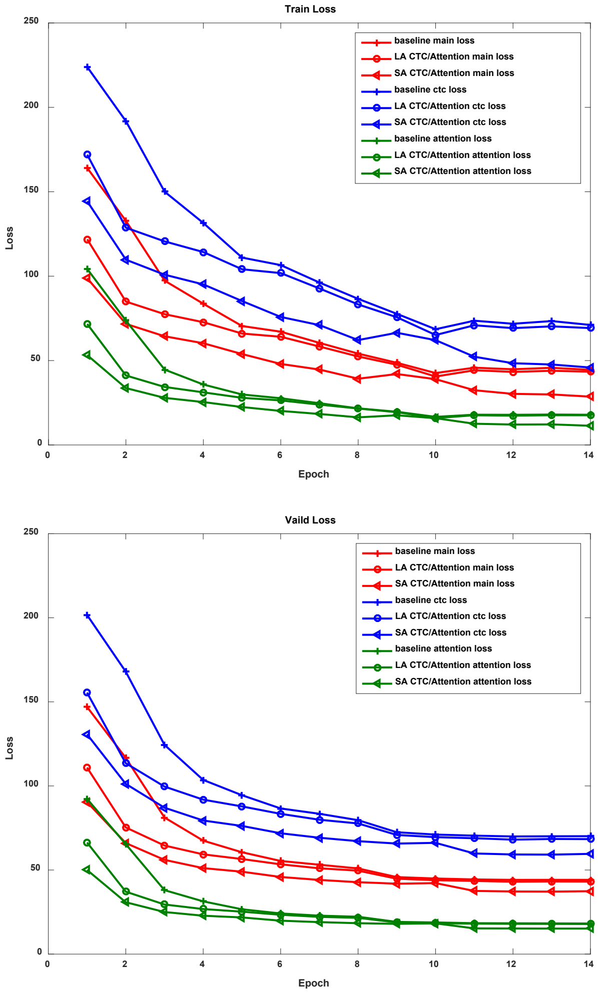

Figure 2. The train and valid losses of the baseline, LA CTC/Attention, and SA CTC/Attention architectures for SWBD.

4.3. CHIME-4

In this subsection, we evaluate the proposed LA CTC/Attention and SA CTC/Attention for multichannel end-to-end speech recognition task on the CHIME-4 dataset. The multichannel ASR algorithm allows the mask-based neural beamformer and ASR components to be jointly optimized to improve the end-to-end ASR objective and leads to an end-to-end framework that works well in the presence of strong background noise. The overall Multichannel Hybrid Attention CTC/Attention ASR framework is shown in Figure 3. A unified architecture [24] was proposed to jointly optimize the neural beamformer and the ASR components. The neural beamformer consists of two parts: ideal ratio mask (IRM) estimator and the Minimum Variance Distortionless Response (MVDR) beamformer. Joint training means that the gradients derived from the ASR loss will back-propagate through all the way from the acoustic model to the complex-valued beamforming and the mask estimation network. Therefore, the improvement of acoustic modeling ability will strengthen the capability of a neural beamforming network. In this paper, we reproduce the multichannel ASR architecture as described in ESPnet firstly. Then, we apply the time-restricted LA CTC/Attention and time-restricted SA CTC/Attention separately to improve the back-end acoustic modeling ability.

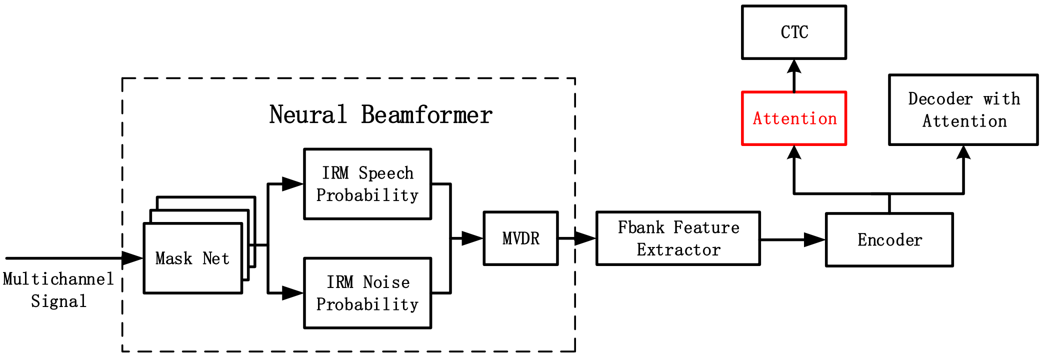

Figure 3. The multichannel time-restricted attention CTC/Attention architecture.

According to the experiments and analyses before, we set the *attentionwindow*, self-attention dim, and self-attention head quantity just as the corresponding parameters in Section 4.2. The detailed results are shown in Table 4.

Table 4. The performances of the SA CTC/Attention for CHIME-4 compared with the LA CTC/Attention and baseline architectures.

CHIME-4	LM Weight = 0							
	CER				WER			
Model	et05simu	et05real	dt05simu	dt05real	et05simu	et05real	dt05simu	dt05real
Baseline	12.2	14.5	**8.7**	8.6	29.5	33.1	**21.5**	21.7
LA CTC/Attention	12.7	15.1	9.5	9.2	30.4	34.2	23.4	22.9
SA CTC/Attention	**11.9**	**14.0**	8.8	**8.5**	**28.4**	**32.2**	21.7	**21.7**
CHIME-4	LM Weight = 1							
	CER				WER			
Model	et05simu	et05real	dt05simu	dt05real	et05simu	et05real	dt05simu	dt05real
Baseline	6.9	9.2	4.1	4.5	13.4	16.9	8.0	9.2
LA CTC/Attention	6.9	8.9	4.5	4.6	13.0	16.2	8.8	9.1
SA CTC/Attention	**6.2**	**8.3**	**4.1**	**4.1**	**11.9**	**15.6**	**8.0**	**8.2**

Firstly, we show the decoding results without RNNLM. Compared with the baseline, LA CTC/Attention generally performs worse than the baseline system. However, the SA CTC/Attention achieves 2.5% and 3.4% relative CER reduction for et05simu and et05real. Moreover, it also achieves 3.7% and 2.7% relative WER reduction for et05simu and et05real. Meanwhile, the CER and WER of SA CTC/Attention in development datasets generally remain steady.

In order to improve the overall performances, RNNLM is employed during decoding. With RNNLM, the LA CTC/Attention acquires 3.0% and 4.1% relative WER reduction for et05simu and et05real. However, it performs poorly in the simulate development dataset. For SA CTC/Attention, however, it acquires consistent WER and CER degradation for all evaluation and development sets. In particular, it achieves 11.2% and 7.7% relative WER reduction for et05simu and et05real. At the same time, it also achieves 10.1% and 9.8% CER reduction for et05simu and et05real. For the development sets, although the WER keeps constant for dt05simu, it obtains 8.9% relative CER reduction and 11.0% relative WER decrease for dt05real.

As the neural beamformer and ASR components are jointly optimized, enhancing the acoustic modeling ability contributes to the capability of neural beamforming network. To evaluate the modeling ability of the neural beamforming, we compute the objective measure PESQ (Perceptual

Evaluation of Speech Quality) of the enhanced speech. Table 5 shows the PESQ results of the three systems. It can be observed that the SA CTC/Attention behaves almost the best in all conditions. This indicates that, while improving the performance of ASR components, the SA CTC/Attention is also conducive to the optimization of the neural beamformer.

In order to further make a comparison between the LA CTC/Attention and SA CTC/Attention, we analyze the losses during training. Figure 4 shows that the SA CTC/Attention framework arrives at the lowest main loss, CTC loss, and attention loss among the three systems. These results indicate that our proposed SA CTC/Attention algorithm works efficiently and improves ASR performance.

Table 5. The PESQ of the SA CTC/Attention for CHIME-4 compared with the LA CTC/Attention and baseline architectures.

CHIME-4	et05simu					dt05simu				
Model	PED	CAF	STR	BUS	MEAN	PED	CAF	STR	BUS	MEAN
Baseline	2.667	2.598	2.681	2.864	2.702	2.736	2.504	2.667	2.827	2.683
LA CTC/Attention	2.653	2.598	**2.682**	2.860	2.698	2.740	2.518	2.667	2.826	2.687
SA CTC/Attention	**2.667**	**2.603**	2.679	**2.865**	**2.703**	**2.743**	**2.524**	**2.672**	**2.831**	**2.692**

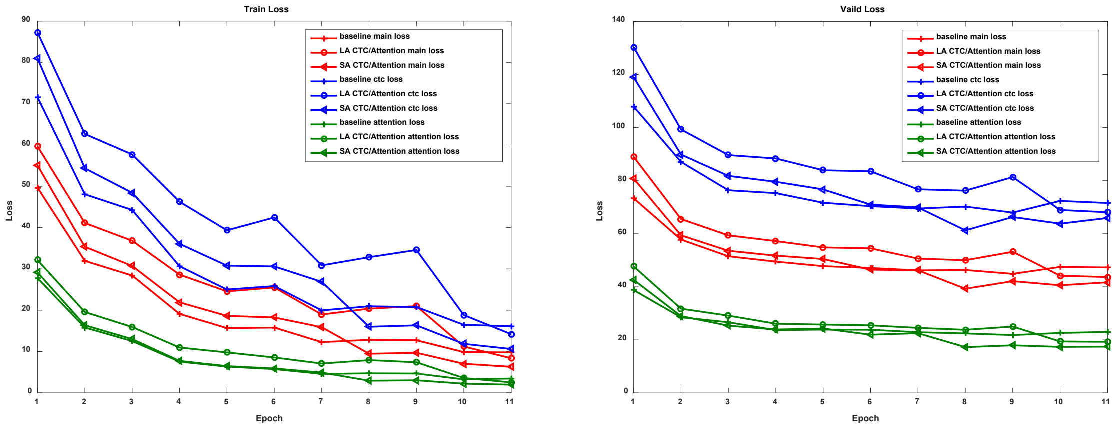

Figure 4. The train and valid loss of the baseline, LA CTC/Attention and SA CTC/Attention architectures for CHIME-4.

5. Conclusions

In order to improve the performance of hybrid CTC/Attention end-to-end ASR, this paper proposes integrating attention mechanism with the CTC branch to address CTC's output independent assumption. Firstly, we explore time-restricted location-aware attention CTC/Attention, establishing the proper time-restricted attention window size. "Time-restricted" indicates that the attention mechanism is conducted on a limited window of frames to the left and right. Inspired by the success of self-attention in machine translation, we further introduce the time-restricted self-attention CTC/Attention that can better model the long-range dependencies among the frames. For a fair comparison, our experiments are carried out on three clean datasets, WSJ, AMI-IHM, and SWBD. For WSJ, the SA CTC/Attention obtains 5.8% and 9.3% relative WER reduction in dev and test with RNNLM. For AMI-IHM, it achieves 4.8% and 4.3% relative WER decrease in dev and eval with RNNLM. For SWBD, it reaches 8.6% and 7.1% relative WER reduction in eval and rt03 with RNNLM. With the exception of WER and CER evaluation criteria, we also exhibit the details of the training process and present the loss curves. The lowest loss curves of SA CTC/Attention also verify the effectiveness of the algorithm. To explore the robustness of this method to noise and reverberation, we also experiment with the SA CTC/Attention framework in CHIME-4. The results indicate that it achieves 11.2% and 7.7% relative WER reduction in et05simu and et05real.

As the recurrence based encoder–decoder architecture is very time-consuming during training and decoding, one future research direction is to investigate utilizing the non-recursive framework. In the meantime, incorporating the attention modeling with CTC in non-recursive framework could be explored. Another future work is to improve the performances of the baseline system.

Author Contributions: Methodology, validation and writing, original draft, L.W. (Long Wu) Review, editing, and formal analysis, T.L. Review and formal analysis, L.W. (Li Wang) Supervision and project administration, Y.Y.

Funding: This work is partially supported by the National Natural Science Foundation of China (Nos. 11590774, 11590770), the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 2016A03007-1), and the IACAS Young Elite Researcher Project (QNYC201602).

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

References

1. Yu, D.; Li, J. Recent Progresses in Deep Learning Based Acoustic Models. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 396–409. [CrossRef]
2. Miao, Y.; Gowayyed, M.; Metze, F. EESEN: End-to-End Speech Recognition Using Deep RNN Models and WFST-Based Decoding. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AR, USA, 13–17 December 2015; pp. 167–174. [CrossRef]
3. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4960–4964. [CrossRef]
4. Prabhavalkar, R.; Rao, K.; Sainath, T.N.; Li, B.; Johnson, L.; Jaitly, N. A Comparison of Sequence-to-Sequence Models for Speech Recognition. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 939–943. [CrossRef]
5. Battenberg, E.; Chen, J.; Child, R.; Coates, A.; Li, Y.G.Y.; Liu, H.; Satheesh, S.; Sriram, A.; Zhu, Z. Exploring Neural Transducers for End-to-End Speech Recognition. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 206–213. [CrossRef]
6. Sak, H.; Shannon, M.; Rao, K.; Beaufays, F. Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017. [CrossRef]
7. Chiu, C.C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778. [CrossRef]
8. Sainath, T.N.; Chiu, C.C.; Prabhavalkar, R.; Kannan, A.; Wu, Y.; Nguyen, P.; Chen, Z. Improving the Performance of Online Neural Transducer Models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5864–5868. [CrossRef]
9. 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—ICML '06, Pittsburgh, PA, USA, 25–26 June 2006; pp. 369–376. [CrossRef]
10. Graves, A.; Jaitly, N. Towards End-To-End Speech Recognition with Recurrent Neural Networks. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1764–1772.
11. Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y. End-to-End Attention-Based Large Vocabulary Speech Recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4945–4949. [CrossRef]
12. 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, Montreal, QC, Canada, 7–12 December 2015.
13. Graves, A. Sequence Transduction with Recurrent Neural Networks. *Comput. Sci.* **2012**, *58*, 235–242.

14. Rao, K.; Sak, H.; Prabhavalkar, R. Exploring Architectures, Data and Units for Streaming End-to-End Speech Recognition with RNN-Transducer. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 193–199. [CrossRef]
15. Das, A.; Li, J.; Zhao, R.; Gong, Y. Advancing Connectionist Temporal Classification with Attention Modeling. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4769–4773. [CrossRef]
16. 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]
17. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Yalta, N.; 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. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762 [cs].
19. Povey, D.; Hadian, H.; Ghahremani, P.; Li, K.; Khudanpur, S. A Time-Restricted Self-Attention Layer for ASR. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5874–5878. [CrossRef]
20. Paul, D.B.; Baker, J.M. The Design for the Wall Street Journal-Based CSR Corpus. In Proceedings of the Workshop on Speech and Natural Language—HLT '91, Harriman, NY, USA, 23–26 February 1992; p. 357. [CrossRef]
21. Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; et al. The AMI Meeting Corpus: A Pre-Announcement. In *Machine Learning for Multimodal Interaction*; Renals, S., Bengio, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3869, pp. 28–39. [CrossRef]
22. Glenn, M.L.; Strassel, S.; Lee, H.; Maeda, K.; Zakhary, R.; Li, X. Transcription Methods for Consistency, Volume and Efficiency. In Proceedings of the LREC, Valletta, Malta, 17–23 May 2010.
23. Vincent, E.; Watanabe, S.; Nugraha, A.A.; Barker, J.; Marxer, R. An Analysis of Environment, Microphone and Data Simulation Mismatches in Robust Speech Recognition. *Comput. Speech Language* **2017**, *46*, 535–557. [CrossRef]
24. Ochiai, T.; Watanabe, S.; Hori, T.; Hershey, J.R.; Xiao, X. Unified Architecture for Multichannel End-to-End Speech Recognition With Neural Beamforming. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1274–1288. [CrossRef]
25. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
26. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.K.; Hannemann, M.; Motlícek, 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.
27. Povey, D.; Peddinti, V.; Galvez, D.; Ghahremani, P.; Manohar, V.; Na, X.; Wang, Y.; Khudanpur, S. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In Proceedings of the INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016. [CrossRef]
28. Xiong, W.; Droppo, J.; Huang, X.; Seide, F.; Seltzer, M.; Stolcke, A.; Yu, D.; Zweig, G. Achieving Human Parity in Conversational Speech Recognition. *arXiv* **2016**, arXiv:1610.05256.
29. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779.
30. Heymann, J.; Drude, L.; Böddeker, C.; Hanebrink, P.; Häb-Umbach, R. Beamnet: End-to-End Training of a Beamformer-Supported Multi-Channel ASR System. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5325–5329. [CrossRef]
31. Braun, S.; Neil, D.; Anumula, J.; Ceolini, E.; Liu, S.C. Multi-Channel Attention for End-to-End Speech Recognition. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018. [CrossRef]