IMPROVING NON-AUTOREGRESSIVE END-TO-END SPEECH RECOGNITION WITH PRE-TRAINED ACOUSTIC AND LANGUAGE MODELS

Keqi Deng[1,2,*], *Zehui Yang*[1,2,*], *Shinji Watanabe*[3], *Yosuke Higuchi*[4], *Gaofeng Cheng*[1], *Pengyuan Zhang*[1,2]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS, China
[2]University of Chinese Academy of Sciences, China
[3]Carnegie Mellon University, USA, [4]Waseda University, Japan

ABSTRACT

While Transformers have achieved promising results in end-to-end (E2E) automatic speech recognition (ASR), their autoregressive (AR) structure becomes a bottleneck for speeding up the decoding process. For real-world deployment, ASR systems are desired to be highly accurate while achieving fast inference. Non-autoregressive (NAR) models have become a popular alternative due to their fast inference speed, but they still fall behind AR systems in recognition accuracy. To fulfill the two demands, in this paper, we propose a NAR CTC/attention model utilizing both pre-trained acoustic and language models: wav2vec2.0 and BERT. To bridge the modality gap between speech and text representations obtained from the pre-trained models, we design a novel modality conversion mechanism, which is more suitable for logographic languages. During inference, we employ a CTC branch to generate a target length, which enables the BERT to predict tokens in parallel. We also design a cache-based CTC/attention joint decoding method to improve the recognition accuracy while keeping the decoding speed fast. Experimental results show that the proposed NAR model greatly outperforms our strong wav2vec2.0 CTC baseline (15.1% relative CER reduction on AISHELL-1). The proposed NAR model significantly surpasses previous NAR systems on the AISHELL-1 benchmark and shows a potential for English tasks.

Index Terms— Non-autoregressive, end-to-end speech recognition, CTC/attention speech recognition

1. INTRODUCTION

End-to-end (E2E) automatic speech recognition (ASR) models simplify the conventional pipeline ASR methods and directly convert input speech into corresponding text [1, 2]. As a way to realize E2E ASR, recurrent neural network transducer (RNN-T) [2] and attention-based encoder-decoder (AED)-based models have been actively studied [1]. These models are categorized as autoregressive (AR), which predict a sequence based on a left-to-right chain rule [3]. On the other hand, non-autoregressive (NAR) models have become popular [3–6] due to its fast inference speed, which can predict tokens simultaneously [4, 5] or iteratively [3, 6].

In general, AR models cannot be efficiently parallelized during inference, since their next token generation process depends on previously predicted tokens and requires incremental computations of a decoder [4, 6]. Although NAR models can greatly improve the efficiency of inference, they still face two challenges [7]. The first is to improve recognition performance since the NAR mechanism often prevents the model from learning the conditional dependencies between output tokens, thus making it fall behind the AR models in performance [6, 7]. The second is to improve the training efficiency [3, 5], as it is difficult to train NAR models and the training convergence is very slow [5].

Among various NAR methods, connectionist temporal classification (CTC) is a popular technique for training a NAR model [8]. CTC achieves a monotonic input-output alignment based on the conditional independence assumption between output tokens. Recently, a self-supervised training method, wav2vec2.0 [9], has achieved promising results on CTC models, and the pre-trained model is shown to accelerate the convergence during the fine-tuning stage. However, even with the pre-trained model obtained by wav2vec2.0, the CTC model needs an external language model (LM) to relax its conditional independence assumption [9, 10]. Several works have investigated incorporating BERT into a NAR ASR model to achieve better recognition accuracies [11–13]. In order to bridge the length gap between the frame-level speech input and token-level text output, [11] and [12] have introduced global attention and a serial continuous integrate-and-fire (CIF) [14], respectively. However, the global attention suffers from poor text length prediction [12, 15], and the serial computation in CIF greatly degrades its training efficiency. Another mismatch lies between the acoustic embedding and the linguistic token embedding of BERT. To solve this, [11] designs a two-stage training strategy and [12] proposes a modal fusion strategy, which both require significantly more training iterations.

In an attempt to improve training efficiency and recognition performance for NAR models, this paper proposes a novel NAR CTC/attention architecture to fully utilize both pre-trained acoustic and language models: wav2vec2.0 and BERT. In addition, a novel modality conversion mechanism (MCM) is proposed to efficiently bridge the gap between speech and text modalities[1]. Unlike previous methods [11, 12], our proposed MCM does not need to greatly increase training iterations. To mitigate the length gap between speech and text sequences, during training, we provide MCM with the ground truth of target length to achieve efficient training. During inference, we employ a CTC branch to generate a target length via greedy search, which enables the BERT to predict tokens in parallel. To further improve the recognition accuracy while keeping fast decoding speed, we design a cache-based CTC/attention joint decoding method. Experimental results show that the proposed model greatly outperforms a strong wav2vec2.0 CTC baseline, and the cache-based CTC/attention joint decoding method is significantly faster than conventional AR-based beam search. The proposed model substantially improves over previous NAR systems on the AISHELL-1 benchmark, while its improvement on the Switchboard is not as significant as on the AISHELL-1.

[*] Equal contribution. This work is partially supported by the National Key Research and Development Program of China (No. 2020AAA0108002).

[1] The MCM works well for character-based systems, but needs some improvements for byte pair encoding (BPE)-based [16] systems.

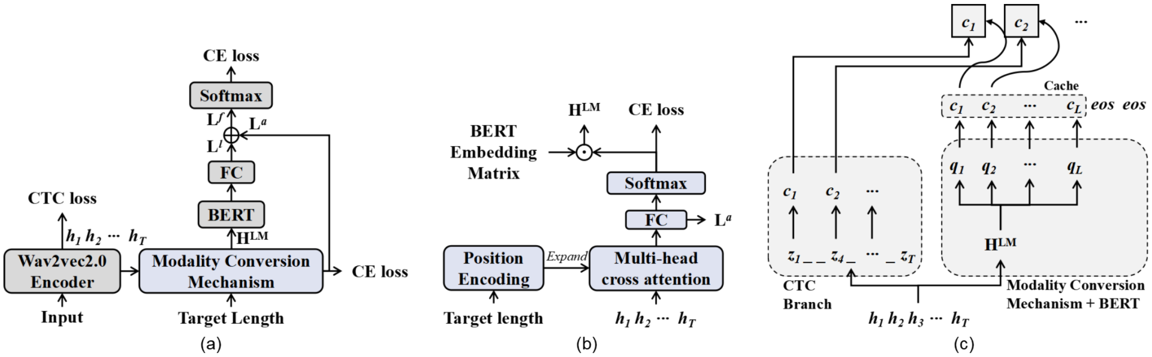

Fig. 1. Illustration of our proposed methods. (a) represents the proposed NAR CTC/attention architecture consisting of a wav2vec2.0 encoder and BERT, (b) explains the modality conversion mechanism (MCM), and (c) shows the cache-based CTC/attention joint decoding method.

2. PROPOSED METHODS

We propose a novel NAR CTC/attention ASR architecture that employs CTC/attention multi-task learning during training. The realization of our model is shown in Fig. 1, where FC represents a fully connected layer, and $\odot$ and $\oplus$ denote the dot product and addition operations, respectively. Our model contains a pre-trained wav2vec2.0 encoder as well as a pre-trained BERT and includes a MCM.

2.1. Wav2vec2.0 Encoder

In our proposed NAR CTC/attention model, the CTC branch plays an important role. It is expected to accurately predict a target length for the modality conversion mechanism and consider all the time boundaries during cache-based CTC/attention joint decoding. Given the promising results recently achieved by wav2vec2.0 [9, 17] based on the CTC criterion [18], we select it as our acoustic encoder.

The wav2vec2.0 encoder consists of a convolutional neural network (CNN)-based feature encoder and a Transformer-based context network. As shown in Fig. 1 (a), we use it to extract an acoustic representation $\mathbf{H}^{\mathrm{AC}} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T)^\top \in \mathbb{R}^{T\times d}$ from the raw speech waveform $\boldsymbol{x}$, where d is the representation dimension and T denotes time steps. During training, we apply a CTC branch after the encoder with the CTC objective $\mathcal{L}_{\mathrm{ctc}}$.

2.2. Modality Conversion Mechanism and BERT

Incorporating BERT into an ASR system requires solving the mismatch between speech and text modalities. As shown in Fig. 1 (b), to efficiently connect acoustic and linguistic representations, we design a novel structure named modality conversion mechanism (MCM), which consists of the following two parts:

First, to bridge the length gap between the acoustic and linguistic representations, we take advantage of the monotonic property of the CTC branch and use it with a greedy search to predict a target length $\hat{L}$. To achieve efficient training, we choose a ground truth L as our target length during training. We implement position encoding over this target length L and broadcast its dimension to be consistent with the acoustic representation $\mathbf{H}^{\mathrm{AC}}$. This broadcast is then fed into a cross attention block together with the $\mathbf{H}^{\mathrm{AC}}$ to capture an alignment between the acoustic representation and its corresponding text. This alignment converts the T-length $\mathbf{H}^{\mathrm{AC}}$ to an L-length embedding:

$$\mathbf{H} = \mathrm{Attention}(\mathbf{H}^{\mathrm{PE}}, \mathbf{H}^{\mathrm{AC}}, \mathbf{H}^{\mathrm{AC}}), \tag{1}$$

where $\mathbf{H}^{\mathrm{PE}} \in \mathbb{R}^{L\times d}$ denotes the positional embedding, and $\mathbf{H} \in \mathbb{R}^{L\times d}$ refers to the obtained acoustic embedding.

Second, we attempt to solve the mismatch between the acoustic embedding $\mathbf{H}$ and the linguistic token embedding of BERT. BERT uses very large character/BPE tokens compared with the ASR system, and its adjustment is difficult. Previous works [11, 12] employ the token embedding of BERT as the learning target of $\mathbf{H}$, but these approaches require a significant increase in the training iterations. Further, BERT's token embedding is available during training and decoding, making it unnecessary to regard it as a learning target. Therefore, instead of adopting these methods [11, 12], we directly perform a dot product operation on the token embedding of BERT. Before using BERT, we first generate a preliminary prediction based on the obtained acoustic embedding:

$$\mathbf{L}^a = \mathbf{H}\mathbf{W}, \tag{2}$$

where matrix $\mathbf{W} \in \mathbb{R}^{d\times V}$ are trainable and $\mathbf{L}^a \in \mathbb{R}^{L\times V}$ denotes the logits of acoustic embedding. Letting V denote the vocabulary size of the ASR system, we then have $\mathbf{M}_{\mathrm{BERT}} \in \mathbb{R}^{V\times d}$, a matrix that contains the token embedding of BERT shared with the ASR system, arranged in the order of tokens list in the ASR system[2]. We can then achieve an efficient modality conversion through dot product:

$$\begin{aligned} \mathbf{W}^a &= \mathrm{Softmax}(\mathbf{L}^a), && (3)\\ \mathbf{H}^{\mathrm{LM}} &= \mathbf{W}^a \cdot \mathbf{M}_{\mathrm{BERT}}, && (4) \end{aligned}$$

where $\mathbf{W}^a \in \mathbb{R}^{L\times V}$ and each element $\mathbf{W}^a_{ij}$ denotes the weight assigned to the j-th BERT token embedding at the i-th step. We then obtain the linguistic representation $\mathbf{H}^{\mathrm{LM}} \in \mathbb{R}^{L\times d}$, which is fed into the BERT encoder. We use a cross-entropy (CE) criterion $\mathcal{L}_{ce1}$ to encourage the $\mathbf{L}^a$ after softmax to generate correct predictions before feeding it into the BERT.

We choose to incorporate BERT [19] into our ASR system due to its powerful text processing capabilities enabled by its embedding layer and a multi-layer Transformer encoder [19, 20]. As shown in Fig. 1 (a), we apply a FC layer on the top of the BERT output to obtain logits of the linguistic representation $\mathbf{L}^l$.

2.3. Training Objective

To utilize both acoustic and linguistic knowledge, we add the $\mathbf{L}^a$ and $\mathbf{L}^l$ together to get the final logits $\mathbf{L}^f$:

$$\mathbf{L}^f = \alpha\mathbf{L}^l + \mathbf{L}^a, \tag{5}$$

[2] It should be noted that this approach works well with character-based BERT for Mandarin, but it needs extensions for BPE tokens used in English as the ASR system and pre-trained LM system have different BPE models, which are related to the vocabulary size of BPE thus cannot be shared.

Algorithm 1 Cache-based joint CTC/attention decoding

```
 1: $\Omega_0 \leftarrow \emptyset$
 2: $\hat{\Omega} \leftarrow \emptyset$
 3: $\hat{L} \leftarrow$ target length generated by CTC greedy search
 4: for $l = 1 \cdots L_{\max}$ do
 5:     $\Omega_l \leftarrow \emptyset$
 6:     while $\Omega_{l-1} \neq \emptyset$ or l = 1 do
 7:         $g \leftarrow \mathrm{HEAD}(\Omega_{l-1})$
 8:         DEQUEUE($\Omega_{l-1}$)
 9:         for each $c \in \mathcal{V} \cup \langle eos \rangle$ do
10:             $h \leftarrow g \cdot c$
11:             if $l > \hat{L}$ then
12:                 $\alpha_f(c =< eos >, x) \leftarrow \log(0.9)$
13:                 $\alpha_f(c \neq< eos >, x) \leftarrow \log(0.1/V)$
14:             $\alpha(h, x) \leftarrow \mu\alpha_{ctc}(h, x) + (1 - \mu)\alpha_f(c, x)$
15:             if $c =< eos >$ then
16:                 ENQUEUE($\hat{\Omega}, h$)
17:             else
18:                 ENQUEUE($\Omega_l, h$)
19:                 if $|\Omega_l| > beamWidth$ then
20:                     REMOVEWORST($\Omega_l$)
21:     if ENDDETECT($\hat{\Omega}, l$) $= \mathrm{true}$ then
22:         break
23: return $\arg\max_{h \in \hat{\Omega}} \alpha(h, x)$
```

where α is a tunable hyper-parameter. We calculate the CE loss $\mathcal{L}_{\mathrm{ce}}$ over the final output $\mathbf{L}^f$ to provide further supervision.

During joint training, the loss function is defined by:

$$\mathcal{L} = (1 - \beta)(\lambda_1 \mathcal{L}_{\mathrm{ce}} + \lambda_2 \mathcal{L}_{\mathrm{ce1}}) + \beta \mathcal{L}_{\mathrm{ctc}}, \tag{6}$$

where λ_1, λ_2, and β are tunable hyper-parameters. $\mathcal{L}_{\mathrm{ctc}}$ and $\mathcal{L}_{\mathrm{ce1}}$ are defined in Sections 2.1 and 2.2, respectively.

2.4. Cache-based CTC/attention Joint Decoding

We employ the CTC branch to generate a target length $\hat{L}$, which enables the BERT to predict tokens in parallel. Inspired by [21], we design a cache-based CTC/attention joint decoding method to further improve the recognition accuracies, as shown in Fig. 1 (c). Unlike the conventional CTC/attention joint decoding [21], we precompute the attention-based score up to $\hat{L}$ length at one step instead of incremental computation, thus enabling our ASR model to achieve high inference speed [3].

The details of our method are shown in Algorithm 1, where Ω_l and $\hat{\Omega}$ denote queues that accept partial hypotheses of length l and complete ones, respectively. $\mathcal{V}$ is the token vocabulary with V size. Each $c \in \mathcal{V} \cup \langle eos \rangle$ is appended to a former partial hypothesis g given by Ω_{l-1}. The attention-based scores of $\hat{L}$ length sequence are calculated at one step and then stored in the cache. When $l > \hat{L}$, we predict $\langle eos \rangle$ with a 0.9 probability. We score each extended hypothesis h=$g \cdot c$ by a CTC prefix score and attention-based scores stored in the cache, as shown in line 14. This hypothesis is stored in either Ω_l or $\hat{\Omega}$ based on the value of c. While $c \neq \langle eos \rangle$, h is added to Ω_l, and Ω_l 's size is then compared with the beam width for pruning by REMOVEWORST($\cdot$). We also utilize an additional function ENDDETECT($\hat{\Omega}, l$) [21] to determine whether to stop the procedure before l reaches $L_{\max}$. If finding $h \in \hat{\Omega}$ with higher scores is almost impossible as l increases, ENDDETECT($\hat{\Omega}, l$) returns $true$.

We apply softmax to $\mathbf{L}^f$ to get an attention-based probability $p_f(c|x)$. We also cumulatively calculate a prefix CTC probability $p_{\mathrm{ctc}}(h, \cdots|x)$. The scores are then computed in log domain:

$$\begin{cases} \alpha_{\mathrm{ctc}}(h, x) = \log p_{\mathrm{ctc}}(h, \cdots|x) \\ \alpha_f(c, x) = \log p_f(c|x) \end{cases} \tag{7}$$

3. EXPERIMENTS

3.1. Corpus

We evaluate our proposed NAR CTC/attention model on the Mandarin AISHELL-1 [22] and English Switchboard [23] corpora. We also use an unlabeled speech training set from the AISHELL-2 corpus to pre-train a Mandarin wav2vec2.0 base model [24].

3.2. Model Descriptions

The ESPnet2 toolkit [25] is used to build our wav2vec2.0 CTC baseline and our proposed NAR CTC/attention model. We employ raw speech as the acoustic input. For Mandarin, we use 4230 Chinese characters with 3 non-verbal symbols: blank, unknown-character, and sos/eos as the modeling units. For English, we utilize 2000 modeling units, including 1997 BPE [16] units and 3 non-verbal symbols.

Our wav2vec2.0 CTC baseline consists of a base wav2vec2.0 [9] encoder and a FC layer as the classifier with a size of 4233 and 2000 for Mandarin and English tasks respectively. If not specified, wav2vec2.0 refers to Mandarin wav2vec2.0 in the Mandarin task. As for our proposed NAR CTC/attention model, the wav2vec2.0 encoder and CTC branch are the same as those in the baseline model, and its MCM contains one-layer multi-head cross attention with 768 model dimensions and 4 heads. The FCs in Fig. 1(a) and Fig. 1(b) are of the same size as the CTC branch. The pre-trained Mandarin BERT (i.e., bert-base-chinese) and the English RoBERTa [4] [26] (i.e., roberta-base) are both provided by Huggingface Transformer Library [27], and the English wav2vec2.0 base model is provided by Fairseq [28]. For the first 5000 and 25000 steps, the parameters of both the wav2vec2.0 and the BERT/RoBERTa are fixed. In our embedding layer and the first 3 and 6 Transformer layers of BERT and RoBERTa, the parameters are always fixed. We set both the λ_1 and λ_2 in Eq. (6) to 0.5. The β in Eq. (6) and α in Eq. (5) are set to 0.3. Real time factor (RTF) is measured on the test set of AISHELL-1 using a P100 GPU.

Following the ESPnet2 recipe [25], we fine-tune the Mandarin BERT and English RoBERTa with subsequent masks as external AR-based LMs for the baseline. During inference, we set the beam size as 10 and the weight of BERT/RoBERTa LM is 0.3 when used.

3.3. Main Results

We compare the performance of our proposed NAR CTC/attention model with the strong wav2vec2.0 CTC baseline and other AR/NAR systems. The results are shown in Table 1, where W2v2 denotes the Mandarin wav2vec2.0. It should be noted that when the CTC baseline model uses an external BERT LM during one-pass decoding with beam search, we classify it as an AR system because the BERT is set to a unidirectional AR structure during the fine-tuning stage. And our NAR CTC/attention model employs greedy search during inference. The results show that our proposed NAR CTC/attention

[3] Strictly speaking, this is not a NAR decoding, but it is used to further improve the proposed model with efficient computation.

[4] RoBERTa is a variant of BERT, and we choose it for our English tasks because its modeling unit is BPE and thus more suitable for our English tasks.

Table 1. The character error rates (CER) (%) of different AR/NAR ASR models on AISHELL-1 corpus.

Model	Epoch	Dev	Test
Autoregressive ASR			
resGSA-Transformer [29]	50	5.4	5.9
ESPnet (Transformer) [25]	50	5.9	6.4
ESPnet (Conformer) [25]	50	4.4	4.7
W2v2 CTC baseline + external BERT LM	20	4.4	4.7
Non-autoregressive ASR			
NAR-Transformer [4]	200	5.3	5.9
A-FMLM [6]	50	6.2	6.7
TSNAT-Big+Two Step Inference [7]	100	5.1	5.6
D-Att shared Enc [30]	50	–	6.5
NAR-BERT-ASR [11]	130	4.9	5.5
CASS-NAT [31]	90	5.3	5.8
LASO-big with BERT [32]	130	5.2	5.8
W2v2 CTC baseline	20	4.8	5.3
Proposed NAR CTC/attention	**20**	**4.1**	**4.5**

Table 2. The CER (%) of our NAR CTC/attention system with different decoding style on AISHELL-1 corpus.

Model	Wav2vec2.0 Encoder	Decoding Style	Dev	Test	RTF
Baseline	English	Greedy search	6.1	6.5	0.016
+ BERT LM	English	Beam search	5.4	5.7	8.098
Baseline	Mandarin	Greedy search	4.8	5.3	**0.015**
+ BERT LM	Mandarin	Beam search	4.4	4.7	8.098
Proposed	English	Greedy search	5.1	5.8	0.044
Proposed	English	Cache CTC/Att	4.8	5.1	0.665
Proposed	Mandarin	Greedy search	4.1	4.5	0.045
Proposed	Mandarin	Cache CTC/Att	**4.0**	**4.3**	0.665

Table 3. The word error rate (WER) (%) of different AR/NAR ASR models on Switchboard corpus.

Model	Epoch	Dev	Eval2000
Autoregressive ASR			
ESPnet (Transformer) [25]	100	–	12.9
ESPnet (Conformer) [25]	150	–	**10.4**
W2v2 CTC baseline + RoBERTa LM	25	**10.0**	10.8
Non-autoregressive ASR			
W2v2 CTC baseline	25	11.3	12.4
Proposed NAR CTC/attention	25	10.6	11.9

model greatly outperforms the strong wav2vec2.0 CTC baseline, yielding 15.1% relative CER reduction.

Furthermore, even when a strong BERT-based AR LM is employed for the baseline to relax its conditional independence assumption, the results show that our proposed NAR CTC/attention model still achieves better recognition accuracy in addition to maintaining the advantage of the NAR model's fast inference speed, which is not achieved by previous works [12, 33]. Comparing with other AR/NAR systems, our NAR model greatly improves over previous NAR systems with much fewer training epochs and it even exceeds the AR-based systems. Therefore, it can be concluded that our NAR CTC/attention model achieves impressive performance with high training efficiency.

3.4. Ablation Studies on Cache-based CTC/attention Decoding

In the main experiments, we employ the greedy search for our NAR model. In this section, we conduct ablation studies to verify the effectiveness of our cache-based CTC/attention joint decoding. The results are shown in Table 2, where Baseline represents our wav2vec2.0 CTC baseline and Cache CTC/Att denotes the cache-based CTC/attention joint decoding (i.e. beam search).

It can be seen from the results that the conclusions are the same whether we use the Mandarin or the English wav2vec2.0 encoder: 1) Our proposed NAR model outperforms the baseline even with vanilla attention-based greedy search. 2) After using the cache-based CTC/attention joint decoding method, our model has been further improved, especially on the English wav2vec2.0 encoder. 3) Although the cache-based CTC/attention joint decoding is slower compared to greedy search, it is significantly faster than the beam search with an AR-based external BERT LM, as our model only needs to be run for one time during inference.

In addition, the improvement on the English wav2vec2.0 encoder indicates that our proposed method may also help assist ASR tasks in low-resource languages based on cross-lingual pre-training.

3.5. Experimental Results on Alphabetic Language

Previous experiments have verified the effectiveness of our proposed model on Mandarin (i.e. logographic language) tasks. However, if we consider applying our method to alphabetic languages (e.g., English), it raises a problem regarding the vocabulary size of BPE, as an ASR system and BERT are likely to have different output units. For example, given the sentence ”*good weather*”, the tokenized input for RoBERTa is ”*_good _weather*”, but the corresponding input for the ASR system is ”*_good _wea ther*”. The BPE vocabulary size of RoBERTa is often word-level, which makes it too large and sparse for the ASR system to train on, especially for the CTC branch. A simple solution is that pre-training RoBERTa with the same BPE size as the ASR system. However, it is not always feasible due to the requirement of large computational resources.

To solve this challenge, first, since BPE units like ”*_wea*” and ”*ther*” also exist in the token list of RoBERTa, we can directly convert these units to their corresponding index in RoBERTa's token list before entering into the RoBERTa model. Second, we can also use CTC greedy search results as input to the MCM during training and decoding to avoid the length mismatch.

We choose the English Switchboard [23] corpus to evaluate our NAR model on alphabetic language. The results in Table 3 show that our method outperforms the strong wav2vec2.0 CTC baseline and AR-based Transformer system. However, after using a RoBERTa-based AR LM, the baseline surpasses our NAR model.

4. CONCLUSION

In this paper, we proposed a NAR CTC/attention model that fully utilizes the pre-trained wav2vec2.0 and BERT. We also designed a novel modality conversion mechanism (MCM) to efficiently bridge the gap between speech and text modalities. During inference, we made use of a CTC branch to generate a target length, which enables BERT to process tokens in parallel. Furthermore, we proposed a cache-based CTC/attention joint decoding method to further improve the recognition accuracy while keeping the decoding speed fast. The experimental results showed that the proposed model improves over our wav2vec2.0 CTC baseline and other NAR models.

5. REFERENCES

[1] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[2] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[3] Y. Higuchi, H. Inaguma, S. Watanabe, T. Ogawa, and T. Kobayashi, "Improved mask-CTC for non-autoregressive end-to-end ASR," in *Proc. ICASSP*, 2021, pp. 8363–8367.

[4] X. Song, Z. Wu, Y. Huang, C. Weng, D. Su, and H. Meng, "Non-autoregressive Transformer ASR with CTC-enhanced decoder input," in *Proc. ICASSP*, 2021, pp. 5894–5898.

[5] Z. Tian, J. Yi, J. Tao, Y. Bai, S. Zhang, and Z. Wen, "Spike-triggered non-autoregressive Transformer for end-to-end speech recognition," in *Proc. Interspeech*, 2020, pp. 5026–5030.

[6] N. Chen, S. Watanabe, J. Villalba, P. Zelasko, and N. Dehak, "Non-autoregressive Transformer for speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 121–125, 2021.

[7] Z. Tian, J. Yi, J. Tao, Y. Bai, S. Zhang, Z. Wen, and X. Liu, "TSNAT: Two-step non-autoregressvie transformer models for speech recognition," *arXiv preprint abs:2104.01522*, 2021.

[8] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, 2014, pp. 1764–1772.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12449–12460.

[10] K. Deng, S. Cao, and L. Ma, "Improving accent identification and accented speech recognition under a framework of self-supervised learning," in *Proc. Interspeech*, 2021, pp. 1504–1508.

[11] F. Yu and K. Chen, "Non-autoregressive transformer-based end-to-end ASR using BERT," *arXiv preprint abs:2104.04805*, 2021.

[12] C. Yi, S. Zhou, and B. Xu, "Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 788–792, 2021.

[13] W. Huang, C. Wu, S. Luo, K. Chen, H. Wang, and T. Toda, "Speech recognition by simply fine-tuning BERT," in *Proc. ICASSP*, 2021, pp. 7343–7347.

[14] L. Dong and B. Xu, "CIF: Continuous integrate-and-fire for end-to-end speech recognition," in *Proc. ICASSP*, 2020, pp. 6079–6083.

[15] L. Dong, C. Yi, J. Wang, S. Zhou, S. Xu, X. Jia, and B. Xu, "A comparison of label-synchronous and frame-synchronous end-to-end models for speech recognition," *arXiv preprint arXiv:2005.10113*, 2020.

[16] P. Gage, "A new algorithm for data compression," *The C Users Journal*, vol. 12, no. 02, pp. 23–38, 1994.

[17] K. Deng, S. Cao, Y. Zhang, and L. Ma, "Improving hybrid ctc/attention end-to-end speech recognition with pretrained acoustic and language models," in *Proc. ASRU*, 2021, pp. 76–82.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[20] K. Deng, G. Cheng, R. Yang, and Y. Yan, "Alleviating asr long-tailed problem by decoupling the learning of representation and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 340–354, 2022.

[21] T. Hori, S. Watanabe, and J. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. ACL*, 2017, pp. 518–529.

[22] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, 2017, pp. 1–5.

[23] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, pp. 517–520.

[24] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming mandarin ASR research into industrial scale," *arXiv preprint abs:1808.10583*, 2018.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, and N. Chen, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:2008.03822*, 2019.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*, 2020, pp. 38–45.

[28] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. NAACL-HLT: Demonstrations*, 2019.

[29] C. Liang, M. Xu, and X.-L. Zhang, "Transformer-based end-to-end speech recognition with residual gaussian-based self-attention," in *Proc. Interspeech*, pp. 2072–2076.

[30] C. Gao, G. Cheng, J. Zhou, P. Zhang, and Y. Yan, "Non-autoregressive deliberation-attention based end-to-end ASR," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[31] R. Fan, W. Chu, P. Chang, and J. Xiao, "CASS-NAT: CTC alignment-based single step non-autoregressive transformer for speech recognition," in *Proc. ICASSP*, 2021, pp. 5889–5893.

[32] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang, "Fast end-to-end speech recognition via non-autoregressive models and cross-modal knowledge transferring from BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1897–1911, 2021.

[33] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2.0 to speech recognition in various low-resource languages," *arXiv preprint abs:2012.12121*, 2020.