CASS-NAT: CTC Alignment-based Single Step Non-Autoregressive Transformer for Speech Recognition

Ruchao Fan[2], Wei Chu[1], Peng Chang[1], Jing Xiao[1]

fanruchao@g.ucla.edu

[1]PAII Inc., USA

[2]Dept. of Electrical and Computer Engineering, University of California, Los Angeles, USA

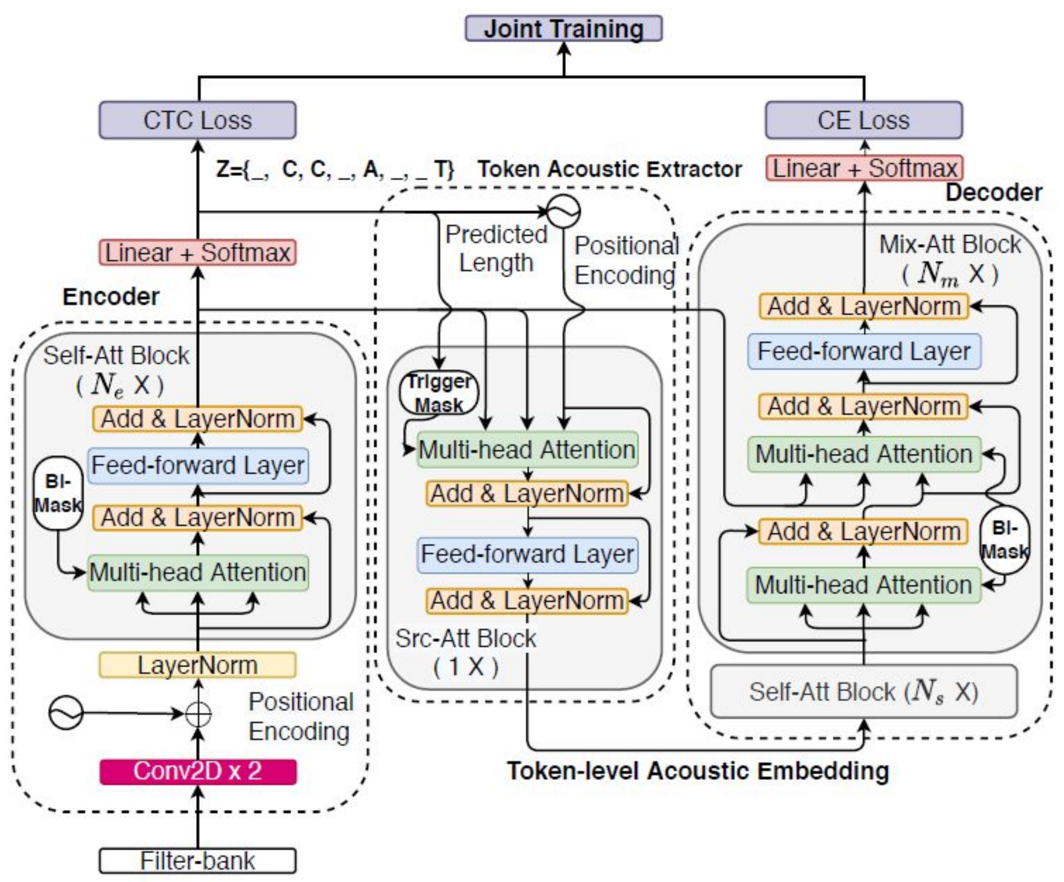
Introduction

- In recent years, autoregressive transformer (AT) achieves great success for automatic speech recognition.
- However, the autoregressive mechanism in transformer decoder slows down the inference speed.
- Non-autoregressive transformer (NAT) was proposed for parallel generation to accelerate the inference.
- **Limitations** for current NAT models:
  - Iterative NAT still needs multiple generation steps, which cannot fully exploit the potential of NAT.
  - Single step NAT extracts incomplete acoustic representations, thus the performance is worse than AT.
- **Novel Contributions:** 1): We propose a novel framework, CTC alignment-based single step NAT (CASS-NAT). 2) An error-based sampling alignment strategy during inference is further proposed to improve the WER performance.
- The proposed CASS-NAT achieves WERs of 3.8%/9.1% on Librispeech test clean/other dataset without an external LM, and a CER of 5.8% on Aishell1 Mandarin corpus.
- Compared to AT baseline, the CASS-NAT has a performance reduction on WER, but is 51.2x faster in terms of RTF.

Proposed CASS-NAT

1. Framework

Figure 1. The proposed CASS-NAT architecture.



- **Encoder:** extract high level representation H
- **CTC:** optimize the CTC alignment that offers auxiliary information for token-level acoustic embedding extraction.
  - Time boundary for each token (trigger mask)
  - Number of tokens for decoder input (NoT)
  - Fix mapping rule when obtaining trigger mask
  - For example, first index of each token is end boundary

Alignment: $Z = \{\_, C, C, \_, A, \_, \_, T, \_\}$

Trigger mask: $[0, 0, 1, 1, 1, 0, 0, 0, 0]$

- **Token-acoustic extractor:**
  - 1 self-attention block
  - Q: sinusoidal positional embedding with NoT
  - K, V: encoder output H
  - Mask: trigger mask from CTC alignment
- **Decoder:**
  - self-att block (not considering H)
  - mix-att block (considering H)
- **CE:** cross entropy loss to optimize the final WER.

2. Training Criterion

- Given $X = \{x_1, x_2, \dots, x_T\}$ and $Y = \{y_1, y_2, \dots, y_U\}$, the CTC alignment Z is introduced, the objective function is:

$$\log P(Y|X) = \log \mathbb{E}_{Z|X}[P(Y|Z, X)], \quad Z \in q.$$

where q is the set of alignments which can be mapped to Y.
- Maximum approximation is applied to reduce computation:

$$\log P(Y|X) \geq \mathbb{E}_{Z|X}[\log P(Y|Z, X)]$$

$$\approx \max_Z \log \prod_{u=1}^{U} P(y_u | z_{t_{u-1}+1:t_u}, x_{1:T})$$

where t_u is the end boundary of token u.
- The final objective function is:

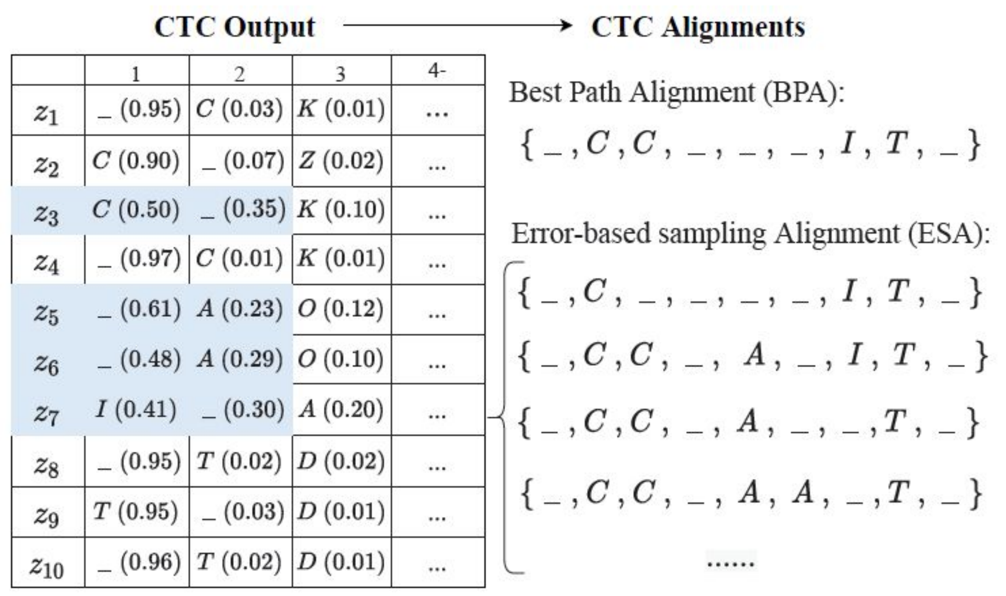
$$L_{\text{joint}} = \max_Z \log \prod_{u=1}^{U} P(y_u | z_{t_{u-1}+1:t_u}, X) + \lambda \cdot \log \sum_{Z \in q} \prod_{i=1}^{T} P(z_i|X)$$

- Semantic modelling is relied on decoder with token-level acoustic embedding as input (assumption).

3. Inference strategy

- Ideally, oracle alignment (obtained using ground truth)
- Best path alignment (BPA)
  - Pro: one step inference Con: alignment is not accurate.
- Beam search alignment (BSA)
  - Pro: alignment is accurate Con: beam search, slow

Figure 2. Illustration of error-based alignment sampling method.



- Error-based sampling alignment (ESA)
  - Sampling over CTC output space is time consuming.
  - Sampling based on best path alignment is easier.
  - If the probability is lower than the threshold (0.7), consider sampling **within top2 tokens**.
  - It is possible to sample alignments with the same number of tokens as oracle alignment.
  - Use AT or LM for ranking different sampled alignments based on decoder outputs.

Experiment - Librispeech

1. Experimental Setup

- Input and output:
  - 80-dim log-mel filter bank features
  - Every 3 frames are concat to form a 240-dim input.
  - Output: 5k word-pieces obtained by SentencePiece [24].
- Model
  - 2 CNNs: 64 filter, kernel size 3, stride 2
  - AT baseline: $N_e = 12, N_d = 6, d_{FF} = 2048, H = 8, d_{MHA} = 512$
  - CASS-NAT:
    - 1-layer token-acoustic extractor
    - Decoder: 3 self-att blocks and 4 mix-attn blocks
  - SpecAug, Label smoothing, **Encoder initialization**

2. Result

Table 1. A comparison of accuracy and speed of Autoregressive Transformer (AT) and non-AT (NAT) algorithms on Librispeech.

		Type	WER (%)				RTF
			dev-clean	dev-other	test-clean	test-other	test-clean
Without LM							
RETURNN [1]		AT	4.3	12.9	4.4	13.5	-
ESPNet		AT	3.2	8.5	3.6	8.4	-
AT (ours)		AT	3.4	8.5	3.6	8.5	0.562
Imputer [16]		NAT	-	-	4.0	11.1	-
CASS-NAT	BPA	NAT	4.4	10.6	4.5	10.7	0.005
	BSA	NAT	3.9	9.6	3.9	9.6	0.655
	ESA	NAT	3.7	9.2	3.8	9.1	0.011
With LM							
RETURNN [1]		AT	2.6	8.4	2.8	9.3	-
ESPNet [25]		AT	2.3	5.6	2.6	5.7	-
AT (ours)		AT	2.5	5.7	2.7	5.8	-
CASS-NAT	ESA	NAT	3.3	8.0	3.3	8.1	-

- ESA decoding reduces WER significantly compared to both BPA and BSA and has a moderate increase of RTF over BPA.
- When no external LM is used, CASS-NAT is 51.2x faster than AT in terms of RTF, while has ~6% relative WER reduction.
- When using an external LM, the gap of WER between AT baselines and CASS-NAT is increasing.

3. Analyse of the performance

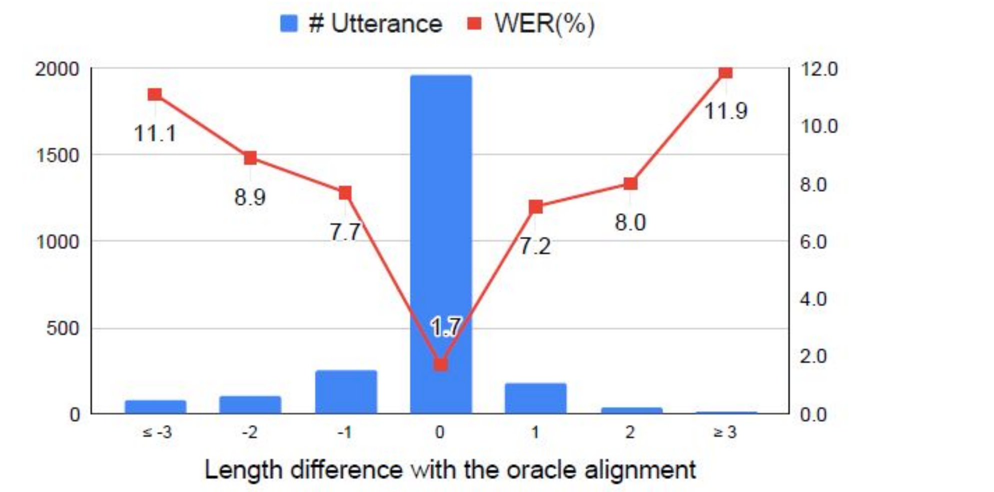
- Mismatch rate (MR): **Deletion and insertion errors** compared to the oracle alignment. Substitution errors do not affect token-level acoustic embedding extraction.
- Length prediction error rate (LPER): Taking the alignment as output and removing blank and repetitions, the ratio of utterances with different length compared to ground truth.

Table 2. A comparison of different alignment generation methods in CASS-NAT decoding without LM.

Alignment	S	WER (%)		MR (%)		LPER (%)	
		test-clean	test-other	test-clean	test-other	test-clean	test-other
Oracle	n/a	2.3	5.8	n/a	n/a	n/a	n/a
BSA	n/a	3.9	9.6	2.2	5.8	27.9	48.3
BPA	n/a	4.5	10.7	2.1	4.9	31.0	51.8
ESA	10	3.9	9.4	2.9	5.7	26.4	42.8
	50	3.8	9.1	3.1	5.8	25.3	41.9
	100	3.8	9.0	3.0	5.8	25.1	41.8
	300	3.8	9.0	3.1	5.8	25.1	41.9

- With oracle alignment, **the lower bound of WER** can be 2.3% for test-clean set.
- For ESA, no further gains are observed when the number of sampled alignments is over 50.
- Correct estimation of the decoder input length is more important for NAT.

Figure 3. Length prediction error distributions and corresponding WERs with ESA(s=50) decoding on the test-clean dataset.



- The WER can be lowered than 2% for the utterances with correct token number estimation.
- The figure shows the importance of length prediction accuracy on the encoder side again.

Experiment - Aishell1

1. **Experimental Setup**

   The setup is almost the same as that for librispeech except:
   - 4230 Chinese characters as output from training set.
   - $N_e = 6$
   - Additionally use **speed perturbation.**

2. **Result**

Table 3. A comparison of WERs on Aishell1 with the existing works.

CER(%)	NAT Type	Dev	Test
AT (ours)	n/a	5.5	5.9
Masked-NAT [13]	iterative	6.4	7.1
Insertion-NAT [15]	iterative	6.1	6.7
ST-NAT [18]	single step	6.9	7.7
LASO [17]	single step	5.8	6.4
CASS-NAT (ours)	single step	5.3	5.8

- Our proposed CASS-NAT is better than previous work.
- CASS-NAT is slightly better than AT, which is promising.
- Our framework general well according to the AT baseline.

Conclusion

- This work presents a novel CASS-NAT framework
  - CTC alignment is used as auxiliary information to extract token-level acoustic embedding.
  - The word embedding in AT is replaced with acoustic embedding for parallel generation.
  - Viterbi-alignment is used for training.
  - Error-based sampling alignment is proposed for inference.
- The importance of length prediction for decoder input is shown by analyzing the relationships between different alignments with the oracle alignment.
- We decrease the gap between AT and NAT, and maintain the acceleration for NAT.

References

The number is appeared as the same in the paper.