VADOI: Voice-Activity-Detection Overlapping Inference For End-to-End Long-Form Speech Recognition

Jinhan Wang[1], Xiaosu Tong[2], Jinxi Guo[2], Di He[2], and Roland Maas[2]

wang7875@g.ucla.edu

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

[2]Amazon Alexa, USA

Number: 4734

icassp 2022 Singapore

Introduction

- End-to-end (E2E) models have shown great performance on the Automatic Speech Recognition (ASR) task.
- E2E models trained on short training segments do not perform well when decoding long-form speech.
- At the inference stage, overlapping inference (OI) and partial overlapping inference (POI) are proposed to align and concatenate overlapped segments after chopping.
- **Limitations** for OI and POI:
  - 50% overlapping percentage doubles computation cost.
  - OI can not tackle low overlapping percentage due to extra cost from non-overlapped region.
  - POI mitigates the above issue but degrades with low overlapping percentage due to lack of common words.
- **Novel Contributions:** 1): Voice-Activity-Detection Overlapping Inference (VADOI) is proposed to introduce more common words around window boundaries to mitigate alignment confusion. 2): We propose Soft-Match to compensate for mismatch between similar but not identical words to further improve alignment quality.
- VADOI achieves equivalent performance as using 50% overlapping percentage, with 20% computation cost reduction on two simulated long-form datasets.

OI and POI

- **Goal:** Minimize pseudo word error rate (WER) between consecutive segments.
- **OI** Figure 1. Illustration of OI decoding scheme (a) and sample alignment graph (b).

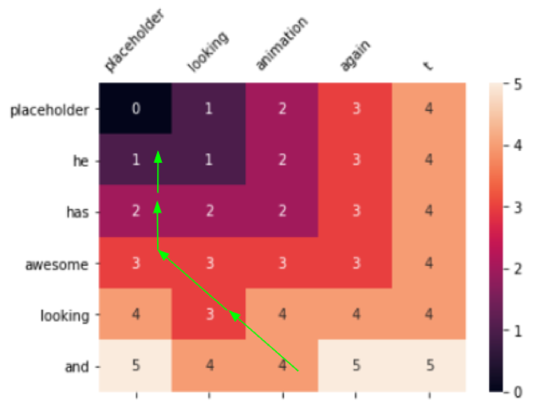
$$C(j,0) = w_{del} \times j, \quad 0 \le j \le n_i \qquad C(0,k) = w_{ins} \times k, \quad 1 \le k \le n_{i+1} \quad (1)$$

$$e_{sub}(j,k) = \begin{cases} w_{match}, & if\ d_i(j) = d_{i+1}(k) \\ w_{sub}, & else \end{cases} \quad (2)$$

$$C(j,k) = min \begin{cases} C(j-1,k) + w_{del}, \\ C(j,k-1) + w_{ins}, \\ C(j-1,k-1) + e_{sub}(j,k), \end{cases} \quad 1 \le j \le n_i, 1 \le k \le n_{i+1}, \quad (3)$$

$$w_{ins}, w_{del}, w_{sub} = 1, \quad w_{match} = 0 \quad (4)$$

(a)



(b)

C(j,k): edit distance between word j (from sentence i) and word k (from sentence i+1)
n_i: length of sentence i   e_sub: matching reward $d_i(j)$: j^{th} word in sentence i
w_{ins}, w_{del}, w_{sub}, w_{match}: corresponding costs

  - **Con:** Non-overlapped region introduces external insertion and deletion errors.
- **POI** Figure 2. Illustration of POI decoding scheme (a) and sample alignment graph (b).

$$C(j,0) = w_{del,non} \times j, \quad 0 \le j \le n_i \qquad C(0,k) = w_{ins} \times k, \quad 1 \le k \le n_{i+1} \quad (5)$$

$$e_{sub}(j,k) = \begin{cases} w_{match}, & if\ d_i(j) = d_{i+1}(k) \\ w_{sub}, & else \end{cases} \quad (6)$$

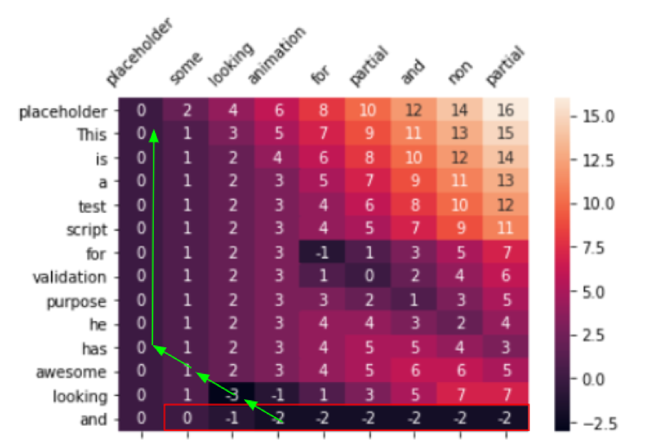
$$C(j,k) = min \begin{cases} C(j-1,k) + w_{del}, \\ C(j,k-1) + w_{ins}, \\ C(j-1,k-1) + e_{sub}(j,k), \end{cases} \quad 1 \le j < n_i, 1 \le k \le n_{i+1}, \quad (7)$$

$$C(j,k) = min \begin{cases} C(j-1,k) + w_{del}, \\ C(j,k-1) + w_{ins,non}, \\ C(j-1,k-1) + e_{sub}(j,k), \end{cases} \quad j = n_i, 1 \le k \le n_{i+1}, \quad (8)$$

$$w_{ins,non}, w_{del,non} = 0, \quad w_{ins}, w_{del} = 2, \quad w_{sub} = 1, \quad w_{match} = -2. \quad (9)$$

(a)

$w_{ins,non}$, $w_{del,non}$: pseudo ins/del cost for non-overlapped region, set to 0.



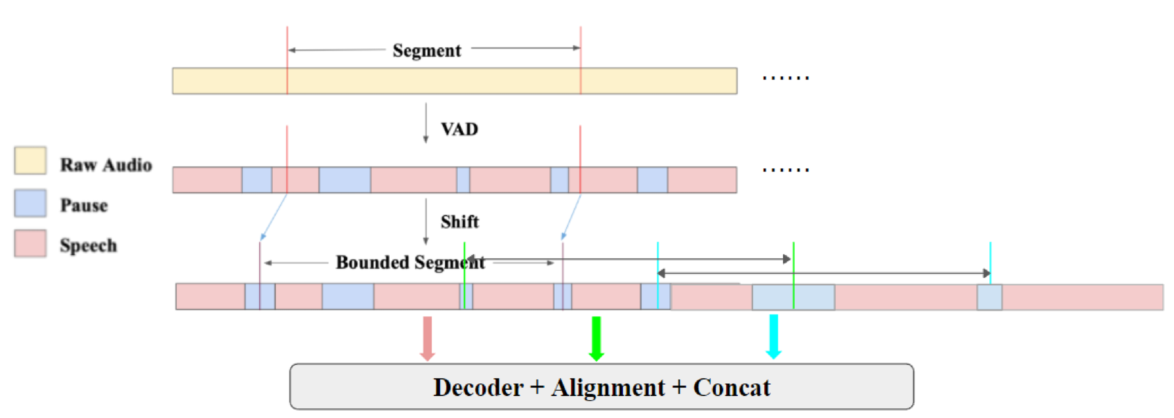
(b)

- **Pros:** 1): Marginal costs are nullified, so lower overlapping percentage is applicable. 2): Instead of word-level, alignment can be done on character-level.
- **Cons:** Lower overlapping percentage degrades performance because of insufficient matching reward/common words.

VADOI

- **Motivation:** Introduce more common words by preventing chopping segments in the middle of a word. Improve alignment quality by mitigating boundary distortion.
- **Proposed VADOI:**

Figure 3. VADOI: Chopping and Shifting



  - Segments generated in first stage pass through a VAD.
  - Starting and end frame are shifted to the closest pause with a length greater than a pre-defined threshold.
    - The existence of overlapped region is guaranteed by restricting frame-shifting distance to be within half of the overlapping region length.
    - If a long-pause is not found within this range, the threshold for length will be cut in half.
    - One special case is the start frame is shifted to right when overlapping percentage is over 40% to prevent triple word-pair.
  - Shifted segments are decoded, aligned and concatenated following POI decoding scheme.

Soft-Match

- **Motivation:** Relax the constraint of Eq.2 and Eq.6, such that similar but not identical words will contribute to a moderate matching reward proportional to similarity.
- **Proposed Soft-Match:**

$$e_{sub} = CER(d_i(j), d_{i+1}(k)) \cdot (w_{sub} - w_{match}) + w_{match}$$

  - Similarity between two words is measured using character error rate (CER), range from 0 (identical) and 1(completely different).
  - CER is projected into a number between w_{sub} and w_{match}.
- **Example:**

Word Pair \ Substitution Cost	wo Soft-Match	Soft-Match
awesome, awesome	-2	-2
looking, booking	1	-1.5714
anime, enemy	1	-0.2

Substitution cost = - matching reward For (looking,booking) and (anime,enemy) pairs, it is more reasonable to assign positive rewards because they are omitted from the same acoustic feature.

Experiment

1. Experimental Setup

- Datasets:
  - Training: 59k hours of mixed public datasets.
  - Testing:
    - MSLT-long: simulated from MSLT with average duration(s) and standard deviation as (121, 5).
    - Lib-Long: concatenated from Librispeech with average duration(s) and standard deviation as (120,3.8).
- Model
  - Input: 64-dim Log-filterbank Energy (LFBE)
  - RNN-T model
    - Encoder: 8x1024 LSTM (layerNorm), 2x16 FLSTM(windows size: 8, stride: 2)
    - Decoder: 2x1024 LSTM
    - Joint Network: feed-forward layer (activation: tanh)
  - SpecAugment, FastEmit Lambda=0.005.
  - Decoding:
    - Segments length: 12s
    - Corresponding costs are set same as Fig.1 and Fig.2.
- Evaluation Protocols:
  - WER
  - Computation Cost:
    - Decoding Time: how many folds needed to decode compared with Baseline/Naive Approach (T)
    - Ovl-Inf Time: Absolute duration for alignment and concatenation (sec/utt)

2. Results (OI and POI)

Table 1. WER(%) and Computation Cost on Various Decoding Schemes on MSLT-Long

WER(%)/Decoding Time/Ovl-Inf Time		OI word	OI char	POI word	POI char
Baseline		20.1/T/NA			
Ovl Percentage	0%	16.4/T/NA			
	50%	13.6/1.87T/0.88	17.0/1.87T/20.5	13.1/1.87T/0.88	13.2/1.87T/20.5
	30%	14.9/1.37T/0.64	54.5/1.37T/14.86	13.3/1.37T/0.64	13.2/1.37T/14.86
	15%	25.2/1.16T/0.53	71.9/1.16T/12.12	13.6/1.16T/0.53	14.1/1.16T/12.12

- WER
  - POI outperforms OI because of better margin conditions.
  - Word-level alignment yields better results than char-level one. For char-level alignment, it might because omitted word not in the vocabulary, which introduces additional sub error.
  - OI is not compatible with char-level alignment because non-overlapped ratio is increased dramatically under char-level.
  - POI has monotonic performance degradation as overlapping percentage decreases.
- Computation Cost:
  - Larger overlapping percentage increases decoding time.
  - Char-level alignment takes significant amount of extra time for alignment and concatenation because exponentially larger dynamic graph size.
  - Word-level POI with 50% overlapping percentage gives the best results but needs additional 87% decoding time.

3. Results (VADOI)

Table 2. WER(%) and Decoding Time of VADOI on MSLT-long.

Exp	VAD	WER(%)	Decoding Time
0%	No	16.40	T
	Yes	14.04	1.05T
50%	No	**13.05**	**1.87T**
	Yes	13.07	2.14T
30%	No	13.27	1.37T
	Yes	**13.02**	**1.50T**
15%	No	13.59	1.16T
	Yes	13.27	1.25T

Table 3. WER(%) and Decoding Time of VADOI on Lib-long.

Exp	VAD	WER(%)	Decoding Time
0%	No	9.70	T
	Yes	7.50	1.06T
50%	No	**6.49**	**1.85T**
	Yes	6.62	2.11T
30%	No	6.79	1.36T
	Yes	**6.58**	**1.48T**
15%	No	7.28	1.13T
	Yes	6.67	1.23T

- Incorporating VADOI under 50% overlapping percentage yields slightly worse performance. We believe it is because additional common words around boundaries are not necessary for the case where overlapped region is sufficiently large.
- With VADOI, equivalent performances are obtained by using 30% as using 50% without VADOI. Computation cost is reduced by 20% relatively on both datasets. Empirical analysis shows that mitigating boundary distortion can greatly improve alignment quality by preventing chopping word in the middle.
- Performance of using VADOI under 15% is not comparable with using 50%, we hypothesis it is because common words are extremely scarce.
- Results with overlapping percentage under 15% are not reported because they start to perform worse than Naive Approach with VADOI.

4. Results (Soft-Match)

Table 4. WER(%) of VADOI with Soft-Match

MSLT	WER(%)	Lib	WER(%)
50%	13.07	50%	6.62
+ Soft	12.99	+ Soft	6.59
30%	13.02	30%	6.58
+ Soft	13.00	+ Soft	6.57
15%	13.27	15%	6.67
+ Soft	13.25	+ Soft	6.63

- Applying Soft-Match constantly yields limited improvement. We suspect it is because the problem expected to be solved by Soft-Match does not prevail.
- The light-weight Soft-Match does not introduce any side effect to the performance and empirical analysis shows it did solve the mis-aligned similar words problem efficiently.

Conclusion

- A comprehensive comparison of OI and POI with various configurations are conducted, and it shows that POI with word-level alignment performs the best.
- We propose VADOI to mitigate boundary distortion, further reduce computation cost. Equivalent performance can be achieved with 20% relative computation cost reduction.
- Soft-Match is proposed to tackle mis-aligned similar words.

References

The number is appeared as the same in the paper.