STREAMING MULTI-SPEAKER ASR WITH RNN-T

Ilya Sklyar*, Anna Piunova*, Yulan Liu
{ilsklyar,piunova,lyulan}@amazon.com

Amazon
Alexa

Paper ID: 2057

Introduction

Problem: multi-speaker overlapped speech recognition by end-to-end (E2E) ASR system

Constraints: low-latency decoding

Proposed model: streaming Recurrent Neural Network Transducer (RNN-T) with multi-output encoder capable to separate and recognize overlapped speech. Training approaches:

- deterministic assignment training (DAT) guided by speaker-order labeling
- permutation invariant training (PIT)

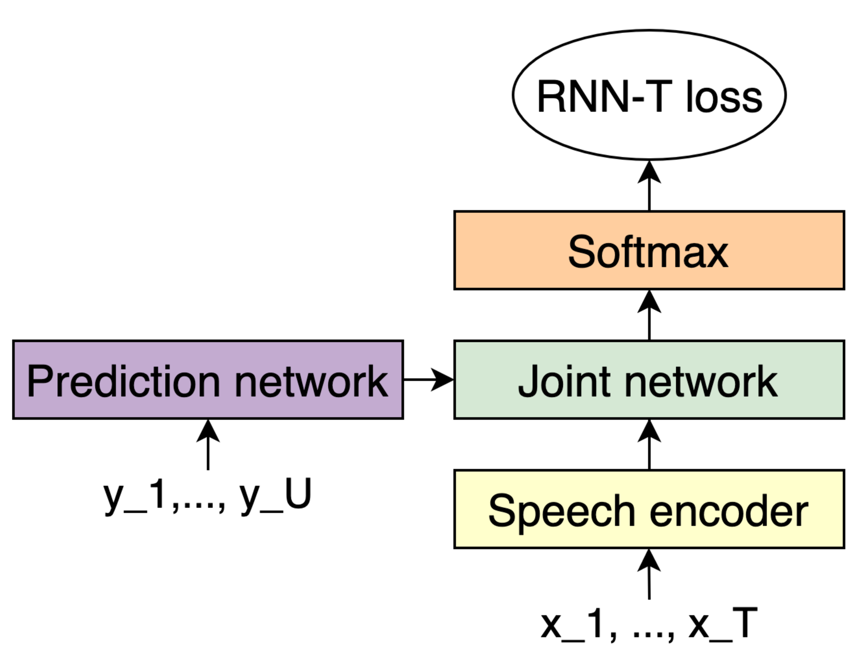
Results: 10.2% WER on 2-speaker LibriSpeechMix, competitive with non-streaming E2E ASR

Data: LibriSpeechMix [12, 13]

- **Train:** artificially mixed LibriSpeech utterances from 960h training set, overall overlap ratio: 28%
- **Dev/Eval:** artificially mixed LibriSpeech utterances from dev/test-clean partitions, overall overlap ratios: 25% (dev) and 24% (eval)

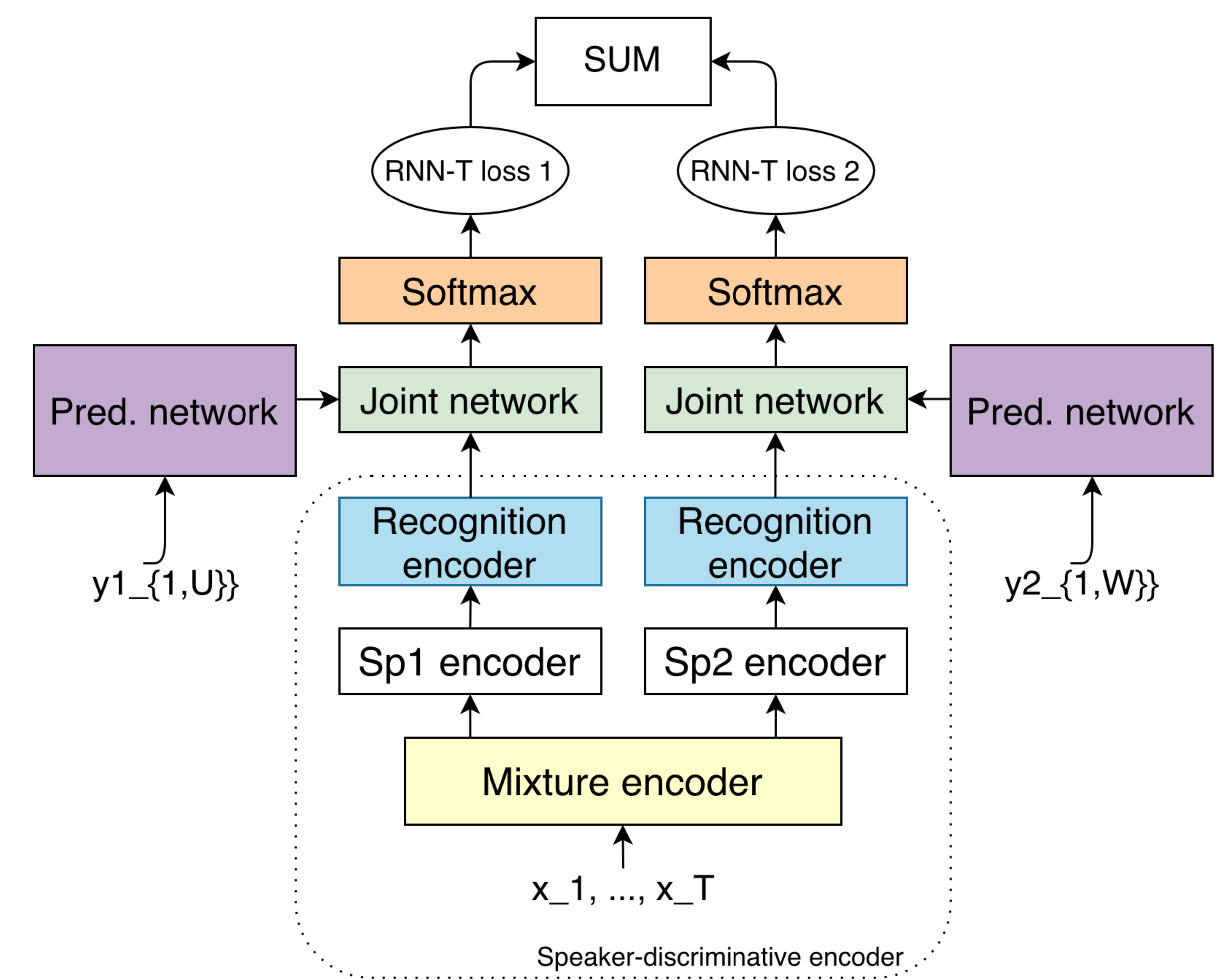
Single-speaker RNN-T

Proposed multi-speaker models are based on a single-speaker RNN-T. Given a sequence of acoustic feature vectors $\mathbf{x} = \{x_1, x_2, ..., x_T\}$ and the corresponding label sequence $\mathbf{y} = \{y_1, ..., y_U\}$ RNN-T estimates conditional probability $P(\mathbf{y}|\mathbf{x})$



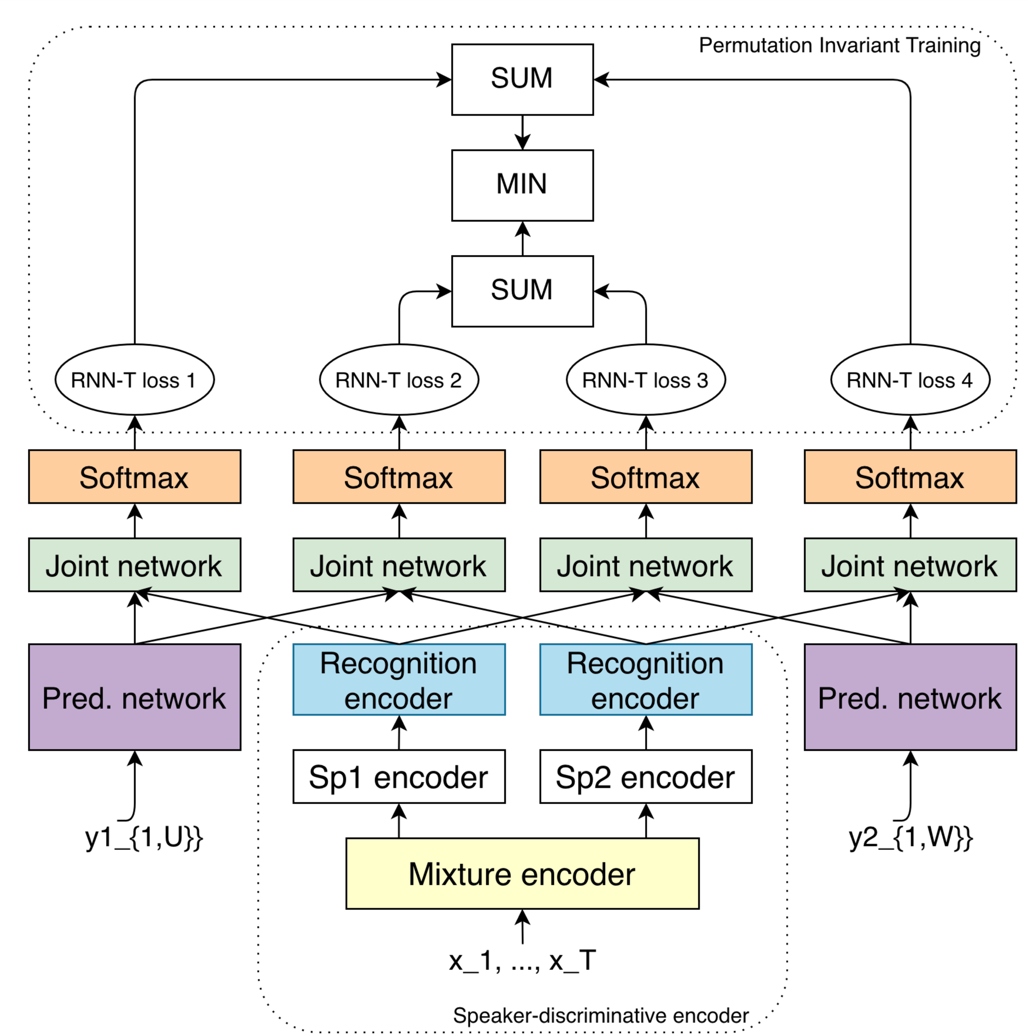
Loss: $\mathcal{L} = -\log P(\mathbf{y}|\mathbf{h})$

Multi-speaker RNN-T: Deterministic assignment training (DAT)



Loss: $$\mathcal{L} = -\sum_s \log P(\mathbf{y}_s|\mathbf{h}_s)$$

Multi-speaker RNN-T: Permutation-invariant training (PIT)



Loss: $$\mathcal{L} = \min_{\pi \in \mathcal{P}} -\sum_s \log P(\mathbf{y}_s|\mathbf{h}_{\pi(s)})$$

References

[12] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition" Interspeech, Oct 2020.

[13] N. Kanda, Y. Gaur, X. Wang, Z. Meng, Z. Chen, T. Zhou, and T. Yoshioka, "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers" Interspeech, Oct 2020.

Evaluation

Optimal edit distance word error rate (WER)

- Ground truths: $\mathbf{R} = [R_1, ..., R_S]$
- Model outputs: $\mathbf{O} = [O_1, ..., O_S]$
- Set of permutations: $\mathcal{P} = \{(1,2), (2,1)\}$

$$WER = \frac{\min_{(i,j)\in\mathcal{P}}(\sum_{i,j}(edits(O_i, R_j))}{\sum_j len(R_j)}$$

Results

Table 1: WER comparison of DAT-MS-RNN-T and PIT-MS-RNN-T on 1,2-spk test sets.

Model	clean	other	2spk	Overall
RNN-T	6.5	15.5	66.3	38.7
DAT-MS-RNN-T	9.2	16.9	11.8	12.4
+ speaker order label	7.7	16.2	11.7	11.8
+multi-style	7.5	15.4	11.0	11.2
PIT-MS-RNN-T	7.9	15.8	10.6	11.2
+multi-style	7.6	15.2	10.2	10.8

Table 2: WER comparison of PIT-MS-RNN-T and non-streaming E2E ASR models on 1,2-spk test sets.

Model	#params	#speakers in training	clean	2spk
PIT-AED[12]	160.7M	1,2	6.7	11.9
SOT-AED [12]	135.6M	1,2,3	4.6	11.2
SOT-AED[13]	135.6M	1,2,3	4.5	10.3
+ speakerID	145.5M	1,2,3	4.2	8.7
PIT-MS-RNN-T	80.9M	1,2	7.6	10.2

Conclusions

- Multi-speaker RNN-T is on-par with non-streaming E2E models reported in literature
- Multi-style training together with explicit speaker-order labeling improve MS-RNN-T generalization to unseen single- and multi-speaker data

*These authors have contributed equally.