Separator-Transducer-Segmenter: Streaming Recognition and Segmentation of Multi-party Speech

Ilya Sklyar, Anna Piunova, Christian Osendorfer

Amazon Alexa

{ilsklyar, piunova, osendorf}@amazon.com

Abstract

Streaming recognition and segmentation of multi-party conversations with overlapping speech is crucial for the next generation of voice assistant applications. In this work we address its challenges discovered in the previous work on multi-turn recurrent neural network transducer (MT-RNN-T) with a novel approach, separator-transducer-segmenter (STS), that enables tighter integration of speech separation, recognition and segmentation in a single model. First, we propose a new segmentation modeling strategy through *start-of-turn* and *end-of-turn* tokens that improves segmentation without recognition accuracy degradation. Second, we further improve both speech recognition and segmentation accuracy through an emission regularization method, FastEmit, and multi-task training with speech activity information as an additional training signal. Third, we experiment with end-of-turn emission latency penalty to improve end-point detection for each speaker turn. Finally, we establish a novel framework for segmentation analysis of multi-party conversations through emission latency metrics. With our best model, we report 4.6% abs. turn counting accuracy improvement and 17% rel. word error rate (WER) improvement on LibriCSS dataset compared to the previously published work.

Index Terms: streaming multi-speaker speech recognition, speech segmentation, separator-transducer-segmenter

1. Introduction

Automatic speech recognition (ASR) of multi-party recordings with overlapping speech has posed a major scientific challenge for many decades [1, 2, 3, 4]. While single-speaker ASR became ubiquitous through applications like voice assistants (Amazon Alexa, Google Home, etc.), its current capability is limited to scenarios with one active speaker at a time. Apart from ASR, speech overlaps also introduce additional challenges to other parts of the traditional speech processing pipeline as speech segmentation and speaker diarization [5].

Multi-speaker ASR problem was attacked before with both independently optimized modules such as speech separation and speech recognition [6, 7, 8, 9, 10], and jointly optimized multi-speaker end-to-end ASR systems [11, 12, 13, 14, 15, 16]. In [17] a joint multi-speaker ASR and speaker change detection system was proposed to tackle speech recognition and segmentation problems simultaneously in the presence of overlapping speech from arbitrary number of speakers. Follow-up work on serialized output training (SOT) [18, 19, 20, 21, 22] extended it to speaker-attributed ASR that can transcribe "who spoke what" in real multi-speaker conversations with a single integrated model.

In parallel, researchers also investigated multi-speaker ASR performance under streaming conditions, which is crucial for applications with minimal latency. In [23] and [24], two conceptually similar streaming multi-speaker ASR systems, multi-speaker recurrent neural network transducer (MS-RNN-T) and

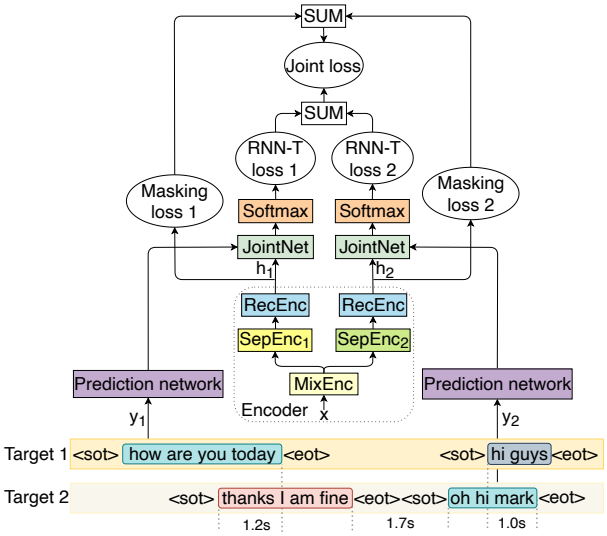

Figure 1: *Separator-Transducer-Segmenter.* `<sot>` *and* `<eot>` *represent start-of-turn and end-of-turn tokens. Model blocks with the same colour have tied parameters, transcripts in the colour-matched boxes belong to the same speaker.*

streaming unmixing and recognition transducer (SURT) were proposed simultaneously to enable time-synchronous decoding of partially overlapping speech from 2 speakers. Both models are based on recurrent neural network transducer (RNN-T) model [25] with implicit speech separation in the encoder and multiple decoding threads (one for each speaker). These approaches were later extended to multi-turn audio processing for any number of speakers in [26] and [27], respectively. SURT was additionally extended to perform speaker identification in [28] and endpoint detection in [29], with limitation to 2-speaker single-turn audio recordings. Recently, an alternative approach to streaming multi-speaker ASR, token-level serialized output training (t-SOT), was proposed in [30], which unified single-speaker and multi-speaker model architectures by mixing tokens from all speakers in one sequence and sorting them by the order of their appearance in the audio.

As reported in [26], a naïve inclusion of *change-of-turn* (`<cot>`) segmentation tag into a streaming multi-speaker ASR model results in a severe underestimation of the real number of turns in the audio. In this work we attack this issue by explicit turn boundary modeling with *start-of-turn* (`<sot>`) and *end-of-turn* (`<eot>`) tokens in a novel separator-transducer-segmenter (STS) model. We further improve recognition and segmentation accuracy of this model through FastEmit [31] and masking loss that penalizes leakage of acoustic information into encoder outputs corresponding to non-active speech frames. Finally, we perform segmentation analysis of STS using start-pointing, end-pointing, first subword and last subword emission latency met-

rics, and apply end-of-turn emission latency penalty to regularize end-pointing emission latency.

2. Separator-Transducer-Segmenter

2.1. Separator-Transducer

STS model inherits its separator and transducer functionalities from a multi-turn RNN-T (MT-RNN-T) [26]. MT-RNN-T extends the standard RNN-T [25] to overlapping speech recognition with multiple output channels N, where N is a maximum number of simultaneously active speakers in the audio. Such design enforces a switch between output channels at speech overlap only and scales to an arbitrary number of speakers. In this work, we only consider cases with $N = 2$, and illustrate the corresponding STS model architecture on Fig. 1.

The encoder of STS has a modular structure containing a mixture encoder (MixEnc), N separation encoders (SepEnc_n) for each output channel $n \in \{1, ..., N\}$ and a recognition encoder (RecEnc) with shared parameters between output channels. The encoder takes acoustic features $\mathbf{x}$ as input and produces high-level disentangled acoustic representations $\mathbf{h}_n$ as output:

$$\mathbf{h}_n = \mathrm{RecEnc}(\mathrm{SepEnc}_n(\mathrm{MixEnc}(\mathbf{x}))). \quad (1)$$

In order to associate $\mathbf{h}_n$ with prediction network outputs for each label sequence $\mathbf{y}_n$ we employ deterministic assignment training (DAT) method [23], which forces the model to learn to associate its output with the speaker order in the audio. In this case, first separation encoder learns to focus on the very first speaker turn, and the second on the follow-up speaker turn if it exists. As a result, DAT computes RNN-T loss only N times:

$$\mathcal{L}_{RNN-T} = -\sum_n \log P(\mathbf{y}_n|\mathbf{h}_n). \quad (2)$$

2.2. Segmenter

On top of the Separator-Transducer model responsible for multi-speaker speech recognition, here we propose a novel Segmenter functionality to perform segmentation of multi-speaker hypotheses into single-speaker hypotheses and estimate turn boundaries for each speaker turn. To achieve this goal, we introduce a new segmentation modelling strategy in Section 2.2.1, and explore various regularization methods to enhance its performance in Sections 2.2.2, 2.2.3 and 2.2.4.

2.2.1. Segmentation modeling

In [26] *change-of-turn* (<cot>) tag was introduced in-between per-turn target transcriptions for each output channel. It was reported that this approach underestimated the number of turns in the audio. In this work, we revisit this design choice and introduce two separate tags for turn segmentation: *start-of-turn* (<sot>) and *end-of-turn* (<eot>). This approach enables joint start-of-turn and end-of-turn detection and allows interpretation of emission timestamps of these tokens as turn boundaries. In the future, these timestamps can be used for other tasks like speaker diarization (as in [32]) or endpoint detection (as in [29]).

2.2.2. FastEmit

Since in STS model emission timings of <sot> and <eot> tokens act as turn boundaries for potential future application in downstream tasks, it becomes important to regularize their emission latency. To achieve this goal, we use FastEmit [31], a sequence-level emission regularization method. It encourages predicting non-blank tokens and suppresses blank tokens across the entire sequence based on transducer forward-backward probabilities.

2.2.3. Multi-task training with masking loss

To further regularize segmentation capability of the STS model, during training we expose it to the ground-truth segmentation information that is encoded in speech activity labels. Inspired by the work of [16], we employ L2 masking loss to penalize recognition encoder outputs $\mathbf{h}_n$ in regions with no active speaker. The model is trained in a multi-task fashion by jointly optimizing RNN-T and masking losses:

$$\mathcal{L} = \mathcal{L}_{RNN-T} + \gamma * \sum_n L2(\mathbf{h}_n \circ \mathbf{m}_n), \quad (3)$$

where $\mathbf{m}_n$ is an inverse binary mask of speech activity for output channel n and γ is a weight of masking loss. This approach enforces encoder outputs $\mathbf{h}_n$ for each output channel n to be close to 0 in frames where there is no active speaker turn assigned to this output channel.

2.2.4. End-of-turn latency penalty

Besides FastEmit, we also explore another emission regularization method, dynamic latency penalty, applied specifically to the <eot> token. This approach was originally proposed for *end-of-speech* token emission latency regularization in [33] and investigated for endpoint detection in multi-speaker ASR in [29]. In our case, for each <eot> token in the target transcription, we apply the following dynamic penalty to the probability of <eot> emission in log domain:

$$\log P(\langle eot \rangle|\mathbf{x}_t) - = max(0, \alpha(t - \tau - t_{end})) \quad (4)$$

where α is a tunable scale of a late <eot> emission penalty, τ is a <eot> token frame buffer and t_{end} is a ground-truth end-of-turn frame. This penalty increases over time and enforces timely emission of the <eot> token.

3. Experimental setup

3.1. Task description

We perform experiments with STS on LibriCSS dataset proposed in [34] for continuous speech separation. It contains 10 one-hour long audio sessions with LibriSpeech utterances played back in a room to simulate meetings with 8 speakers, and it is divided into 6 partitions. 0S and 0L partitions exclude overlapped speech but contain short (S) and long (L) silence gaps between speaker turns, respectively. The remaining partitions represent different overlap ratios from 10% to 40%: OV10, OV20, OV30 and OV40. We use Session 0 of this dataset as a development set to tune decoding hyper-parameters and select best checkpoints, while the remaining Sessions 1-9 are used to report performance.

Following previous work in [26], we adopt an utterance group evaluation protocol (proposed in [20]) for experiments on this dataset. This evaluation protocol enforces segmentation of the original one-hour long audio sessions into utterance group segments using oracle silence boundary information. It ensures the existence of utterance groups containing only one speaker turn (0S, 0L) and utterance groups containing multiple partially overlapping turns. We are aware of the fact that parallel works on streaming multi-speaker ASR [27, 30] recently adopted an alternative continuous input evaluation protocol from [34], and we plan to address this discrepancy in the future work.

3.2. Training setup

The model topology of STS closely follows the one established in the previous work on MS-RNN-T and MT-RNN-T [23, 26].

Table 1: *WER and turn counting accuracy benchmarking of the STS model variants against the baseline on LibriCSS.*

Model	Turn counting accuracy [%]		WER [%]						
	Overall	>2 turns	0L	0S	OV10	OV20	OV30	OV40	full
MT-RNN-T [26]	85.6	28.0	14.7	14.8	20.7	25.3	33.2	36.4	25.3
STS	89.0	43.0	14.8	14.7	18.1	24.0	32.6	38.8	25.0
+FastEmit	90.1	47.5	13.0	13.5	16.0	21.3	29.3	31.3	21.7
+Masking loss	90.2	50.6	13.0	14.0	15.9	18.8	28.6	30.7	21.1

We use 2 LSTM layers in each recurrent module of the architecture (mixture encoder, 2 separation encoders, recognition encoder, prediction network) with 1024 units in each layer. Layer normalization [35] is performed after each LSTM layer in the model architecture. Output layers in the recognition encoder and the prediction network have 640 units. The joint network has a single feed-forward layer with 512 units. The output softmax layer has a dimensionality of 2503 which corresponds to the blank label, <sot> token, <eot> token and 2500 wordpieces that represent the most likely subword segmentation from a unigram word piece model [36].

Acoustic features are 64-dimensional log-mel filterbanks with a frame shift of 10ms which are stacked and downsampled by a factor of 3. We use SpecAugment with LibriFullAdapt policy [37] for feature augmentation. We use the Adam algorithm [38] with the warm-up, hold and decay schedule proposed in [39] for the optimization of all models. All experiments with enabled FastEmit are done with $\lambda_{FastEmit} = 0.005$.

STS model is pre-trained with a single separation encoder on the LibriSpeech dataset. We use on-the-fly data simulation pipeline developed in [26] for subsequent training on multi-speaker data. For each simulated example, we sample random number of utterances uniformly from the range $\{1, \ldots, 5\}$, scale them to achieve desired energy ratio (sampled from the range between -5 dB and 5 dB) and convolve with an acoustic impulse response (AIR) before adding to the mixture. Simulated examples longer than 30 seconds are filtered out to avoid out-of-memory errors in the RNN-T loss.

As an additional optimization on top of the segmentation strategy described in Section 2.2.1, we remove <sot> from the first turn and <eot> from the last turn in target transcriptions. We motivate this design choice by the absence of leading and trailing silence segments in our experimental setup, which makes modeling of <sot> and <eot> redundant and arguably detrimental in such cases.

4. Results

We report speech recognition and turn counting performance of the STS model variants on the LibriCSS dataset in Table 1. We measure speech recognition performance in optimal reference combination WER (ORC WER)[26], which effectively factors out the reference-hypothesis pairing errors from the actual word recognition errors. We measure turn counting performance in terms of accuracy of the correct number of turn prediction for 2 cases: overall accuracy and accuracy on utterances with > 2 turns. The latter is of particular interest for us, since cases with 1 or 2 turns are easily tackled by the model with 2 outputs, and do not require explicit segmentation. We select MT-RNN-T with the *change-of-turn* token from [26] as a baseline in this experiment. For experimental models we consider 3 STS variants: vanilla STS with turn boundary modeling through <sot> and <eot> tokens, STS with FastEmit, and STS trained with both FastEmit and masking loss.

As shown in Table 1, STS significantly improves turn counting accuracy by 3.4% abs. (85.6→ 89.0). This improvement is especially pronounced on utterances with $>$ 2 turns, where we report 15% abs. gain in performance (28.0 → 43.0). We observe some fluctuations of speech recognition performance among different data partitions, but WER on the full dataset remains on par. This observation clearly shows the benefit of the proposed segmentation modeling strategy for turn counting performance.

STS model with FastEmit further improves turn counting accuracy for utterances with more than 2 turns by 4.5% abs., and achieves overall relative word error rate reduction (WERR) of 13%. The latter is attributed to halved deletion rate (from 12% to 6%), and shifted ratio between insertion and deletion errors (from 0.18 to 0.48). Evidently, it also helps with more reliable turn count estimation, as the model is less prone to delete the whole turn in the worst-case scenario.

Multi-task training with L2-loss brings an additional boost to the turn counting accuracy on utterances with $>$ 2 turns, which is improved by 3% abs. Moreover, due to strong regularization effects, multi-task training leads to rel. WERR on LibriCSS partitions with high ratio of overlapped speech, i.e. 12% on OV20, 2% on OV30 and 2% on OV40. To better understand the behaviour of the masking loss, we compare per-frame L2 norms of recognition encoder outputs in regions with and without speech activity. We observe that masking loss changes the ratio between the average per-frame norm in "active" and "non-active" regions from 1.2 to 5.3. This observation shows that leakage of acoustic information into non-active regions can be detrimental to the model performance, but it can be partially mitigated by the masking loss.

5. Segmentation analysis

5.1. Motivation

Results in Section 4 demonstrate the benefit of the proposed segmentation modeling strategy for the turn counting accuracy. However, in the realistic scenario, we are not only interested in the correct prediction of the number of turns in the audio, but also in the turn boundary estimation, i.e. prediction of start-of-turn and end-of-turn timestamps. End-of-turn timestamps can be used for endpoint detection, i.e. to close a current turn and propagate its transcription to downstream services such as natural language understanding (NLU). Both start-of-turn and end-of-turn timestamps can assist speaker diarization in assigning a speaker label to each turn. Therefore, in this section, we propose a methodology for performing comprehensive segmentation analysis of the STS model.

5.2. Methodology

To better understand token emission behavior of <sot> and <eot> tokens, we extract emission timings for both output channels of the STS model. For each analysis we pick utterances with $>$ 2 turns as the remaining cases are trivial for a two-output system like ours. Moreover, we only focus on ut-

Table 2: *Segmentation analysis of the STS model variants on LibriCSS. pX is a X-th percentile of emission latency (EL).*

Model	Emission latency (EL) [ms]											
	End-pointing (EP)			Last subword (LS)			Start-pointing (SP)			First subword (FS)		
	Mean	p50	p90	Mean	p50	p90	Mean	p50	p90	Mean	p50	p90
STS	1428	1100	2611	267	60	230	712	386	666	907	603	794
+FastEmit	1509	1100	2793	359	10	192	479	211	572	787	555	728
+Masking loss	1288	980	2711	74	-1	145	332	263	537	570	561	730

terances with correctly estimated number of turns. We consider the following emission latency metrics for this analysis:

End-pointing emission latency (EP EL) – difference between ground-truth end-of-turn timestamp and emission timing of the <eot> token. Last turn is omitted.

Last subword emission latency (LS EL) – difference between ground-truth end-of-turn timestamp and emission timing of the last subword token in this turn. Last turn is omitted.

Start-pointing emission latency (SP EL) – difference between ground truth start-of-turn timestamp and emission timing of the <sot> token. First turn is omitted.

First subword emission latency (FS EL) – difference between ground truth start-of-turn timestamp and emission timing of the first subword token in this turn. First turn is omitted.

On Fig. 2 an example emission latency analysis is depicted. It contains STS model output timestamps for each word and special token as well as ground-truth start-of-turn (blue dashed lines) and end-of-turn (red dashed lines) timestamps for all turns taken into consideration. As seen from this example, LS EL sets a lower bound on EP EL, while FS EL sets an upper bound on SP EL. Difference between SP and FS EL also shows how much audio context STS model needs to open the next turn without predicting the first subword token.

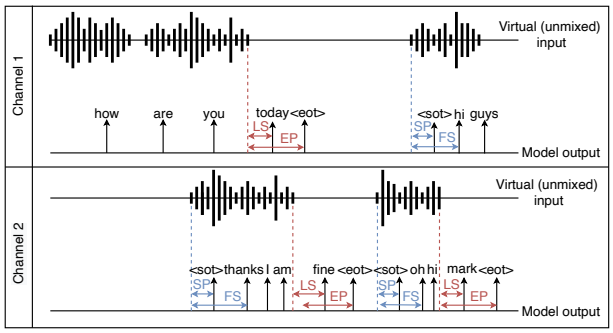

Figure 2: *Segmentation analysis example for the proposed STS model. For this analysis, 4 emission latency (EL) metrics are considered: end-pointing (EP), start-pointing (SP), first subword (FS) and last subword (LS).*

5.3. Results

In Table 2 we present the results of the segmentation analysis for the STS model variants. Vanilla STS shows the benefit of start-pointing modeling through <sot> token, as SP EL for half of utterances is 35% smaller than FS EL (386ms vs. 603ms). On a different note, a sizeable gap between LS EL and EP EL reveals a potential caveat of the proposed end-of-turn modeling approach through <eot> token. Average EP EL is around 1.5 sec, which shows that in most cases vanilla STS model significantly delays prediction of <eot>.

FastEmit brings substantial improvements to almost all considered latency metrics. SP EL p50 is almost halved (386ms $\rightarrow$ 211ms), while LS EL p50 is improved by 50ms (60ms $\rightarrow$ 10ms). However, FastEmit does not have an expected emission regularization impact on the <eot> token. This observation is in-line with what was reported in [29] in the context of *end-of-speech* modeling for multi-speaker ASR, and it motivates us to experiment with a dedicated end-of-turn latency penalty in Section 5.4. Interestingly, multi-task training with masking loss effectively stabilizes emission latency distribution, which manifests itself in almost 5-fold LS EL reduction (359ms $\rightarrow$ 74ms).

Table 3: *Impact of end-of-turn latency penalty frame buffer τ on end-pointing emission latency (EL) and WER. pX is a X-th percentile of EL.*

τ	End-pointing EL [ms]						WER
	Mean	p50	p60	p70	p80	p90	[%]
-	1428	1100	1440	1828	2136	2611	25.0
3	1031	14	50	73	134	7458	27.6
5	1322	80	100	140	204	7507	26.3
7	1748	130	158	190	288	9415	27.3
10	1058	196	230	260	300	5388	27.7

5.4. Experiments with end-of-turn latency penalty

To specifically improve EP EL, we apply the end-of-turn latency penalty approach described in Section 2.2.4. We use vanilla STS as a baseline in this experiment, and apply the end-of-turn latency penalty with a fixed scale $\alpha = 1$ and different values of end-of-turn frame buffer τ. As shown in Table 3, it successfully reduces EP EL p50 at least by the order of magnitude for 50-th, 60-th, 70-th and 80-th percentiles. Relaxed end-of-turn frame buffer τ results in a delayed end-point detection. However, we observe WER and EL EP p90 degradation with all explored τ values. A more detailed error analysis reveals that they originate from a few end-point detection failures, which lead to the hallucinated hypothesis duplicates from the parallel output channel in affected turns. We tried to address it by combining end-of-turn latency penalty with the best STS model that incorporates both FastEmit and masking loss, with limited success. The major culprit is a numerical instability of the training with both FastEmit and end-of-turn latency penalties, and we plan to address it in the future work.

6. Conclusion

In this paper, we proposed Separator-Transducer-Segmenter (STS) model for joint recognition and segmentation of multi-party speech through prediction of turn boundary tokens <sot> and <eot>. It improved turn counting accuracy by 15% abs. on partially overlapping LibriCSS utterances with $>$ 2 turns and enabled segmentation analysis based on emission latency of these tokens. On top of STS, three additional modeling changes were explored: an emission regularization method FastEmit, a multi-task training approach with speech activity signal and end-of-turn emission latency penalty. The former two combined additionally improved turn counting accuracy by 7.6% abs. on utterances with $>$ 2 turns and overall WER by 16% rel., while the latter significantly improved end-pointing emission latency for most turns at the cost of slight WER degradation.

7. References

[1] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa, and M. Lincoln, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP 2007*, vol. 4, 2007, pp. IV–357–IV–360.

[2] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The Sheffield wargames corpus," in *Proc. Interspeech 2013*, 2013, pp. 1116–1120.

[3] Y. Liu, C. Fox, M. Hasan, and T. Hain, "The sheffield wargame corpus - day two and day three," *Interspeech*, Sep 2016.

[4] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *Interspeech*, Sep 2018.

[5] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech Language*, vol. 72, p. 101317, 2022.

[6] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *Interspeech*, Sep 2016.

[7] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," *Interspeech*, Sep 2019.

[8] S. Settle, J. L. Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *ICASSP*, 2018, pp. 4819–4823.

[9] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," *ICASSP*, May 2020.

[10] T. v. Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR," *Interspeech 2020*, Oct 2020.

[11] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *ICASSP*, Mar 2017.

[12] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *Interspeech*, Aug 2017.

[13] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, p. 1–11, Nov 2018.

[14] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[15] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," *ICASSP*, May 2019.

[16] A. Tripathi, H. Lu, and H. Sak, "End-to-end multi-talker overlapping speech recognition," in *ICASSP*, 2020, pp. 6129–6133.

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

[18] 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.

[19] N. Kanda, G. Ye, Y. Wu, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Large-scale pre-training of end-to-end multi-talker ASR for meeting transcription with single distant microphone," in *Proc. Interspeech 2021*, 2021, pp. 3430–3434.

[20] N. Kanda, X. Chang, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings," in *Proc. SLT 2021*, 2021, pp. 809–816.

[21] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "End-to-end speaker-attributed ASR with transformer," *Interspeech*, Sep 2021.

[22] N. Kanda, X. Xiao, J. Wu, T. Zhou, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "A comparative study of modular and joint approaches for speaker-attributed asr on monaural long-form audio," *ASRU*, 2021.

[23] I. Sklyar, A. Piunova, and Y. Liu, "Streaming multi-speaker ASR with RNN-T," in *ICASSP*, 2021, pp. 6903–6907.

[24] L. Lu, N. Kanda, J. Li, and Y. Gong, "Streaming end-to-end multi-talker speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 803–807, 2021.

[25] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, vol. abs/1211.3711, 2012.

[26] I. Sklyar, A. Piunova, X. Zheng, and Y. Liu, "Multi-turn rnn-t for streaming recognition of multi-party speech," in *ICASSP*, 2022.

[27] D. Raj, L. Lu, Z. Chen, Y. Gaur, and J. Li, "Continuous streaming multi-talker asr with dual-path transducers," in *ICASSP*, 2022.

[28] L. Lu, N. Kanda, J. Li, and Y.Gong, "Streaming multi-talker speech recognition with joint speaker identification," in *Proc. Interspeech 2021*, 2021, pp. 1782–1786.

[29] L. Lu, J. Li, and Y. Gong, "Endpoint detection for streaming end-to-end multi-talker asr," in *ICASSP*, 2022.

[30] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming multi-talker ASR with token-level serialized output training," *CoRR*, vol. abs/2202.00842, 2022. [Online]. Available: https://arxiv.org/abs/2202.00842

[31] J. Yu, C.-C. Chiu, B. Li, S.-y. Chang, T. N. Sainath, Y. He, A. Narayanan, W. Han, A. Gulati, Y. Wu, and R. Pang, "Fastemit: Low-latency streaming asr with sequence-level emission regularization," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6004–6008.

[32] W. Xia, H. Lu, Q. Wang, A. Tripathi, I. Lopez-Moreno, and H. Sak, "Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection," *CoRR*, vol. abs/2109.11641, 2021. [Online]. Available: https://arxiv.org/abs/2109.11641

[33] B. Li, S.-y. Chang, T. N. Sainath, R. Pang, Y. He, T. Strohman, and Y. Wu, "Towards fast and accurate streaming end-to-end asr," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6069–6073.

[34] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP*, 2020, pp. 7284–7288.

[35] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *arXiv*, 2016.

[36] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Nov. 2018, pp. 66–71.

[37] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "SpecAugment on large scale datasets," *ICASSP*, May 2020.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*, 2015.

[39] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, Sep 2019.