INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Joint Autoregressive Modeling of End-to-End Multi-Talker Overlapped Speech Recognition and Utterance-level Timestamp Prediction

Naoki Makishima, Keita Suzuki, Satoshi Suzuki, Atsushi Ando, Ryo Masumura

NTT Computer and Data Science Laboratories, NTT Corporation, Japan

naoki.makishima@ntt.com

Abstract

This paper proposes autoregressive modeling of the joint multi-talker automatic speech recognition (ASR) and timestamp prediction. Autoregressive modeling of multi-talker ASR is a simple and promising approach. However, it does not predict utterance timestamp information despite its being important in practice. To address this problem, our key idea is to extend autoregressive-modeling-based multi-talker ASR to predict quantized timestamp tokens representing the start and end time of an utterance. Our method estimates transcription and utterance-level timestamp tokens of multiple speakers one after another. This enables joint modeling of multi-talker ASR and timestamps prediction without changing the simple autoregressive modeling of the conventional multi-talker ASR. Experimental results show that our method outperforms the ASR performance of conventional autoregressive multi-talker ASR without timestamp prediction and achieves promising timestamp prediction accuracy.

Index Terms: multi-talker automatic speech recognition, timestamp prediction, autoregressive modeling

1. Introduction

Our natural conversations and meetings often include speech where several people speak simultaneously. The task of multi-talker automatic speech recognition (ASR) is to transcribe each utterance of this overlapped speech into text. Several studies have tackled the issue of how to accurately transcribe overlapped speech [1–13]. One of the most popular approaches uses serial pipeline processing that combines speech separation and ASR [1–7]. This approach first separates the overlapped speech into non-overlapped speech using speech separation methods such as deep clustering [1, 14] or Transformer-based methods with permutation invariant training (PIT) [12, 15] and then applies typical single-talker ASR [16–20] that is trained to transcribe non-overlapped speech. Although powerful speech separation methods are available, one of their weaknesses is that they cannot handle the dependency among utterances of each speaker, because the output separated speech is independently processed by single-talker ASR, which leads to poor ASR results, including duplicate hypotheses.

In contrast, a recent promising approach for end-to-end multi-talker ASR is autoregressive modeling with serialized output training (SOT) [11, 21, 22]. Instead of having independent multiple output layers that output transcriptions of each speaker [1–7], it generates the transcription of multiple speakers recursively one after another with a single output layer. This enables simple and natural modeling of the dependency among the outputs for multiple speakers; that is, the transcription of the next speaker is predicted given the transcription of the previous speaker's speech, just like in our natural conversations, with the same architecture as simple single-talker ASR. In addition, this not only avoids the maximum speaker number constraints but also helps avoid the generation of duplicate hypotheses. Specifically, it outputs serialized transcriptions of multiple speakers by introducing a special symbol representing the speaker change and concatenating the transcriptions using the special symbol.

However, a limitation is that their prediction does not include utterance timestamp information. In other words, although we understand who speaks what, we cannot figure out when the utterance is spoken, despite its being important in practice. For example, the overlap ratio indicates how active a meeting is and it becomes easier to follow the flow of the discussions with timestamps when reading the record. To address this problem, our key idea is to extend autoregressive-modeling-based multi-talker ASR to predict quantized timestamp tokens representing the start and end time of an utterance. The idea to predict quantized timestamp tokens is first explored in single-talker ASR [20]. In [20], they directly estimate the quantized timestamp tokens representing the start time of the utterance, the transcription, and the quantized timestamp token representing the end time of the utterance one after another. This approach is promising because it enables the timestamp prediction with single-talker ASR without changing the simple autoregressive modeling of the conventional single-talker ASR that does not predict timestamp information. On the other hand, to the best of our acknowledge, this approach is not verified in multi-talker overlapped ASR settings.

In this paper, we propose joint autoregressive modeling of multi-talker ASR and timestamps prediction. The proposed method outputs serialized transcriptions of multiple speakers and their utterance-level timestamps one after another by introducing speaker change symbols and quantized timestamp tokens. Since the task to estimate quantized timestamp tokens is solved as a classification problem, our method solves both multi-talker ASR and timestamps prediction with a single output layer; that is, the same output layer is used to estimate transcription and quantized timestamps, which enables the same simple autoregressive modeling as the conventional multi-talker ASR without timestamp prediction. The proposed method estimates both start and end timestamp tokens of each speaker before ASR results to learn their dependencies. We conducted experiments to evaluate the ASR performance and timestamp prediction performance with the Corpus of Spontaneous Japanese (CSJ) [23]. We experimentally show that the joint modeling of transcription and timestamps improve ASR performance and that training to transcribe overlapped speeches with more than one speaker improves ASR performance and timestamp prediction accuracy even under a typical single-talker setting. Moreover, we compare our method with neural speaker diariza-

10.21437/Interspeech.2023-564

tion [24], which estimates the frame-level speech activation of each speaker to evaluate the prediction accuracy of the timestamps, and show that the proposed method achieves better results.

2. Conventional methods

2.1. Single-talker ASR with autoregressive modeling

We denote the acoustic feature of the input speech and its textual token as $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$ and $\boldsymbol{W} = (w_1, \dots, w_N)$, respectively, where $\boldsymbol{x}_t \in \mathbb{R}^F$ denotes the tth frame of the feature, F denotes its dimension, T denotes the length of acoustic features, $w_n \in \mathcal{V}$ denotes the nth textual token, $\mathcal{V}$ denotes the vocabrary set, and N denotes the length of the token. Typical single-talker ASR with autoregressive modeling estimates generation probability of $\boldsymbol{W}$ given $\boldsymbol{X}$ as follows:

$$P(\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta}_{\mathrm{ST}}) = \prod_{n=1}^{N} P(w_n|\boldsymbol{w}_{1:n-1}, \boldsymbol{X};\boldsymbol{\Theta}_{\mathrm{ST}}), \quad (1)$$

where $\boldsymbol{\Theta}_{\mathrm{ST}}$ denotes the parameters of the single-talker ASR model and $\boldsymbol{w}_{1:n-1} = (w_1, \dots, w_{n-1})$. The parameter $\boldsymbol{\Theta}_{\mathrm{ST}}$ is optimized with the following cross-entropy function:

$$L_{\mathrm{ST}} = -\log P(\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta}_{\mathrm{ST}}). \quad (2)$$

When estimating timestamps in single-talker ASR [20], all timestamps are quantized and additional timestamp tokens are added to the ASR result; the start time token is predicted before transcription, and the end time token is predicted after it. In [20], all timestamps are quantized to the nearest 20 ms, and the output hypothesis becomes $(< t_{\mathrm{s}} >, w_1, \dots, w_N, < t_{\mathrm{e}} >)$, where $< t_{\mathrm{s}} >$ and $< t_{\mathrm{e}} >$ denote the quantized start time token and end time token, respectively.

2.2. Multi-talker ASR with autoregressive modeling

We denote utterance-level textual tokens of multiple speakers as $\boldsymbol{W}^{1:K} = (\boldsymbol{W}^1, \dots, \boldsymbol{W}^K)$, where K denotes the number of speakers in the overlapped speech, $\boldsymbol{W}^k = (w_1^k, \dots, w_{N^k}^k)$ denotes the kth speaker's textual token, and N^k denotes the length of the token. Since there are permutation ambiguities in the order of k when predicting $\boldsymbol{W}^k$, the first-in, first-out approach [10, 11] is adopted in SOT. In this approach, $\boldsymbol{W}^{1:K}$ is sorted by their utterance start times. Moreover, to recognize multiple utterances with a single output layer, $\boldsymbol{W}^{1:K}$ is serialized into a single token sequence with special symbol [sep] representing speaker change. The serialized token $\boldsymbol{S} \in \{\mathcal{V} \cup \mathcal{O}\}$ is given as

$$\begin{aligned}\boldsymbol{S} = (&w_1^1, \dots, w_{N^1}^1, [\mathrm{sep}], w_1^2, \dots, w_{N^2}^2, [\mathrm{sep}], \\ &\dots, w_{N^{K-1}}^{K-1}, [\mathrm{sep}], w_1^K, \dots, w_{N^K}^K, [\mathrm{eos}]),\end{aligned} \quad (3)$$

where [eos] denotes the end of a sentence, $\mathcal{O} = \{[\mathrm{sep}], [\mathrm{eos}]\}$, and we assume that $\boldsymbol{W}^{1:K}$ is sorted in order of utterance start times for simplicity.

Multi-talker ASR with autoregressive modeling estimates generation probability of $\boldsymbol{S}$ given acoustic feature $\boldsymbol{X}$ in the same manner as single-talker ASR as

$$P(\boldsymbol{S}|\boldsymbol{X};\boldsymbol{\Theta}_{\mathrm{MT}}) = \prod_{l=1}^{|S|} P(s_l|\boldsymbol{s}_{1:l-1}, \boldsymbol{X};\boldsymbol{\Theta}_{\mathrm{MT}}), \quad (4)$$

where s_l denotes the lth token of $\boldsymbol{S}$, $\boldsymbol{s}_{1:l-1} = (s_1, \dots, s_{l-1})$, $|S|$ denotes the length of $\boldsymbol{S}$, and $\boldsymbol{\Theta}_{\mathrm{MT}}$ denotes the parameter of the multi-talker ASR model. The parameter $\boldsymbol{\Theta}_{\mathrm{MT}}$ is optimized with the following cross-entropy function:

$$L_{\mathrm{MT}} = -\log P(\boldsymbol{S}|\boldsymbol{X};\boldsymbol{\Theta}_{\mathrm{MT}}). \quad (5)$$

3. Proposed method

3.1. Strategy

As discussed in Section 1, since our natural conversations or meetings usually contain overlapped speech, joint modeling of multi-talker ASR and timestamp prediction is useful in many situations, such as for creating easy-to-read records of conversations and measuring the activity of meetings. Moreover, the joint modeling enables ASR to consider which segment includes the target utterance speech by predicting the start and end timestamps, which is expected to improve ASR performance, especially under overlapped settings.

Figure 1 shows an overview of the proposed method. In our autoregressive modeling of joint multi-talker ASR and timestamps prediction, we predict the joint generation probability of multiple transcriptions $\boldsymbol{W}^{1:K}$ and timestamps from single-channel overlapped speech. We use a unified autoregressive model with a single output layer to jointly handle the ASR task and timestamp prediction task.

3.2. Formulation

We denote the start time token and the end time token of multiple speakers as $\boldsymbol{T}_{\mathrm{s}}^{1:K} = (< t_{\mathrm{s}}^1 >, \dots, < t_{\mathrm{s}}^K >)$ and $\boldsymbol{T}_{\mathrm{e}}^{1:K} = (< t_{\mathrm{e}}^1 >, \dots, < t_{\mathrm{e}}^K >)$, respectively, where $< t_{\mathrm{s}}^k > \in \mathcal{T}$ and $< t_{\mathrm{e}}^k > \in \mathcal{T}$ denote the kth speaker's start time token and end time token, respectively, and $\mathcal{T}$ denotes the quantized time token label set. The quantized time token is obtained by rounding the continuous timestamp values to the nearest quantized value every Q seconds.

To efficiently model the joint generation probability of $\boldsymbol{W}^{1:K}$, $\boldsymbol{T}_{\mathrm{s}}^{1:K}$, and $\boldsymbol{T}_{\mathrm{e}}^{1:K}$, we serialize them into a single label sequence as SOT [11]. In our preliminary experiments, placing the start time token and end time token before transcription achieved slightly better ASR performance compared to the one described in Section 2.1. Thus, we serialize the labels and obtain the single label sequence $\tilde{\boldsymbol{S}} \in \{\mathcal{V} \cup \mathcal{O} \cup \mathcal{T}\}$ as

$$\begin{aligned}\tilde{\boldsymbol{S}} = (&< t_{\mathrm{s}}^1 >, < t_{\mathrm{e}}^1 >, w_1^1, \dots, w_{N^1}^1, [\mathrm{sep}], \\ &< t_{\mathrm{s}}^2 >, < t_{\mathrm{e}}^2 >, w_1^2, \dots, w_{N^2}^2, [\mathrm{sep}], \\ &\dots, < t_{\mathrm{s}}^K >, < t_{\mathrm{e}}^K >, w_1^K, \dots, w_{N^K}^K, [\mathrm{eos}]).\end{aligned} \quad (6)$$

The joint generation probability of $\boldsymbol{W}^{1:K}$, $\boldsymbol{T}_{\mathrm{s}}^{1:K}$, and $\boldsymbol{T}_{\mathrm{e}}^{1:K}$ given multi-talker overlapped speech $\boldsymbol{X}$ is obtained as

$$\begin{aligned}P(\boldsymbol{W}^{1:K}, \boldsymbol{T}_{\mathrm{s}}^{1:K}, \boldsymbol{T}_{\mathrm{e}}^{1:K}|\boldsymbol{X};\boldsymbol{\Theta}) &= P(\tilde{\boldsymbol{S}}|\boldsymbol{X};\boldsymbol{\Theta}) \\ &= \prod_{l=1}^{|\tilde{\boldsymbol{S}}|} P(\tilde{s}_l|\tilde{\boldsymbol{s}}_{1:l-1}, \boldsymbol{X};\boldsymbol{\Theta}),\end{aligned} \quad (7)$$

where $\tilde{s}_l$ denotes the lth token of $\tilde{\boldsymbol{S}}$, $\tilde{\boldsymbol{s}}_{1:l-1} = (\tilde{s}_1, \dots, \tilde{s}_{l-1})$, $|\tilde{\boldsymbol{S}}|$ denotes the length of $\tilde{\boldsymbol{S}}$, and $\boldsymbol{\Theta}$ denotes the parameter of the model in the proposed method.

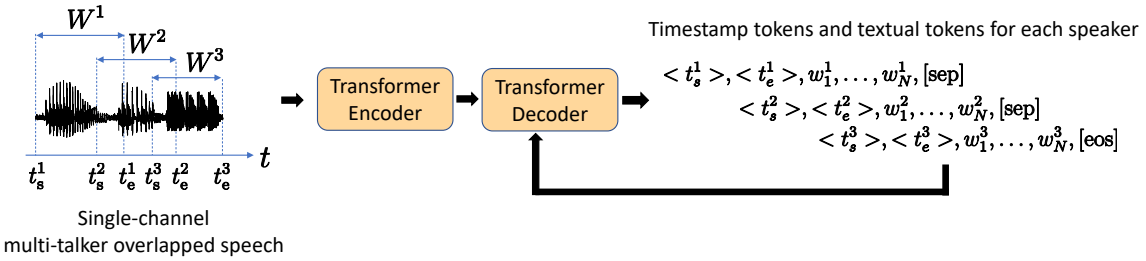

Figure 1: *Overview of the proposed method.*

3.3. Modeling

We use a Transformer-based ASR model [19, 25]. The joint generation probability is obtained as follows:

$$\boldsymbol{H} = \mathrm{TransformerEnc}(\boldsymbol{X}; \boldsymbol{\theta}_{\mathrm{enc}}), \tag{8}$$

$$P(\tilde{s}_l | \tilde{\boldsymbol{s}}_{1:l-1}, \boldsymbol{X}; \boldsymbol{\Theta}) = \mathrm{TransformerDec}(\boldsymbol{H}, \tilde{\boldsymbol{s}}_{1:l-1}; \boldsymbol{\theta}_{\mathrm{dec}}), \tag{9}$$

where $\mathrm{TransformerEnc}(\cdot)$ is a Transformer encoder that consists of a pre-net, a positional encoding layer, and multiple multi-head self-attention blocks; $\boldsymbol{\theta}_{\mathrm{enc}}$ denotes its parameters, $\mathrm{TransformerDec}(\cdot)$ is a Transformer decoder that consists of an embedding layer, a positional encoding layer, and multiple-multi-head self-attention and encoder-decoder attention blocks; and $\boldsymbol{\theta}_{\mathrm{dec}}$ denotes its parameters. We describe the detailed architecture of the model in Section 4.2. The parameter $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{\mathrm{enc}}, \boldsymbol{\theta}_{\mathrm{dec}}\}$ is optimized with the cross-entropy function that is defined as

$$L = -\log P(\tilde{\boldsymbol{S}} | \boldsymbol{X}; \boldsymbol{\Theta}). \tag{10}$$

4. Experiment

4.1. Dataset

We evaluated the proposed method by conducting single-talker and multi-talker ASR tasks with timestamp prediction. We used the CSJ [23] for our experiments. First, we divided CSJ into training, validation, and test data. Training data consists of 1,388 speakers, and its size is 522 h. Validation data consists of 10 speakers, and its size is 1.3 h, and test data consists of 10 speakers, and its size is 1.9 h. Since the CSJ is a dataset for single-talker ASR, we created two-speaker and three-speaker simulated mixtures by mixing the utterances of different speakers for multi-talker ASR experiments. When mixing the audio signals, the original volume of each utterance was kept unchanged, resulting in an average signal-to-interference ratio of about 0 dB. As for the delay applied to each utterance, the delay values were randomly chosen under the constraints as in [11]. First, the start times of individual utterances differed by 0.5 s or longer. Second, every utterance in each mixed audio sample had at least one speaker-overlapped region with other utterances. The average overlap rate was about 35 %. Figure 2 shows the histograms of speech duration. We used 80 log mel-scale filterbank coefficients as acoustic features, which were extracted using a 20-ms-long Hann window with a 10-ms-long shift. For ASR, this paper used characters as textual tokens. We set Q as 0.5, which means continuous timestamps are rounded every 0.5 s.

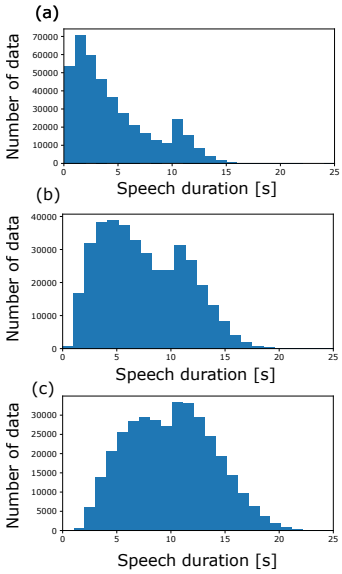

Figure 2: *Histogram of speech duration of (a) single-speaker dataset, (b) simulated two-speaker dataset, and (c) simulated three-speaker dataset.*

4.2. Implementation

We used a Transformer-based ASR model [19, 25] in this paper. The acoustic feature was first passed to layers composed of two 1×1 convolutions with 1×1 strides, two max pooling with a stride of 2, two 3×3 depthwise convolutions with 1×1 strides, and two long-short term memory layers with outputs of 256 dimensions. Then, we stacked ten-layer Transformer encoder blocks, where the number of heads in the multi-head attention was set to 4, the dimensions of the output continuous representations were set to 256, and the dimensions of the inner output in the position-wise feed-forward networks were set to 1,024. For decoder layers, we stacked two-layer Transformer decoder blocks, where the settings were the same as for the encoder blocks. For the activation function, we used Swish [26].

4.3. Settings

We compared the following methods (listed in Table 1): conventional single-talker ASR with timestamp prediction [20], conventional multi-talker ASR without timestamp prediction [11], and the proposed multi-talker ASR with timestamp prediction. Moreover, to compare the direct timestamp prediction adopted in the proposed method with frame-level time estimation adopted in end-to-end neural diarization (EEND) of the

Table 1: *Evaluation results*

Number of speakers in test dataset	Methods	ASR	Timestamps	CER (%)	DER (%)	SCA (%)
1	Conventional single-talker ASR [20]	✓	✓	8.31	0.23	100
	Conventional multi-talker ASR [11]	✓		7.67	-	100
	EEND-EDA [24]		✓	-	0.16	99.2
	Proposed method w/o transcription		✓	-	0.01	99.9
	Proposed method	✓	✓	7.33	0.07	99.9
2	Conventional multi-talker ASR [11]	✓		8.92	-	97.8
	EEND-EDA [24]		✓	-	0.68	81.5
	Proposed method w/o transcription		✓	-	0.22	97.9
	Proposed method	✓	✓	8.56	0.19	98.8
3	Conventional multi-talker ASR [11]	✓		12.95	-	92.8
	EEND-EDA [24]		✓	-	4.69	72.1
	Proposed method w/o transcription		✓	-	0.36	93.8
	Proposed method	✓	✓	12.28	0.60	92.0

overlapped speech [24, 27–31], we also compared timestamp prediction with autoregressive modeling (the proposed method w/o transcription) and EEND with encoder-decoder based attractors (EEND-EDA) [24]. In the proposed method without transcription, the model was trained to estimate the quantized timestamp tokens $\boldsymbol{T}_{\mathrm{s}}^{1:K}$ and $\boldsymbol{T}_{\mathrm{e}}^{1:K}$ and the speaker change token $[\mathrm{sep}]$ recursively. EEND-EDA was trained following [24]. The other models were optimized by using the RAdam [32] algorithm with a minibatch size of 32. We set the learning rate of the algorithm to 0.0001. The training steps were stopped if the loss on the validation set did not decrease for ten epochs in succession. We applied label smoothing with the smoothing weight of 0.1 [33]. For testing, we used a beam search algorithm whose beam size was set to 20.

We used the character error rate (CER), diarization error rate (DER), and speaker count accuracy (SCA) to evaluate the total performance of the methods. Note that we used DER to evaluate prediction accuracy of the timestamps. Since our interest in this paper is not the estimation of the speaker identity, we created simulated mixtures so that the new speaker speaks after the speaker change in our dataset. In other words, the dataset consisted of one utterance per speaker, which enabled us to calculate the prediction accuracy of the timestamps with DER. When comparing hypothesized boundaries to references, we used a tolerance of 500 ms. SCA was calculated as the ratio of the number of test samples for which each method correctly counted the speaker to the total number of test samples. In multi-talker overlapped ASR settings, we compared hypotheses with references while considering the order of utterances. When calculating CER, we only evaluated textual tokens excluding the special token $\mathcal{O}$ and the time-stamp token $\mathcal{T}$.

4.4. Results

Table 1 shows the evaluation results for each method. Note that since we trained the models using single-speaker, two-speaker, and three-speakers datasets except for conventional single-talker ASR, we used the same parameters when evaluating the test dataset of different numbers of speakers. First, the proposed method achieves the best CER performance of all under all settings. Interestingly, when there is only one speaker in the test dataset (typical single-talker setting), the proposed method improves CER by 0.98% compared to conventional single-talker ASR. This suggests that training with more than one speaker is effective even when the model is used for single-talker ASR. Moreover, compared to conventional multi-talker ASR without timestamp prediction, the proposed method improves CER by 0.34%, 0.36%, and 0.67% when the number of speakers is one, two, and three. This is probably because the predicted timestamps help make the ambiguous overlapped speech boundaries clear, which leads to more accurate ASR when the number of speakers in overlapped speech is large.

Second, with respect to DER and SCA, the proposed method without transcription achieves better results compared to frame-level estimation of EEND-EDA, which suggests that autoregressive modeling of direct timestamp prediction is a promising approach to estimate start and end time of an utterance especially when the number of speakers in overlapped speech is large. In the typical single-talker setting, the proposed method improves DER by 0.16 % compared to conventional single-talker ASR. This suggests that training with the multi-talker dataset improves timestamp prediction performance as well as ASR performance.

5. Conclusions

In this paper, we proposed autoregressive modeling of joint multi-talker ASR and timestamps prediction under the multi-talker ASR setting. The proposed method outputs serialized transcriptions of multiple speakers and their quantized timestamp tokens one after another with a single layer, which enables simple and natural modeling of multi-talker ASR and timestamp prediction without changing the simple autoregressive modeling of the conventional multi-talker ASR. Experimental results show that the proposed method outperforms the conventional single-talker ASR and the conventional multi-talker ASR. This indicates that predicted timestamps are helpful when estimating transcription and that training with overlapped speech from more than one speaker improves typical single-talker ASR performance and timestamp prediction performance. Moreover, DER and SCA of the proposed method outperformed EEND-EDA, which indicates that the autoregressive modeling of direct timestamp prediction is a promising approach when estimating the start and end time of an utterance.

6. References

[1] Y. Z. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549.

[2] H. Seki, T. Hori, S. Watanabe, J. L. Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. ACL*, 2018, pp. 2620–2630.

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

[4] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," in *Proc. ICASSP*, 2019, pp. 6256–6260.

[5] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. SLT*, 2021, pp. 897–904.

[6] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. ICASSP*, 2021, pp. 5749–5753.

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

[8] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *Proc. ICASSP*, 2020, pp. 6134–6138.

[9] P. Denisov and N. T. Vu, "End-to-end multi-speaker speech recognition using speaker embeddings and transfer learning," in *Proc. Interspeech*, 2019, pp. 4425–4429.

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

[11] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *Proc. Interspeech*, 2020, pp. 2797–2801.

[12] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *Proc. Interspeech*, 2017, pp. 2456–2460.

[13] N. Kanda, X. Xiao, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR," in *Proc. ICASSP*, 2022, pp. 8082–8086.

[14] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[15] M. Kolbaek, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[16] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, vol. 32, 2014, pp. 1764–1772.

[17] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Advances in NIPS*, 2015, pp. 577–585.

[18] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016, pp. 4960–4964.

[19] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. ICASSP*, 2018, pp. 5884–5888.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[21] 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," in *Proc. Interspeech*, 2020, pp. 36–40.

[22] R. Masumura, D. Okamura, N. Makishima, M. Ihori, A. Takashima, T. Tanaka, and S. Orihashi, "Unified autoregressive modeling for joint end-to-end multi-talker overlapped speech recognition and speaker attribute estimation," in *Proc. Interspeech*, 2021, pp. 2591–2595.

[23] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC*, 2000.

[24] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1493–1507, 2022.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[26] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," in *Proc. ICLR Workshop*, 2018.

[27] L. E. Shafey, H. Soltau, and I. Shafran, "Joint speech recognition and speaker diarization via sequence transduction," in *Proc. Interspeech*, 2019, pp. 396–400.

[28] A. Zhang, Q. Wang, Z. Zhu, J. W. Paisley, and C. Wang, "Fully supervised speaker diarization," in *Proc. ICASSP*, 2019, pp. 6301–6305.

[29] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech*, 2019, pp. 4300–4304.

[30] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech*, 2020, pp. 274–278.

[31] H. H. Mao, S. Li, J. J. McAuley, and G. W. Cottrell, "Speech recognition and multi-speaker diarization of long conversations," in *Proc. Interspeech*, 2020, pp. 691–695.

[32] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. ICLR*, 2020.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 2818–2826.