INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone

Naoyuki Kanda[1], *Guoli Ye*[1], *Yu Wu*[2], *Yashesh Gaur*[1], *Xiaofei Wang*[1], *Zhong Meng*[1], *Zhuo Chen*[1], *Takuya Yoshioka*[1]
[1]Microsoft Cloud+AI, USA
[2]Microsoft Research Asia, China
{Naoyuki.Kanda,guoye,Wu.Yu,Yashesh.Gaur,Xiaofei.Wang,Zhong.Meng}@microsoft.com,
{zhuc,tayoshio}@microsoft.com

Abstract

Transcribing meetings containing overlapped speech with only a single distant microphone (SDM) has been one of the most challenging problems for automatic speech recognition (ASR). While various approaches have been proposed, all previous studies on the monaural overlapped speech recognition problem were based on either simulation data or small-scale real data. In this paper, we extensively investigate a two-step approach where we first pre-train a serialized output training (SOT)-based multi-talker ASR by using large-scale simulation data and then fine-tune the model with a small amount of real meeting data. Experiments are conducted by utilizing 75 thousand (K) hours of our internal single-talker recording to simulate a total of 900K hours of multi-talker audio segments for supervised pre-training. With fine-tuning on the 70 hours of the AMI-SDM training data, our SOT ASR model achieves a word error rate (WER) of 21.2% for the AMI-SDM evaluation set while automatically counting speakers in each test segment. This result is not only significantly better than the previous state-of-the-art WER of 36.4% with oracle utterance boundary information but also better than a result by a similarly fine-tuned single-talker ASR model applied to beamformed audio.

Index Terms: multi-talker speech recognition, speaker counting, serialized output training

1. Introduction

Meeting transcription with a distant microphone has been widely studied as one of the most challenging problems for automatic speech recognition (ASR) [1, 2, 3]. The audio is noisy and reverberant due to the distance between the speaker and microphone and often includes overlapped utterances [4]. Meanwhile, the sentences are usually less grammatical in verbal communication, which creates additional difficulties for ASR. While various approaches have been proposed especially for microphone array settings (e.g. [5, 6]), meeting transcription with only a single distant microphone (SDM) is still highly challenging. For example, the best reported word error rate (WER) for the AMI meeting corpus [2] is still over 35% for the SDM setting even with oracle utterance boundary information [7, 8, 9, 10].

The meeting transcription system needs to recognize utterances from a variable number of speakers from audio which may contain overlapped utterances. To handle this, one approach is to first apply speaker diarization [11, 12, 13] to detect the utterance boundaries for each speaker and then perform ASR for each utterance. This method, however, could suffer from the accuracy degradation in overlapped regions because the ASR system is usually designed to recognize single-speaker speech. Another approach is to apply a speech separation system, followed by the ASR system (e.g., [14, 15]). However, a speech separation system is usually trained with a signal-level criterion, which is not necessarily optimal for ASR.

To overcome this suboptimality, there has been a series of studies for multi-talker ASR that directly transcribes multiple utterances from overlapped speech. One popular approach is using a neural network that has multiple output layers, each of which recognizes one speaker [16, 17, 18, 19, 20, 21, 22]. Permutation invariant training (PIT) [23] is usually used to train such a multiple-output models. One drawback of such method is, however, that the number of recognizable speakers is limited by the number of the output layers. PIT also requires $O(S^3)$ of computation with respect to the number of recognizable speakers S, which makes it inefficient to conduct large-scale training. Recently, serialized output training (SOT) was proposed to recognize any number of utterances in speaker-overlapped audio [24]. The SOT-based ASR was shown to outperform PIT-based ASR while automatically counting the speakers in the speaker-mixed audio. [24] also showed that the SOT-based ASR could be trained with $O(S)$ efficiency.

While promising results were shown, all the previous studies were limited to either simulated data [16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27] or small-scale real data [7, 9, 8, 10][1]. It is due to the difficulty in collecting real meeting recordings with precise transcriptions at large-scale. One potential approach to the data scarcity problem is large-scale pre-training, which has been studied for *single-talker* ASR (e.g., with labeled data [29, 30] or with unlabeled data [31, 32]). However, it is still an open question if we can learn a good representation for *multi-talker audio which is sometimes heavily overlapped.*

To further advance the SDM-based meeting transcription, we extensively explore the supervised pre-training of SOT-based multi-talker ASR system with large-scale simulation. Experiments are conducted with 75 thousand (K) hours of single-talker data to simulate a total of 900K hours of multi-talker audio segments for pre-training. For the AMI-SDM evaluation set, the proposed multi-talker model with large-scale pre-training achieves a substantially better WER than the previously known results after fine-tuning on the AMI-SDM training set.

2. SOT-Based Multi-Talker ASR

2.1. ASR based on attention encoder decoder

Given input $X \in \mathbb{R}^{f^a \times l^a}$, where f^a and l^a are the feature dimension and the sequence length, respectively. The goal of ASR system is to estimate transcription $Y = (y_n \in \{1, ..., |\mathcal{V}|\} | n = 1, ..., N)$, where $|\mathcal{V}|$ is the size of the vocabulary $\mathcal{V}$, and N is the number of estimated tokens.

[1]Concurrently with our work, Chan et al. [28] proposed to mix various corpora to pre-train a single large ASR, showing a WER of 21.7% for AMI-SDM evaluation set.

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-102

In this paper, we use the attention-based encoder-decoder (AED) [33, 34] as the backbone of the ASR system, which is represented as follows:

$$H = \text{Encoder}(X), \tag{1}$$

$$o_n = \text{Decoder}(y_{[1:n-1]}, H). \tag{2}$$

The Encoder module first converts X into a sequence of hidden embeddings $H \in \mathbb{R}^{f^h \times l^h}$ for ASR (Eq. (1)), where f^h and l^h are the embedding dimension and the sequence length, respectively. At each decoder step n, the Decoder module calculates the output distribution $o_n \in \mathbb{R}^{|\mathcal{V}|}$ given previous token estimates $y_{[1:n-1]}$ and H (Eq. (2)). The posterior probability of token i (i.e. the i-th token in the vocabulary $\mathcal{V}$) at the n-th decoder step is represented as

$$Pr(y_n = i | y_{[1:n-1]}, X) = o_{n,i}, \tag{3}$$

where $o_{n,i}$ represents the i-th element of o_n. The posterior probability of token Y given input X is represented as,

$$Pr(Y|X) = \prod_{n=1}^{N} Pr(y_n | y_{[1:n-1]}, X). \tag{4}$$

In this paper, we use a modified version of Conformer network [35] for Encoder module, and a conventional Transformer-based decoder [36] for Decoder module. The modifications we have made to the Conformer network are as follows: (i) we insert a squeeze and excitation module [37] just before the dropout of the convolution module; (ii) we do not use batch normalization in the convolution module; and (iii) we add one more point-wise convolution after depth-wise convolution. These changes have been made based on our preliminary test.

2.2. SOT for multi-talker ASR

SOT was proposed for the AED to recognize a variable number of speakers from possibly overlapped audio [24]. In the SOT framework, the multiple utterances are concatenated to form a single token sequence by inserting a special symbol $\langle sc \rangle$ representing a speaker change. For example, for the three-speaker case, a reference token sequence is given as $R = \{r_1^1, .., r_{N^1}^1, \langle sc \rangle, r_1^2, .., r_{N^2}^2, \langle sc \rangle, r_1^3, .., r_{N^3}^3, \langle eos \rangle\}$, where r_i^j represents the i-th token of the j-th speaker. Here, $\langle eos \rangle$, a token for sequence end, is used only at the end of the entire sequence. In the inference, the decoding process is iterated until $\langle eos \rangle$ is detected so that the SOT-based model can theoretically transcribe utterances of any number of speakers while automatically estimating the number of speakers.

There are multiple ways to determine the speaker order to form reference label sequence R. One simple yet effective approach proposed in [24] is sorting the reference labels by their start times, which is called "first-in, first-out" (FIFO) training. FIFO training works with complexity of $O(S)$ with respect to the number of speakers S while showing superior accuracy than a scheme that exhaustively considers all possible permutations [24]. In this paper, we always use this FIFO training scheme.

3. Experimental Settings

3.1. Data

3.1.1. Simulated multi-talker audio based on 75K-hour data

We used 64 million anonymized and transcribed English utterances, totaling 75K hours, for the data simulation to pre-train the multi-talker ASR models. The data includes audio from various domains such as voice search and dictation. Each audio is assumed to contain single speaker's voice. Although the data could contain untranscribed interference speaker's speech in the background noise, we did not apply any filtering.

For the pre-training, we simulated multi-talker recordings on-the-fly in our data loader module. Specifically, we picked N audio samples from the 75K-hour data by assuming they are from different speakers and mixed them by adding delay to each sample, where N was uniformly sampled from 1 to 5. The delay amount was also randomly sampled under the constraints that there was at least 0.5 sec of difference in the start time of each audio sample and that each audio sample had at least one overlapping region with another sample. After mixing the speech samples, we applied speed perturbation [38] of 0.9–1.1x to further increase the data variation. Thanks to the on-the-fly simulation, there was almost no duplication of simulated data through the entire pre-training process.

Table 1: *Statistics of AMI corpus.*

(a) utterance

	# of speakers	# of segments	Average dur. (sec)	Total dur. (hr)	# of words
train	1	107,319	2.6	76.9	793,856
dev	1	13,098	2.5	8.9	94,953
eval	1	12,643	2.5	8.7	89,666

(b) utterance group

	# of speakers	# of segments	Average dur. (sec)	Total dur. (hr)	# of words
train	1	33,646	3.0	28.5	298,157
	2	12,914	5.5	19.6	239,820
	3	5,450	8.2	12.5	171,266
	4	1,779	11.6	5.7	84,514
	5	3	6.7	0.006	99
	Total	53,792	4.4	66.2	793.856
dev	1	4,280	2.9	3.4	36,745
	2	1,578	4.9	2.2	27,675
	3	680	7.3	1.4	19,895
	4	254	9.5	0.7	10,638
	Total	6,792	4.1	7.6	94,953
eval	1	3,956	3.0	3.3	34,076
	2	1,347	5.1	1.9	24,036
	3	631	8.0	1.4	20,276
	4	203	13.2	0.7	11,278
	Total	6,137	4.3	7.3	89,666

3.1.2. AMI meeting corpus

We used the AMI meeting corpus [2] for the evaluation as well as the fine-tuning of the pre-trained models. The corpus comprises approximately 100 hours of meeting recordings, each containing three to five participants. The audio was recorded by an 8-ch microphone array, which is often called multiple distant microphone (MDM). The first channel of the MDM audio is used for monaural ASR evaluation, referred to as a single distant microphone (SDM) setting. The AMI corpus also contains the recordings from independent headset microphones (IHM) worn by each participant.

As mentioned earlier, the primary focus of this paper is the evaluation with SDM recordings. Meanwhile, we sometimes used the MDM or IHM recordings for analysis purposes. We used scripts in Kaldi toolkit [39] to partition the AMI corpus into training, development and evaluation recordings. The statistics about the AMI meeting corpus are shown in Table 1. The definition of "utterance group" in the table will be explained in the next section.

Table 2: *WER (%) for AMI test set with various configurations. Single Distant Microphone (SDM) evaluation is our primary focus in this paper, and the results with Multiple Distant Microphone (MDM), i.e., 8-ch microphone array, are shown as reference numbers. A result with † is reported in a concurrent work [28] in which a mixed corpus including AMI is used.*

Config. ID	Audio device	ASR Architecture	ASR Pre-training	ASR Fine-tuning	Front-end	Evaluation segment	WER (%) dev	WER (%) eval
Kanda et al. [10]	SDM	CNN-TDNN-BLSTM hybrid	-	AMI	-	utterance	33.4	36.4
Kanda et al. [10]	MDM	Multi-ch CNN-TDNN-BLSTM hybrid	-	AMI	BeamformIt	utterance	30.1	32.3
Chan et al. [28]	SDM	Conformer RNN-T	Mixed (incl. AMI)	-	-	utterance	-	21.7†
1	SDM	Single-talker Conformer AED	-	AMI	-	utterance	47.5	51.1
2	SDM	Single-talker Conformer AED	75K	-	-	utterance	35.3	40.1
3	SDM	Single-talker Conformer AED	75K	AMI	-	utterance	23.0	25.8
4	MDM	Single-talker Conformer AED	75K	AMI	BeamformIt	utterance	22.2	24.4
5	SDM	SOT Multi-talker Conformer AED	-	AMI	-	utterance group	57.2	59.1
6	SDM	SOT Multi-talker Conformer AED	75K	-	-	utterance group	41.6	44.2
7	SDM	SOT Multi-talker Conformer AED	75K	AMI	-	utterance group	**18.4**	**21.2**

Table 3: *Speaker counting accuracy (%) by SOT Multi-talker AED for each utterance group of AMI-SDM evaluation set.*

Model	Actual # of Speakers	Estimated # of Speakers (%) 0	1	2	3	4	>5
Before fine-tuning (Config. 6)	1	2.1	**97.8**	0.1	0.0	0.0	0.0
	2	0.1	96.7	**3.3**	0.0	0.0	0.0
	3	0.0	95.6	4.3	**0.2**	0.0	0.0
	4	0.0	94.1	5.9	0.0	**0.0**	0.0
After fine-tuning (Config. 7)	1	0.2	**97.2**	2.5	0.1	0.0	0.0
	2	0.0	13.7	**80.5**	5.9	0.0	0.0
	3	0.0	2.4	32.6	**60.2**	4.8	0.0
	4	0.0	0.0	9.9	51.2	**38.9**	0.0

Table 4: *WER (%) for AMI-SDM evaluation set by SOT Multi-talker AED w.r.t the number of speakers in utterance group.*

	Pre-training	Fine-tuning	# of speakers in the segment 1	2	3	4	Total
Config. 5	-	AMI	37.8	59.5	76.8	88.8	59.1
Config. 6	75K	-	22.8	42.0	62.7	77.4	44.2
Config. 7	75K	AMI	14.7	19.6	25.7	35.5	21.2

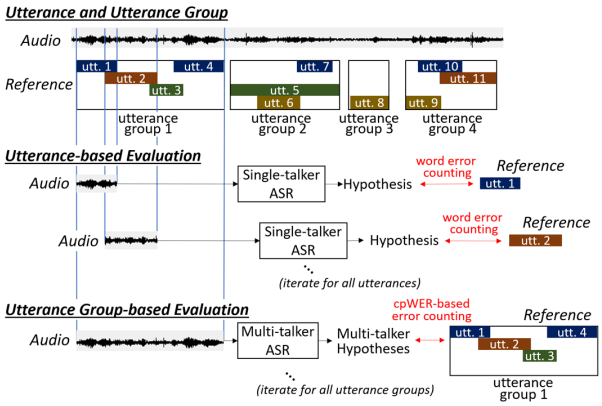

Figure 1: *Utterance and utterance group for segmentation.*

3.2. Evaluation metric

In this paper, we introduce a notion of "utterance group" to appropriately evaluate multi-talker ASR models. The relationship between the utterance and utterance group is illustrated in Fig. 1 top. The utterance group is defined as a set of utterances that are connected by speaker overlap regions. In other words, the utterance groups can be formed by segmenting the recording at silence positions or non-overlapping utterance boundaries.

The utterance-based evaluation is a standard way to evaluate single-talker ASR models for AMI. As shown in the middle of Fig. 1, ASR is applied to each speech segment obtained with reference utterance boundaries. In the MDM setting, beamforming is used before the signals are fed into the single-talker ASR system. Mis-recognized words are counted for each utterance, and the errors are summed up to calculate a final WER.

The utterance group-based evaluation is newly introduced in this paper to evaluate the multi-talker ASR models. As shown in Fig. 1 bottom, multi-talker ASR is applied to each utterance group segment without information about the utterance boundaries inside the segment. The multi-talker ASR system generates hypotheses consisting of one or more utterances from the utterance group audio. We calculate the number of mis-recognized words based on the concatenated minimum-permutation word error rate [6]. Specifically, we first concatenate the reference transcriptions of the same speaker in the utterance group (i.e., in the example of Fig. 1, utterance #1 and #4 in utterance group #1). Then, the best alignment between the hypotheses and the concatenated references is selected among all possible permutations[2] for error counting. Finally, the errors from each utterance group are summed up and divided by the number of total reference words to calculate the WER. Note that the denominator to calculate the WER, i.e. the number of the total reference words, is the same for the utterance-based evaluation and the utterance group-based evaluation.

3.3. ASR model configuration

In our investigation, we trained single-talker ASR models as well as SOT-based multi-talker ASR models. We used the same encoder-decoder architecture for all cases. Specifically, the encoder consisted of 2 layers of convolution layers that subsample the time frames by a factor of 4, followed by 18 conformer layers. Each conformer layer consisted of two 1024-dim feed forward layers in a sandwitch structure, a multi-head attention with 8 heads, a depth-wise convolution with kernel size 3, and a squeeze-and-excitation network with reduction factor 8 [37]. The embedding dimension was set to 512. The decoder consisted of 6 layers, each of which had a multi-head attention with 8 heads and a 2048-dim feed forward layer. 4K subwords [40] were used as a recognition unit. We used a 80-dim log mel filterbank extracted every 10 msec for the input feature.

The single-talker model was pre-trained by the 75K-hour data while the multi-talker model was pre-trained by the 75K-hour-based simulated multi-talker data. For both models, we

[2]If the number of hypotheses are larger than that of references, all hypotheses that don't have corresponding references are counted as insertion errors. Similarly, if the number of references are larger than that of hypotheses, all unmatched references are counted as deletion errors.

performed 425k training iterations with 32 GPUs, each of which consumed mini-batches of 24,000 frames (roughly corresponding to 900K hours of data simulation and consumption). We used Adam optimizer with a linear decay learning rate schedule with a peak learning rate of 1e-3 after 25k warm up iterations. In the fine-tuning stage, the pre-trained single-talker model was fine-tuned by the AMI-SDM *utterance* segments while the pre-trained multi-talker model was fine-tuned by the AMI-SDM *utterance group* segments with the speaker-based FIFO training scheme [26]. For both cases, we used speed perturbation and SpecAugment [41], and conducted 25k training iterations with 16 GPUs, each of which consumed mini-batches of 6,000 frames. A linear decay learning rate schedule starting at a learning rate of 1e-4 was used. Besides the pre-training-based two-stage approach, we also trained models on the AMI-SDM data without pre-training. In this case, we used mini-batches of 6,000 frames and trained the models for 110k iterations with 16 GPUs, with a linear decay learning rate schedule with a peak learning rate of 1e-4 after 10k warm up iterations.

4. Evaluation Results

4.1. Main results

Table 2 shows our main results comparing single-talker ASR and SOT-based multi-talker ASR models. The two-stage approach of performing both pre-training and fine-tuning significantly improved the WER for both single-talker (Config. 3) and multi-talker models (Config. 7). The two-stage models outperformed both the 75K-hour-only models and AMI-only models with significant margins. Interestingly, although the single-talker model outperformed the multi-talker model before fine-tuning (Configs. 2 and 6), the multi-talker model showed superior performance after fine-tuning (Configs. 3 and 7). Table 3 shows that fine-tuning significantly improved the speaker counting accuracy of the multi-talker model. This improvement in speaker counting led to the larger WER improvement of the multi-talker model especially for segments with many speakers as shown in Table 4 (Config. 6 vs. Config. 7). Table 4 also shows the improvement provided by the pre-training (Config. 5 vs. Config. 7), where large gains were observed for segments with more speakers. This is because such segments were more scarce in the real data as shown in Table 1 (b).

While our main focus in this paper is SDM-based recognition, we also evaluated the fine-tuned single-talker ASR model for the MDM setting using BeamformIt beamformer [42]. The result is shown in Config. 4 of Table 2. We observed that the proposed SOT multi-talker AED model (Config. 7) outperformed the combination of the single-talker AED and MDM-based beamforming while the SOT multi-talker AED model even does not require the utterance-level boundary information. This result shows the effectiveness of the end-to-end multi-talker ASR model that jointly performs speaker counting, speech separation, and speech recognition.

4.2. Why fine-tuning by real data is effective

To better understand why fine-tuning worked so effectively, we conducted an experiment using the AMI-IHM training data for fine-tuning. Table 5 shows the results. Here, we prepared two new datasets for fine-tuning: "IHM real-mix" and "IHM rand-mix". "IHM real-mix" was generated by first mixing all IHM recordings in the same meeting and then segmenting the mixed audio by the utterance groups. The difference from "SDM" to "IHM real-mix" lies only in the noise and reverberation characteristics. On the other hand, "IHM rand-mix" was generated by randomly mixing the AMI IHM utterances for up to 4 utterances. The difference between "IHM real-mix" and "IHM rand-mix" resides in the speaker overlap pattern. The biggest WER difference was observed between the no-fine-tuning (1st row) and the fine-tuning by "IHM rand-mix" (2nd raw), which suggested the fine-tuning of the language model (LM), English accent, and the accurate overlap of speech region[3] had the biggest impact. The second largest WER difference was observed between "IHM rand-mix"-based fine-tuning and "IHM real-mix"-based fine-tuning, showing the importance of including the real overlapping pattern. The difference from "IHM real-mix" to "SDM" was relatively small presumably because the pre-trained representation was already robust to noisy samples.

Table 5: *WER (%) of AMI-SDM test set by SOT Multi-talker AED with different fine-tuning data.*

Training data for fine-tuning	LM, accent, & accurate overlap	Overlap pattern	Noise & reverb.	dev	eval
-	-	-	-	41.6	44.2
IHM rand-mix	√	-	-	25.0	29.0
IHM real-mix	√	√	-	19.6	23.0
SDM	√	√	√	18.4	21.2

Table 6: *WERs (%) from IHM to SDM settings. The single-talker AED was used for the utterance evaluation while the multi-talker AED was used for the utterance group evaluation.*

Evaluation data	Evaluation segment	Microphone distance?	Speech overlapped?	Utterance boundary?	WER (%) dev	eval
IHM	utterance	close	no	given	12.8	12.2
IHM-mix	utterance group	close	yes	n/a	13.5	14.9
SDM	utterance group	distant	yes	n/a	18.4	21.2

4.3. Gap between SDM and IHM-based recognition

We further conducted the evaluation to understand the gap between SDM-based recognition and IHM-based recognition, the results of which are shown in Table 6. We prepared two additional evaluation sets, "IHM" and "IHM-mix". "IHM-mix" was generated by first mixing all IHM recordings in the same session and then segmenting the mixed audio based on the utterance group. The difference between "SDM" and "IHM-mix" is whether the audio is recorded by SDM or IHM. On the other hand, the difference between "IHM-mix" and "IHM" includes whether the speech is overlapped and whether we use additional utterance boundary information. While we observed a significant gap between SDM and IHM-mix, the WER difference between IHM-mix and IHM was relatively small. This indicates that the SOT AED already worked well in terms of speaker counting and multi-talker transcription and that the refinement of the model capacity or the data simulation is rather necessary to better cope with noisier data.

5. Conclusions

In this paper, we extensively explored the pre-training of SOT-based multi-talker ASR with large-scale simulation. In the evaluation, the proposed multi-talker model achieved state-of-the-art WER of 21.2% for AMI-SDM evaluation set, even outperforming the combination of MDM-based speech enhancement and strong single-talker ASR.

[3] In pre-training, the simulated mixed audio may not have the overlap of speech region due to the silence of the original audio. Also, the audio used in the simulation could contain interference speech without transcription. We think these factors may be also refined by fine-tuning.

6. References

[1] A. Janin *et al.*, "The ICSI meeting corpus," in *Proc. ICASSP*, vol. 1, 2003, pp. I–I.

[2] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*, 2005, pp. 28–39.

[3] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*, 2007, pp. 373–389.

[4] Ö. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *Proc. Interspeech*, 2006.

[5] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang *et al.*, "Advances in online audio-visual meeting transcription," in *Proc. ASRU*, 2019, pp. 276–283.

[6] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. CHiME 2020*, 2020.

[7] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[8] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2017.

[9] S. Ganapathy and V. Peddinti, "3-d CNN models for far-field multi-channel speech recognition," in *ICASSP*, 2018, pp. 5499–5503.

[10] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, and S. Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *Proc. ICASSP*, 2019, pp. 6630–6634.

[11] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. on ASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[12] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. on ASLP*, vol. 20, no. 2, pp. 356–370, 2012.

[13] 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," *arXiv preprint arXiv:2101.09624*, 2021.

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

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

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

[17] 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.

[18] W. Zhang, X. Chang, Y. Qian, and S. Watanabe, "Improving end-to-end single-channel multi-talker speech recognition," *IEEE/ACM Trans. on ASLP*, vol. 28, pp. 1385–1394, 2020.

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

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

[21] I. Sklyar, A. Piunova, and Y. Liu, "Streaming multi-speaker asr with rnn-t," *arXiv preprint arXiv:2011.11671*, 2020.

[22] L. Lu, N. Kanda, J. Li, and Y. Gong, "Streaming end-to-end multi-talker speech recognition," *arXiv preprint arXiv:2011.13148*, 2020.

[23] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[24] 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.

[25] 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.

[26] N. Kanda *et al.*, "Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings," in *Proc. SLT*, 2021, pp. 809–816.

[27] X. Chang, N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Hypothesis stitcher for end-to-end speaker-attributed asr on long-form multi-talker recordings," in *Proc. ICASSP*, 2021.

[28] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "SpeechStew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[29] Y. Huang, D. Yu, C. Liu, and Y. Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Proc. Interspeech*, 2014, pp. 2977–2981.

[30] V. Joshi, R. Zhao, R. R. Mehta, K. Kumar, and J. Li, "Transfer learning approaches for streaming end-to-end speech recognition system," in *Proc. Interspeech*, 2020, pp. 2152–2156.

[31] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020.

[33] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *NIPS Workshop on Deep Learning*, 2014.

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

[35] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," *Proc. Interspeech*, pp. 5036–5040, 2020.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NIPS*, 2017, pp. 6000–6010.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[38] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[39] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[40] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[41] 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," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[42] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.