THE ROYALFLUSH SYSTEM OF SPEECH RECOGNITION FOR M2MET CHALLENGE

Shuaishuai Ye, Peiyao Wang, Shunfei Chen, Xinhui Hu, and Xinkang Xu

Hithink RoyalFlush AI Research Institute, Zhejiang, China

{yeshuaishuai, wangpeiyao, chenshunfei, huxinhui, xuxinkang}@myhexin.com

ABSTRACT

This paper describes our RoyalFlush system for the track of multi-speaker automatic speech recognition (ASR) in the M2MeT challenge. We adopted the serialized output training (SOT) based multi-speakers ASR system with large-scale simulation data. Firstly, we investigated a set of front-end methods, including multi-channel weighted predicted error (WPE), beamforming, speech separation, speech enhancement, etc., to process training, evaluation, and test sets. However, according to their experimental results, we only selected the WPE and beamforming approach as our front-end methods. Secondly, we made great efforts in the data augmentation for multi-speaker ASR, including adding noise and reverberation, overlapped speech simulation, multi-channel speech simulation, speed perturbation, front-end processing, etc., which brought us a significant performance improvement. Finally, to make full use of the performance complementary of different model architecture, we trained the standard conformer based joint CTC/Attention (Conformer) and U2++ ASR model with a bidirectional attention decoder, a modification of Conformer, to fuse their results. Compared with the official baseline system, our system got a 12.22% absolute Character Error Rate (CER) reduction on the evaluation set and 12.11% on the test set.

Index Terms— Multi-channel and multi-speaker ASR, Overlapped speech, Data augmentation, Conformer, U2++, Model fusion

1. INTRODUCTION

Although ASR has achieved superior performance in many tasks [1, 2], meeting transcription is still considered as one of the most challenging tasks [3]. Due to the distance between the speaker and microphone [4], Meeting speech often contains noise and reverberation. In the meantime, multiple speakers talking in a free speaking style cause overlapped speech, and the sentences are usually less grammatical in verbal communication [3]. What is more, the number of speakers cannot be determined. These problems drastically decrease the accuracy and robustness of the ASR system, especially in real-world applications.

Far-field speech with complicated noise and reverberation in large meeting rooms is complex in the acoustic domain. Many researchers have made great efforts in improving the accuracy and robustness of far-field ASR from various aspects. Front-end processing techniques play an essential role in multi-channel ASR, such as WPE dereverberation [5] and various beamforming methods (MVDR [6], neural network beamformer [7]). In addition, data augmentation techniques also play a crucial role in addressing data sparsity issues and the robustness of the model for far-field ASR [2, 8, 9]. Some researches pay more attention to data augmentation based model training [10–12]. In building ASR models,

the front-end speech enhancement and the back-end ASR has been combined in some studies [13, 14].

Overlapped speech is another common phenomenon in multi-talker meeting, and it has a great impact on the accuracy of ASR. Many researches have made tremendous efforts towards overlapped ASR. A standard solution is to construct a cascaded system with speech separation and ASR [15, 16]. However, due to the inconsistency between the objectives of multiple modules and the training data, the performance of the cascaded system tends to decline in the real world. Permutation invariant training(PIT) has been introduced into the multi-speaker ASR since it can consider all possible permutations of speakers [17, 18]. Although PIT-based ASR system works well for overlapped speech recognition, the output layer of the model is limited to the number of speakers. It cannot handle the dependency among utterances of multiple speakers [19, 20]. In recent years, the SOT method was proposed for solving the overlapped ASR by introducing a special symbol to represent the speaker change [19–21]. It can solve the disadvantage of PIT and naturally model the dependency among the outputs for multiple speakers.

In this work, we adopted SOT based multi-speakers ASR system. In order to make full use of the performance complementary of different model architecture, we built two acoustic models of ASR system with the joint CTC/Attention End-to-End (E2E) framework: U2++ [22] and Conformer [23]. we employed various data augmentation techniques for ASR. (1) We first use the single-channel data to simulate large multi-channels data. (2) The single-speaker data from near and far-field data is used for simulating large multi-talker speech data. (3) Based on two simulated data and real training data, we exploited strategies of front-end data processing methods to process multi-channel data, such as beamforming, WPE, and simulated equalization (EQ). (4) After that, we also used conventional data augmentation methods for all data, such as adding noise and reverberation, speed perturbation, pitch perturbation, and spectrum augmentation. Using these augmentations, We obtained about 18,000-hour effective training data. In addition, an LSTM based language model was also trained for decoding. Finally, we employed the ROVER tool to fuse the multiple results. Comparing with the official baseline system, the fusion result achieved a 12.22% absolute CER reduction on the evaluation set, from 29.7% to 17.48% and 12.11% on test set, from 30.9% to 18.79%.

The rest of the paper is organized as follows. In Section 2, we will introduce the M2MeT challenge and the details of data preparation for the challenge. Section 3 presents our speech recognition system. The experiments and results are shown in Section 4. Finally, Section 5 concludes the paper and discusses the future work.

2. M2MET CHALLENGE AND DATA PREPARATION

In this section, we first introduce the Multi-channel Multi-party Meeting Transcription (M2Met) challenge, which we participated
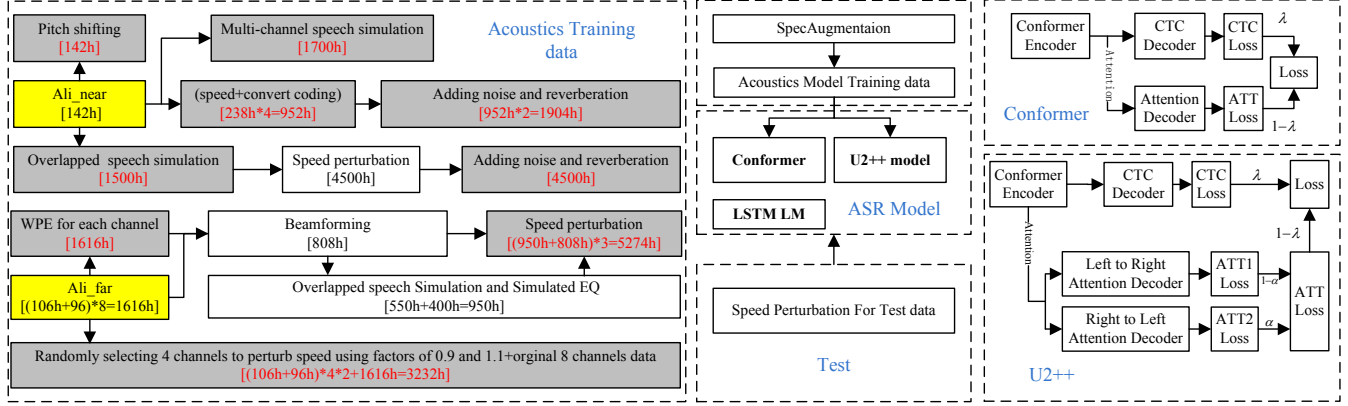
Fig. 1. The left part is the acoustics training data, where we listed our all data augmentation methods, the yellow blocks are original data set and the grey blocks were used to train our models. In the training stage, the utterances greater than 30 seconds were excluded, so the final duration of training set is about 18,000 hours. The middle part is our system architecture, including training data, ASR model and test data processing. The right part is two ASR model architectures.

in, and its dataset, namely AliMeeting and Aishell-4. Then, we will describe our data augmentation methods, including speed perturbation, adding noise and reverberation, pitch shifting, multi-channel far-field speech simulation, overlapped speech simulation, spectrum augmentation and simulated equalization.

2.1. M2Met challenge

In order to inspire research on advanced meeting rich transcription, Alibaba Group launched the M2MeT challenge, as an ICASSP2022 Signal Processing Grand Challenge [24]. The challenge consists of two tracks, namely speaker diarization and multi-speaker ASR. For the track of multi-speaker ASR, the organizer also set up two sub-track: the first sub-track limits the data usage while the second sub-track allows the participants to use extra constrained data. This work describes our submitted system to the challenge, which is only for the first subtrack of the multi-speaker ASR.

2.2. Data preparation

In the M2MeT challenge, AliMeeting [24] dataset was presented along with Aishell-4 [25] as the training data. The AliMeeting dataset is a sizeable Mandarin meeting corpus containing 118.75 hours of real meeting data recorded by 8-channels directional microphone array and headset microphone. The detailed recorded configuration refers to [24]. The Aishell-4 dataset, an 120 hours open-source dataset for speech enhancement, separation, recognition, and speaker diarization in conference scenarios, covers a variety of aspects in real-world meetings, including diverse recording conditions, various number of meeting participants, various overlap ratios noise types, and meets the modern requirement of accurate labeling. In the training phase, the AliMeeting is divided into 104.75 hours for training, 4 hours for evaluation, and 10 hours as a test set, and the Aishell-4 is divided into 107.5 hours for training, and 12.72 hours for evaluation.

2.3. Data Augmentation

Data augmentation plays an important role in model training, especially in the low resource scenario, and it can significantly increase data diversity and quantity. Our data augmentation methods are as follows:

- **Speech perturbation**: Speech perturbation [8] is an audio-level speech augmentation method. In this work, It produces 3 variants of the original signal with speed factors of 0.9, 1.0, and 1.1.

- **Adding noise and reverberation**: Adding noise and reverberation to corrupt clean speech can simulate the noisy and reverberant conditions for some rooms. In our system, the noise sources are composed of MUSAN corpus [26], excluding bable noise and AliMeeting noise source extracted from the AliMeeting training set according to timestamp. The reverberation sources include two parts, RIRS_NOISES corpus [27] and AliMeeting single-channel reverberation source simulated based on image method [28] according to the room sizes of AliMeeting using a public open-source code[1].

- **Pitch shifting**: One copy was created for each original data by modifying the pitch. We used the *Sox* audio manipulation tool [29] to realize pitch shifting.

- **Multi-channel far-field speech simulation**: For speech data captured by a headset microphone, we conducted a multi-channel data simulation to make full use of it under the far-field environment. Concretely speaking, we generated 17,400 multi-channel room impulse responses (RIRs) based on the configuration of the microphone array with the image method [28]. 2400 of these RIRs were generated according to the AliMeeting room information, and the others were generated according to random room size, which ranges from $3*3*2.5m^3$ to $10*10*3.5m^3$. Then we took the headset data as the sound source, converted it into 8-channel far-field data with the simulated RIRs, and added noise from different directions to simulate the array data.

- **Overlapped speech simulation**: According to [24], the training and evaluation data's average speech overlap ratio is 42.27% and 34.76%, respectively. In order to improve ASR model performance for recognizing the overlapped speech, we employed the non-overlapped speech in the data sets recorded by an 8-channel directional microphone array and headset microphone to simulate a large amount of overlapped speech data with proportions of 50% for 2 speakers, 30% for 3 speakers, 20% for 4 speakers, respectively.

- **Spectrum augmentation (SpecAugment)**: For improving the model performance and the robustness of our ASR model, we

---

[1] https://github.com/ehabets/RIR-Generator

applied the SpecAugment method [2] to all input features of a neural network at a mini-batch level. The SpecAugment includes warping the features, masking blocks of frequency channels, and masking blocks of time steps, and we only utilized the last two tricks in our system. We set the parameters of SpecAugment to default values in WeNet [22] and EspNet [30] tools.

- **Simulated equalization(EQ)**: Inspired by [31], we used various filters to process data, including weighted low-pass/high-pass filter, de-emphasis filter, simulated frequency response curve filter and so on.

3. SYSTEM DESCRIPTION

In this section, we will elaborate on our ASR system submitted to the M2MeT challenge. The system diagram is shown in the Figure 1. Our ASR system mainly consists of front-end processing, data preparation, acoustic model, language model, and model fusion. For the reason that data augmentation has been introduced in section2, here we will describe the remaining parts.

3.1. Front-end processing

The WPE [5] has been shown to be an effective method for speech dereverberation in far-field scenarios, thus improving the recognition performance. It is a kind of long-term linear prediction method to estimate the reverberation tail in the speech signal and subtract it from the observed signal. The original assumption of the WPE method is that there is no noise and overlapped speakers. However, in our experiments, it is found that the WPE can improve the performance of the speech recognition system even in an environment with additive noise and overlapped speech. The same conclusion was also obtained in [7].

After the WPE was used for dereverberation, the weighted delay-and-sum beamforming and scheme reference from the beamformIt [32] were further utilized for processing. We select a reference microphone by calculating the crosscorrelations of two microphone pairs. Then it uses Generalized Cross Correlation with Phase Transform (GCC-PHAT) method to calculate the time delays of arrival(TDOA) between each microphone and the reference microphone. Furthermore, the best TDOA is estimated by a 2-step Viterbi search method. Finally, the signals are aligned and summed by each channel's estimated TDOA and weights.

3.2. Acoustic model

Different model architectures have different modeling capabilities for ASR, so they can complement each other to improve system performance. For this reason, we adopted two model architectures in our system, namely the standard Conformer based joint CTC/Attention ASR model (Conformer) [23] and the Conformer based U2++ ASR model [33], as shown in the right part of Figure 1.

The Conformer consists of: a shared encoder, an attention decoder, and a CTC module seeing the right-top part of Figure1. The U2++ model has been made two modifications based on the Conformer mentioned above; one is to modify the attention decoder into a bidirectional decoder, seeing the right-bottom part of the Figure1 and the other is to propose attention rescoring, a new decoding mode. The more descriptions for Conformer and U2++ refer to [23] and [33].

3.3. Language model

For the Conformer model, a recurrent neural network language model (LM) was trained by the Espnet [30]. The training transcripts consist of two parts: One part is the original training data, and the other one is an augmentation text by concatenating several consecutive sentences [34]. The model architecture consist of 2 LSTM layers with 1024 dimensions and the model unit is character.

3.4. Model fusion

To fuse the above mentioned ASR systems, the fusion strategy Rover [35] is used. It is a system developed at NIST to produce a composite ASR system output when the outputs of multiple ASR systems are available. The Rover system implements a voting or rescoring process to reconcile differences in ASR system outputs.

We prepared the candidate results from U2++ and conformers without/with LM and performed speed perturbation to the test set with a speed factor of 0.9 and 1.1, respectively.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

The results of M2MeT are summarized in [36], where our system is denoted as 'R62'. Our experiments and results are summarized as follows.

All our experiments were conducted on the EspNet for the Conformer and the WeNet for U2++ model. The acoustic feature of 83 dimensions, 80-dimensional Fbank plus the 3-dimensional pitch feature was extracted from every frame with a frame length of 25ms and a frame shift of 10ms.

The standard conformer based on joint CTC/Attention model is composed of a 12-layer encoder with 2048-dim feed-forward and depth-wise convolution with kernel size 3, a 6-layer decoder with 2048 units and 8 heads attention with 512 dimensions. The U2++ model was configured with the same settings as the aforementioned standard model, expect for an 18-layer encoder and modifying a 6-layer decoder into a 3-layer left-to-right decoder and a 3-layer right-to-left decoder. Both models employed the characters in training data as the modeling unit. We set both the CTC weight and the reverse weight to 0.3 in the training phase and set 0.3 and 0.6 respectively in the decoding phase. The beam size was set to 20. We trained 30 epochs for the U2++ model and 11 epochs for the Conformer model with the same learning rate of 0.001. During the training stage of the Conformer model, the first 10 epochs is trained using 10000 hours and the last epoch is trained using 14000 hours of the acoustics training data. We averaged the best 10 and the best 8 models during the decoding stage based on the loss value as the final model for the U2++ and Conformer models respectively.

4.2. Results and analysis

Due to the different hyper-parameters of the two models trained in our system, their results are not comparable. In this section, we mainly conducted different experiments on two models, respectively. Our experiments were conducted on the following several aspects to investigate the influence on the ASR model: (1) SOT scheme; (2) the amount of training data; (3) the number of the channel used to train model; (4) front-end methods, namely the WPE and beamforming approach; (5) model fusion.

Table 1. The results of different models on the AliMeeting far-field evaluation and test set. 'wb' refers to 'wpe-bf', and 'Eval' refers to evaluation set. 'U2++_*channel' represents the number of channels. 'U2++_big_data' represents the final model trained by all data.

Group	System	Eval (CER%)						Test (CER%)		
		Eval-1c	Eval-wpe	Eval-bf	Eval-wb	Eval-wb-sp0.9	Eval-wb-sp1.1	Test-wb	Test-wb-sp0.9	Test-wb-sp1.1
Group1	baseline [24]	30.8	-	29.7	-	-	-	30.9		
Group2	Conformer_Ali-near	47.44	-	-	-	-	-	-	-	-
	+SOT	46.06	-	-	-	-	-	-	-	-
	Conformer_5k-hours	24.02	24.99	23.32	22.77	-	-	23.15	-	-
	Conformer_10k-hours	21.49	21.71	20.27	**19.06**	**19.74**	**19.31**	**20.14**	**20.65**	**20.68**
	+LSTM LM	21.64	21.84	20.52	**19.14**	19.90	19.59	**20.29**	-	-
Group3	U2++_1channel	28.49	28.95	26.83	26.13	-	-	26.78	-	-
	U2++_4channels	27.93	28.47	26.56	26.04	-	-	26.64	-	-
	U2++_8channels	27.45	27.86	26.15	25.50	-	-	26.20	-	-
	U2++_big_data	21.10	21.08	19.92	**18.68**	**18.82**	**19.12**	**19.99**	**20.10**	**20.46**
Group4	Fusion*	**17.48**						**18.79**		

* The numbers with bold font are fused as the final result using the *ROVER* tool.

4.2.1. SOT scheme

In the experiments of the conformer model, we verified the effectiveness of the SOT method on the same duration of the simulated overlapped near-field speech data Ali-near. The results on the evaluation set are shown in first two lines of Group2 in Table 1. Compared with a model trained using the Ali-near, we can find that the SOT method can significantly improve the performance. Therefore, the following experiments were conducted by adopting the SOT method. More detailed experimental verification for SOT refers to [20].

4.2.2. The number of channel for training model

In order to verify whether the number of channel used to train the model will affect the model's performance, we carried out three experiments on U2++ ASR model, and the results are shown in the first three lines of Group3 in Table 1. With the increasing number of channels, we can see that the performance has also been improved to varying degrees. We inferred that the gain is mainly caused by channel differences which can improve the robustness of the ASR model. So, we used 8-channel far-field data to simulate overlapped speech and trained the ASR models in the following experiments.

4.2.3. The amount of training data

With the increase of training data, the performance of the ASR model generally becomes improving. We conducted 2 experiments with a 5000-hour and a 10000-hour of training data for the Conformer model as shown in the Group2 of Table 1 and 4 experiments in the Group3 for U2++ model. Compared with the baseline model [24], we find that the conformer model and the U2++ model achieved 9.31% and 9.7% absolute CER reduction on the Ali-1c without front-end data, and 10.61% and 10.91% absolute CER reduction on the test set, respectively. The significant performance improvement was mainly attributed to the data augmentation. Based on these facts, we see that a more considerable amount of training data can obtain better performance. Therefore, we can conclude that large-scale training data with a set of data augmentation methods can effectively improve the performance and robustness of the model.

4.2.4. Front-end

Since the WPE can only reduce the reverberation tail, its benefit to the improvement of the ASR performance is not so large. The beamforming can mask part of noise and reverberation relying on spatial information, and it is capable of improving the signal-to-noise ratio (SNR) in the speaker direction. So, by combining the WPE and beamforming methods, both their advantages can be utilized. As shown in Table 1, the beamforming can significantly decrease the CER of the evaluation and test set. Furthermore, the performance can be prominently improved when WPE and beamforming are used successively.

4.2.5. Model fusion

In order to further reduce the CER on the evaluation set, we fused 7 subsystems with ROVER tool, as the bold font in Table 1. Compared with the best single system, the fusion scheme brought a CER absolute reduction of 1.2% from 18.68% to 17.48% on the evaluation set and 1.2% from 19.99% to 18.79% on the test set. The possible reasons for the improvement were summarized as follows: (1) With the perturbing processing, the incorrectly recognized characters caused by those speeches that are too fast or too slow can be recognized correctly. (2) As a supplementary result to the original utterance, the LSTM based language model can correct some syntax error.

5. CONCLUSION

This paper described our multi-speaker ASR system for the M2MeT challenge. Two E2E ASR frameworks, Conformer and U2++, were adopted with SOT method. Various front-end processing methods, data augmentations, model fusion, and LSTM LM, were investigated. Experimental results showed that the WPE and beamforming based on multi-channel can effectively decrease the CER. We found that using a variety of data augmentations to expand the training set can significantly improve performance and robustness. We also found that the fusion of different models and speed perturbation can achieve 1.2% absolute CER reduction on the evaluation set compared with the best result of the U2++ ASR model. Comparing with the official baseline system, our system got a 12.22% absolute CER reduction on the evaluation set, from 29.7% to 17.48% and 12.11% absolute CER reduction on the test set, from 30.9% to 18.79%.

6. REFERENCES

[1] K. J. Han, R Prieto, and T. Ma, "State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions," in *ASRU*, 2019.

[2] Daniel S. Park, William Chan, Yu Zhang, and et al., "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep 2019.

[3] N. Kanda, G. Ye, Y. Wu, and et al., "Large-scale pre-training of end-to-end multi-talker asr for meeting transcription with single distant microphone," 2021.

[4] Z. Etin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition," in *INTERSPEECH 2006 - ICSLP*, 2006.

[5] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 10, pp. p.2707–2720, 2012.

[6] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2009.

[7] Lukas Drude, Christoph Boeddeker, Jahn Heymann, and et al., "Integrating neural network based beamforming and weighted prediction error dereverberation.," in *Interspeech*, 2018, pp. 3043–3047.

[8] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech 2015*, 2015.

[9] E. Tsunoo, K. Shibata, C. Narisetty, and et al., "Data augmentation methods for end-to-end speech recognition on distant-talk scenarios," 2021.

[10] K. Kumar, B. Ren, Y. Gong, and J. Wu, "Bandpass noise generation and augmentation for unified asr," in *Interspeech 2020*, 2020.

[11] Shucong Zhang, Cong-Thanh Do, Rama Doddipatla, and Steve Renals, "Learning noise invariant features through transfer learning for robust end-to-end speech recognition," in *ICASSP*, 2020, pp. 7024–7028.

[12] J. Du, Y. H. Tu, L. Sun, and et al., "The ustc-nelslip systems for chime-6 challenge," in *The 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020.

[13] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," 2018, pp. 5739–5743.

[14] Tom O'Malley, Arun Narayanan, Quan Wang, and et al., "A conformer-based asr frontend for joint acoustic echo cancellation, speech enhancement and speech separation," 2021.

[15] Naoyuki Kanda, Xiong Xiao, Jian Wu, and et al., "A comparative study of modular and joint approaches for speaker-attributed asr on monaural long-form audio," 2021.

[16] N. Kanda, Y. Fujita, S. Horiguchi, and et al., "Acoustic modeling for distant multi-talker speech recognition with single- and multi-channel branches," in *ICASSP*, 2019.

[17] Y. Dong, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," 2017.

[18] I. Sklyar, A. Piunova, and Y. Liu, "Streaming multi-speaker asr with rnn-t," in *ICASSP*, 2021.

[19] R. Masumura, D. Okamura, and et al., "Unified autoregressive modeling for joint end-to-end multi-talker overlapped speech recognition and speaker attribute estimation," 2021.

[20] N. Kanda, Y. Gaur, X Wang, and et al., "Serialized output training for end-to-end overlapped speech recognition," 2020.

[21] N. Kanda, Y. Gaur, X. Wang, and et al., "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," 2020.

[22] Zhuoyuan Yao, Di Wu, Xiong Wang, and et al., "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, Brno, Czech Republic, 2021.

[23] A. Gulati, J. Qin, C. C. Chiu, and et al., "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[24] Fan Yu, Shiliang Zhang, Yihui Fu, and et al., "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*. IEEE, 2022.

[25] Y. Fu, L. Cheng, S. Lv, and et al., "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," 2021.

[26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *Computer Science*, 2015.

[27] T. Ko, V. Peddinti, D. Povey, and et al., "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics*, 2017.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1998.

[29] Sox, "audio manipulation tool," [Online]. Available: `http://sox.sourceforge.net/`.

[30] Shinji Watanabe, Takaaki Hori, Shigeki Karita, and et al., "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[31] Xinlei Ren, Xu Zhang, Lianwu Chen, and et al., "A causal u-net based neural beamforming network for real-time multi-channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 1832–1836.

[32] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio Speech & Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[33] Di Wu, Binbin Zhang, Chao Yang, and et al., "U2++: Unified two-pass bidirectional end-to-end model for speech recognition," *arXiv preprint arXiv:2106.05642*, 2021.

[34] K. Wei, P. Guo, H. Lv, and et al., "Context-aware rnnlm rescoring for conversational speech recognition," 2020.

[35] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 2002.

[36] Fan Yu, Shiliang Zhang, Pengcheng Guo, and et al., "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *Proc. ICASSP*. IEEE, 2022.