ConferencingSpeech 2021 Challenge Evaluation Plan

Wei Rao[1], *Lei Xie*[2], *Yannan Wang*[1], *Tao Yu*[3], *Shinji Watanabe*[4,5], *Zheng-Hua Tan*[6],
Hui Bu[7], *Shidong Shang*[1]

[1]Tencent Ethereal Audio Lab, Shenzhen, China
[2]Northwestern Polytechnical University, Xi'an, China
[3]Tencent Ethereal Audio Lab, Seattle, USA
[4]Carnegie Mellon University, USA
[5]Johns Hopkins University, USA
[6]Aalborg University, Denmark
[7]AI Shell Foundation, Beijing, China
ConferencingSpeech@tencent.com

1. Introduction

With the advances in video conferencing technology, we are able to seamlessly connect with people of our choice anytime anywhere in the world. Video conferencing helps break barriers of distance among people. However, during video conference, the speech quality will be significantly affected by background noise, reverberation, number of recording microphones, the layout of microphone array, the acoustic and circuit design of microphone arrays, and so on. Effective speech enhancement plays an important role in the video conferencing system. Although the performance of speech enhancement has been improved dramatically in the past several decades, there are still a set of open research problems that should be further addressed in the far-field and complex meeting room environments, which includes but not limited to:

- Multi-channel speech enhancement with single microphone array
- Multi-channel speech enhancement with multiple distributed microphone arrays
- Multi-channel speech enhancement with noisy and reverberant environments
- Multi-channel speech enhancement with low-latency and zero look-ahead (casual system)

ConferencingSpeech 2021 challenge is proposed to stimulate research in the areas mentioned above and aims for processing the far-field speech from microphone arrays in the video conferencing rooms. Besides background noise and reverberation affecting the speech quality during video conference, the layout, acoustic and circuit design of microphone array also has the impact. The challenge takes this factor into consideration by using different types of microphone arrays for recording and also explores the far-field multi-channel speech enhancement using multiple distributed microphone arrays in the real meeting room scenario. In addition, ConferencingSpeech 2021 challenge has the following features:

- To focus on the development of algorithms, the challenge requires the *close* training condition. In other words, only provided list of open source clean speech datasets and noise dataset could be used for training.
- Different from the conventional speech enhancement (denoising), the speech enhancement tasks in this challenge focus on processing the reverberant and noisy speech.
- The challenge aims to explore real-time multi-channel speech enhancement methods to achieve superior perceptual quality and intelligibility of enhanced speech with low latency and zero looking ahead.
- Targeting the real video conferencing room application, the ConferencingSpeech 2021 challenge database is recorded from real speakers. The number of speakers and the distances between speakers and microphone arrays vary according to the sizes of meeting rooms. 12 different sizes and decorated materials of rooms with the presence of common meeting room noises are used for recording, which makes the reverberation and noises as the dominating factors affecting the speech quality.
- Multiple microphone arrays from 3 different types are allocated in each recording environment, which can boost the both single microphone array based and distributed microphone array based far-field speech enhancement research.
- The final ranking of the challenge will be decided by the subjective evaluation. The subjective evaluation will be performed using Absolute Category Ratings (ACR) to estimate a Mean Opinion Score (MOS) through Tencent Media Subjective Evaluation platform.

ConferencingSpeech 2021 challenge is open to all. There is no cost to participate in the challenge. The provided datasets, scripts, and baseline system will be available free of charge. The information of challenge registration can be found on the challenge website[1]

2. Task Description

The challenge will have the following two tasks. Each registered team could participate in any one or both tasks. Participating in both tasks are encouraged. The topology of microphone arrays will be given in this plan. The details of microphone arrays can be referred to Section 3.

- **Task1: Multi-channel speech enhancement with single microphone array**. This task is focusing on processing the speech from single linear microphone array with non-uniform distributed microphones and considering practical application with real time requirement.

[1] https://tea-lab.qq.com/conferencingspeech-2021

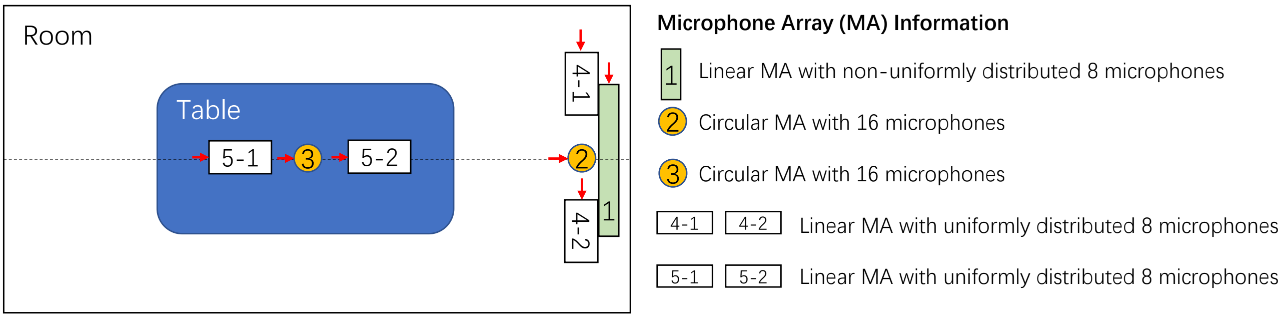
Room
Table
5-1
3
5-2
2
4-1
4-2
1
Microphone Array (MA) Information
1 Linear MA with non-uniformly distributed 8 microphones
2 Circular MA with 16 microphones
3 Circular MA with 16 microphones
4-1 4-2 Linear MA with uniformly distributed 8 microphones
5-1 5-2 Linear MA with uniformly distributed 8 microphones

Figure 1: *The setup of microphone arrays in the meeting rooms*

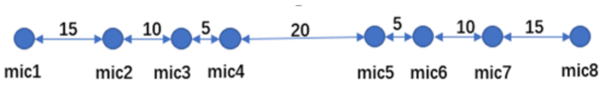

Figure 2: *The information of MA No.1.The unit in this figure is centimeter.*

No future information (zero look-ahead) could be used in Task 1. Frame length should be less than or equal to 40ms. The real time factor of algorithms must less and equal than one on the single thread of an Intel Core i5 machine clocked at 2.4GHz or equivalent processors. The real time factor F_{rt} is formulated as follows:

$$F_{rt} = \frac{T_p}{T_t} \tag{1}$$

where T_p is the processing time of given test clip; T_t is the time of test clip.

- **Task2: Multi-channel speech enhancement with multiple distributed microphone arrays**. This task is focusing on processing the speech from multiple distributed microphone arrays. There are five microphone arrays from three different geometric typologies. All speech signals from these five microphone arrays are synchronized. This task is non-real-time track and does not have any constraints so that participants could explore any algorithms to obtain high speech quality.

3. Data Description

3.1. ConferencingSpeech 2021 Challenge Database

Aiming at the real video conferencing room scenario, the ConferencingSpeech 2021 Challenge Database is recorded with real speakers and all recording facilities are located by following the real setup of video conferencing room. Specifically, 12 rooms of different sizes and decorated materials are used for recording to simulate most of video conferencing room scenarios. The decorated materials of rooms could be categorized into four type: without glass wall, with 1 glass walls, with 2 glass walls, and with 3 glass walls. The database contains two languages: English and Chinese. The sampling rate is 16kHz.

Multiple microphone arrays from three different types (including geometry and manufactory) are distributed in the recording rooms to explore the impact of layout, acoustic and circuit design of microphone arrays on the speech quality. Figure 1 shows the recording setup of ConferencingSpeech 2021 Challenge database. The recording devices include 5 microphone arrays from 3 different geometric topology. The red arrow in Figure 1 points to the first channel of microphone arrays. The microphone arrays in each meeting room follow the allocation of Figure 1, but the distances among microphone arrays (MAs) vary according to the sizes of meeting room. All recordings from these 5 MAs are synchronized. The information of MAs in Figure 1 are concluded as follows:

- No.1 is a linear MA with non-uniformly distributed 8 microphones. The interval among microphones could be referred to Figure 2.
- No.2 and No.3 are circular MA with 16 microphones. The radius of circular MA is 5cm.
- No.4 and No.5 are linear MA. Each MA is composed of two small linear MA with uniformly distributed 8 microphones. The interval among microphones in the small linear MA is 1.1cm.

The recording conditions of this database could be categorized into the following two parts:

- Semi-real recording: the data is recorded in the real meeting room scenario, but is not in the real meeting. The speech and noise data are separately recorded at all 12 quiet meeting rooms. Both are collected from playback and real speakers. Then these speech and noise data are used for simulation. It contains two language: English and Chinese. The benefit of semi-real recording is that more meeting scenarios could be considered.
- Real recording: the data is recorded during the real meeting in the real meeting room. All recorded speeches and noises are from real speakers. The recording language is in Chinese.

3.2. Training Set

To focus on the development of algorithms, the challenge requires the *close* training condition. In other words, except Room Impulse Responses (RIR), only provided list of open source clean speech datasets and noise dataset could be used for training.

3.2.1. Clean Speech

Clean training speech are collected from four open source speech databases: AISHELL-1 [1], AISHELL-3 [2], VCTK [3], Librispeech(train-clean-360)[4]. The speech utterances with SNR larger than 15dB are selected for training. The total duration of clean training speech is around 550 hours. The list of selected clean open source data will be released to participants.

3.2.2. *Noise Set*

The Noise set is composed of two parts. Part I is selected from MUSAN [5] and Audioset[2]. The total duration is around 120 hours. Part II is the real meeting room noises recorded by high fidelity devices. The total number of clips is 98. The list of Part I and Part II noise set will be released to participants.

3.2.3. *Room Impulse Responses (RIR)*

Image method is performed to simulate RIR. The room size ranges from $3 * 3 * 3\ m^3$ to $8 * 8 * 3\ m^3$. More than 2500 rooms are covered. The microphone array is randomly placed in the room with height ranges from 1.0 to 1.5 m. The sound source, including speech and noise, comes from any position in the room with height ranges from 1.2 to 1.9 m. The angle between two sources are wider than 20°. The distance between sound source and microphone array ranges from 0.5 to 5.0 m. There are totally around 20,000 RIRs.

3.3. Development test set

Development test set could be categorized into three parts: Simulation clips, Semi-real recordings, and Real recordings. Semi-real and real recordings are selected from 3 rooms' recordings of ConferencingSpeech 2021 Challenge Database.

3.3.1. *Simulation clips*

The simulation set is provided for participants to develop the systems and estimate the objective scores, which contains two sets: (1) single MA set and (2) multiple MA set.

For single MA set, we simulate 1,588 clips for three types of MA, respectively. These three types of MA are Circular MA with 8 uniformly distributed microphones (radius = $5cm$), linear MA with 8 uniformly distributed microphones (interval = $1.1cm$) and linear MA with 8 non-uniformly distributed microphones. The details can be referred to Section 3.1. Similar as single MA set, multiple MA set also consists of simulation clips from these three MAs. The only difference is that these three MAs are assumed in the same room during simulation.

1,624 clean speech selected from AISHELL-1, AISHELL-3, and VCTK and 800 noise clips selected from MUSAN are used for the simulation of both two sets. The simulated SNR ranges from 0dB to 30dB and the duration of clips is 6 seconds.

3.3.2. *Semi-real recordings*

As mentioned in Section 3.2, the speech sources could be divided into playback and real speaker. The Semi-real recordings consists of 2.35 hours of playback English speech segments and 2.31 hours of real speaker's Chinese speech segments. Each audio clip contains multiple channel information. All five MAs' recordings are provided for participants to develop their systems.

3.3.3. *Real recordings*

More than 200 real recording clips are provided, which are from 12 real speakers and their ages range from 18 to 50 years old. Similar as Semi-real recordings, each audio clip contains multiple channel information and all five MAs' recordings are provided.

3.4. Evaluation test set

The evaluation set contains Task 1 and Task 2 test sets. Audio clips from other 9 rooms' recordings of ConferencingSpeech 2021 Challenge Database are selected. The test set of Task 1 is focusing on the single microphone array. No. 1 MA in Figure 1 is selected for Task 1. Task 2 test set is focusing on the multiple distributed microphone arrays. The clips from five MAs in Figure 1 are provided for Task 2.

Different from development test set, the evaluation test set only contains the semi-real and real recordings. Specifically, evaluation set for each task is composed of three parts: semi-real recordings from playback, semi-real recordings from real speakers, and real meeting recordings.

4. Evaluation Methodology

The final ranking of this challenge will be decided by the subjective evaluation and computational complexity. The subjective evaluation will be performed using Absolute Category Ratings (ACR) to estimate a Mean Opinion Score (MOS) through Tencent Media Subjective Evaluation platform which is similar as the online subjective evaluation framework in DNS challenge [6]. The submission with the highest average MOS and satisfying the challenge requirements will be the winner.

5. Challenge Rules and Requirements

5.1. Registration

The registration link is available in the website of challenge[3]. We kindly request participants to use institutional email for registration. Otherwise, the registration may be invalid. Each participant must sign the challenge agreement and send it to the organizers. Once the registration is confirmed, participants will receive the confirmation letter of registration, and your email address will be recorded for the access of downloading the challenge database.

Please note that any deliberate attempts to bypass the submission limit will lead to automatic disqualification. For example, the same participants create multiple teams and submit multiple results. In case of any issue, the final interpretation right belongs to the organizing committee.

5.2. Submission

5.2.1. *Submission of enhanced evaluation test set*

Participants are required to submit the enhanced evaluation test set for each participating task. The filename should keep the same as the provided evaluation test set. The details of rule and way of submission will be notified by organizers in March 2021.

5.2.2. *Submission of System description*

Each registered team is required to submit a technical system description report. Please submit this report using the Interspeech 2021 paper template. Reports must be written in English. The system description does not need to repeat the content of the evaluation plan, such as the introduction of database, evaluation metric, etc. The system description must include the following items:

- a complete description of the system components, including the acoustic feature parameters, algorithm modules along with their configurations, etc.

[2] https://research.google.com/audioset/

[3] https://tea-lab.qq.com/conferencingspeech-2021

- a complete data description for training.
- the objective scores of simulation clips in development test set, including PESQ, STOI, E-STOI, and SI-SDR.
- a report of the model size, real time factors (single threaded CPU) as well as the amount of memory used to process a single clip. Both Task 1 and Task 2 are needed to report.

5.2.3. *Paper submission*

Each participating team must submit a paper to Interspeech 2021 special session - Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing (ConferencingSpeech 2021). Only the teams submitting the paper to ConferencingSpeech special session will be considered for the final ranking of challenge. Please submit your paper by 26 Mar 2021 to the Interspeech 2021 paper submission system and choose this special session (ConferencingSpeech 2021). The papers will undergo the standard peer-review process of Interspeech 2021.

5.3. Other Important Rules

- Participants must abide by the requirements in Section 2 for each task.
- Participants can only use the provided training data and noise sets for training.
- No restrictions on the algorithm. Participants could develop any algorithm for the tasks.
- Participants must send the results achieved by their developed models to the organizers. The details of submissions can be referred to Section 5.2.
- The organizers will use the submitted enhanced clips with no alteration to perform subjective evaluation and select the winners based on the results.
- Participants are forbidden to use the evaluation test set to fine-tune or retrain their models. They should not submit results using other speech enhancement methods which are not submitted to ConferencingSpeech 2021. Failing to adhere to these rules will lead to disqualification from the challenge.
- Winners will be selected based on the subjective evaluation.

6. Timeline

The schedule of ConferencingSpeech 2021 challenge is planed as follows:

- 22 Jan 2021: Release of the list of training data, noise dataset, development test set, and scripts for simulation
- 05 Feb 2021: Release of baseline system
- 07 Mar 2021: Deadline of challenge registration
- 08 Mar 2021: Release of evaluation test set
- 13 Mar 2021: Deadline of submitting the results for subjective evaluation on the evaluation test set
- 23 Mar 2021: Notification of the results of participants
- 26 Mar 2021: Interspeech paper submission deadline
- 02 Apr 2021: Interspeech paper update deadline
- 02 Jun 2021: Paper acceptance/rejection notification
- 05 Jun 2021: Notification of the winners
- 31 Aug 2021: Opening of Interspeech 2021

7. References

[1] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *Oriental COCOSDA 2017*, 2017.

[2] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, “AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines,” 2020. [Online]. Available: https://arxiv.org/abs/2010.11567

[3] V. Christophe, Y. Junichi, and M. Kirsten, “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” *The Centre for Speech Technology Research (CSTR)*, 2016.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[5] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” 2015, arXiv:1510.08484v1.

[6] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, and S. M. et al, “The interspeech 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results,” in *Proc. INTERSPEECH 2020*, 2020.