AISHELL-4

A Free Mandarin Multi-channel Meeting Speech Corpus

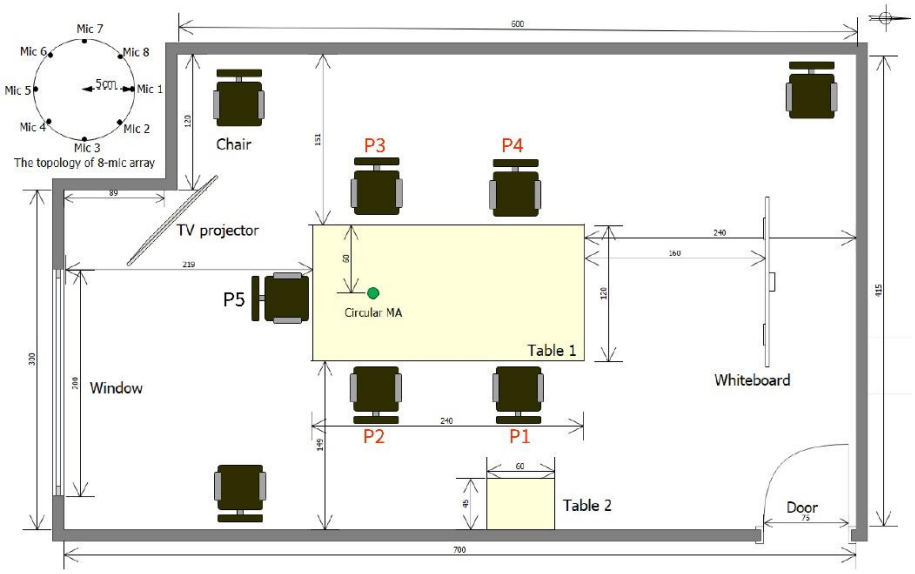
1 Dataset Information

AISHELL-4 is part of the AISHELL-ASR0055 corpus. The recording was put in a conference environment, using 7 different devices in parallel: high fidelity microphone (44.1kHz, 16-bit); circular microphone array (16kHz, 16-bit); linear microphone array (16kHz, 16-bit); headset microphone (16kHz, 16-bit); Android-system Pad (16kHz, 16-bit); Android-system mobile phone (16kHz, 16-bit), iOS-system mobile phone (16kHz, 16-bit). AISHELL-4 chooses 8-channel audio data record by the circular microphone array.

AISHELL-4 contains 120 hours of speech data, divided into 107.50 hours of training and 12.72 hours evaluation set. The training and evaluation set includes 191 and 20 sessions, respectively. Each session consists of a 30-minute discussion by a group of participants. The total number of participants in training and evaluation sets are 36 and 25, with balanced gender coverage.

The dataset is collected in 10 conference venues. The conference venues are divided into three types: small, medium, and large room, whose size range from $7 * 3 * 3$ to $15 * 7 * 3$ m^3 . The type of wall material of the conference venues covers cement, glass, etc. Other furnishings in conference venues include sofa, TV, blackboard, fan, air conditioner, plants, etc. During recording, the participants of the conference sit around the microphone array which is placed on the table in the middle of the room and conduct a natural conversation. The microphone-speaker distance ranges from 0.6 to 6.0 m. All participants are native Chinese speakers speaking Mandarin without strong accents. During the conference, various kinds of indoor noise including but not limited to clicking, keyboard, door opening/closing, fan, bubble noise, etc., are made by participants naturally. For the training set, the participants are required to remain in the same position during recording, while for the evaluation set, the participants may move naturally within a small range.

There is no room overlap and only one speaker overlap between training and evaluation set. An example of the recording venue from the training set, including the topology of microphone array, is shown in Fig.1.



Meeting Room Schema [MR001]

Figure 1: An example of recording venue of training set and the topology of microphone array.

The number of participants within one conference session ranges from 4 to 8. To ensure the coverage of different overlap ratios, we select various meeting topics during recording, including medical treatment, education, business, organization management, industrial production and other daily routine meetings. The average speech overlap ratios of training and evaluation sets are 19.04% and 9.31%, respectively. More details of AISHELL-4 is shown in Table 1. A detailed session-level overlap ratio distribution of training and evaluation sets is shown in Table 2.

Table 1: *Details of data to release.*

	Training	Evaluation
Duration (h)	107.50	12.72
#Session	191	20
#Room	5	5
#Participant	36	25
#Male	16	11
#Female	20	14
Overlap Ratio (Avg.)	19.04%	9.31%

Table 2: *Session-level overlap ratio distribution. 0%-10%, 10%-20%, 20%-30%, 30%-40% and 40%-100% indicate the range of overlap ratio. The numbers followed indicate the number of sessions with corresponding overlap ratio in training and evaluation sets, respectively.*

Overlap Ratio	Training	Evaluation
0%-10%	41	12
10%-20%	76	6
20%-30%	44	2
30%-40%	20	0
40%-100%	10	0

2 Speech Content Annotation

We also record the near-field signal using headset microphones for each participant. To obtain the transcription, we first align the signal recorded by headset microphone and the first channel of microphone array and then select the signal with higher quality for manual labeling. Before labeling the scripts, we apply automatic speech recognition on the recorded data to assist transcribers to for accurate labeling results. Then inspectors will double-check the labeling results of each session from transcribers and decide whether meet the acceptance standard. Praat is used for further calibration to check the accuracy of speaker distribution and to avoid the miss cutting of speech segments. We also pay special attention to accurate punctuation labeling. Each session is labeled by three professional annotators on average as secondary inspection to improve the labeling quality.

All scripts of the speech data are prepared in textgrid format for each session, which contains the information of the session duration, speaker information (number of speaker, speaker-id, gender, etc.), the total number of segments of each speaker, the timestamp and transcription of each segment, etc. The non-speech events are transcribed as well, such as pauses, laughing, coughing, breathing, etc. The overlapping and non-overlapping segments are also identified.

Data annotator listens to the audio content to write, in order to make the text and audio content consistent with pronunciation. General guidelines are shown as below:

1) Transliteration and heard speech content must be entirely consistent, not more, fewer, or wrongly written a word.

2)To transfer into digital form Chinese characters, such as "一二三", instead of "123". Pay attention to distinguish between "一" and "幺", "二" and "两".

3) Audio in English pronunciation should be written in the corresponding Chinese characters or English. Specific is divided into the following situations:

All the letters or words contained in the URL are capitalized. For example, the pronunciation content for the "www.abc.com" should transfer to "三 W 点 A B C 点 com"

The English pronunciation contains all lowercase words, transliteration.

English pronounce as words should transcript as lowercase.

English pronounce as spelling should transcript as uppercase.

For some proper nouns or some English abbreviations, all transcript as uppercase with a space mark, such as C E O, C C T V, etc.

4) The integrity of the content should be consistent with the actual pronunciation and shall not be deleted.

5) Extra tags information.

Num.	Tag	Usage
1	<sil>	Pause. These are uttered when the speaker is hesitating, stalling for time, collecting his/her thoughts.
2	<->	Used for a partial word
3	<$>	mark an occurrence of laughter
4	<_>	Sentence truncation
5	<%>	cough
6	<#>	Invalid Speech
7	&	Overlap tag

3 Corpus Catalog

3.1 Directory Structure

Directory tree		
Corpus Catalog		
|AISHELL-4 Data-Specification.pdf		**Corpus Information**
└─DOC		**DOC File**
| ├─all_wav_list.txt		**Audio list**
| ├─spk_info.xlsx		**Speaker Information**
└─L		**Venue Type**
| ├─L_R003		**Venue ID**
| ├─wav		**Audio Data File**
| | | 20201129_L_R003S06C01.wav		**Audio**
| ├─TextGrid		**Transcript File**
| | | 20201129_L_R003S06C01.TextGrid		**TextGrid**
| ├─L_R003.jpg		**Venue Schema**

3.2 Naming Rule

3.2.1 Directory Naming Rules

/< VENUE_TYPE>/< VENUE_ID >/<WAV>/<AUDIO_ID>

e.g. L/L_R003/wav/20201129_L_R003S06C01.wav

Directory	Content	Note
VENUE_TYPE file	Small / Medium / Large	Meeting venue type
VENUE_ID file	L_R003	Venue ID
AUDIO file	wav/20201129_L_R003S06C01.wav	WAV file

Chart 3-2-1

3.2.2 File Naming Rules

<TIME_ID>_<VENUE_TYPE>_< VENUE_ID >< MEETING_ID >_<DEVICE_ID>.wav

e.g. 20201129_L_R003S06C01.wav

文件	内容	备注
TIME_ID	20201129	Recording Time
VENUE_TYPE	_S/M/L_	Venue type
VENUE_ID	R001~500	Venue ID
MEETING_NUM	S01~08	Meeting ID
DEVICE_ID	C01 & 02	Device ID

Chart 3-2-2

4 Copyright