Open Source Mandarin Speech Corpus

[AISHELL-ASR0009-OS1]

Training and Test Data

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail： bd@aishelldata.com

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail： bd@aishelldata.com

1 Product Instruction

This Open Source Mandarin Speech Corpus, AISHELL-ASR0009-OS1, is 173 hours long. It is a part of AISHELL-ASR0009, of which utterance contains 11 domains, including smart home, autonomous driving, and industrial production.

The whole recording was put in quiet indoor environment, using 3 different devices at the same time: high fidelity microphone (44.1kHz, 16-bit); Android-system mobile phone (16kHz, 16-bit), iOS-system mobile phone (16kHz, 16-bit).

400 speakers from different accent areas in China were invited to participate in the recording. The manual transcription accuracy rate is above 95%, through professional speech annotation and strict quality inspection. The corpus is divided into training, development and testing sets.

This database is free for academic research, not in the commerce, if without permission.

2 Recording Text

2.1 Text Pool

2.1.1 Text Pool Content

Considering the application of speech recognition in smart home, autonomous driving, industrial production, and other fields, the corpus is selected from 11 domains. (Chart 2-1)

Serial 6 to serial 10 are included in AISHELL-ASR0009-OS1.

Serial Number	Domain
1	Smart Home Voice Control
2	POI (Geographic Information)
3	Music (Voice Control)
4	Digital String (Voice Control)
5	TV Play and Film Names
6	Finance
7	Science and Technology
8	Sports
9	Entertainments
10	News
11	English Spelling

Char 2-1 Text Pool Content

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

2.1.2 Text Pool Processing

- Off-sensitivity. Delete politically sensitivity, personal privacy, and pornographic violence such kind of content.
- Delete < , > , [,] , ～ , / , \ , =, such kind of mark.
- Delete content in languages other than Chinese and English.
- Unified format.

2.2 Structure Design of Recording Text

In view of speech coverage and phoneme balance, the recording text of AISHELL-ASR0009 is designed by the allocation of 500 sentences, extracted from the text pool, and structured as follow.

AISHELL-ASR0009-OS1 contains 5 domains, from serial 6 to 10. (Chart 2-2)

Serial Number	Domain	Allocation /#Sentence
1	Smart Home Voice Control	5
2	POI (Geographic Information)	30
3	Music (Voice Control)	46
4	Digital String (Voice Control)	29
5	TV Play and Film Names	10
6	Finance	132
7	Science and Technology	85
8	Sports	66
9	Entertainments	27
10	News	66
11	English Spelling	4
Total	**11 domains**	**500 sentences**

Chart 2-2

3 Speaker Information

3.1 Speaker Information Registration

Speaker information is comprised of Task ID, Age, Gender, Accent Area and Birth Place. (Chart 3-1)

Task ID	Age	Gender	Birth Place	Accent Area

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

C0002	18	M	Hebei	North

Chart 3-1

Task ID: Each speaker can only fulfill 1 Task, while each Task corresponding to 1 recording text.

Gender: M defined as male, while F defined as female.

Birth Place: Duplicate from every speaker's Citizen ID Card.

Accent Area: Divided into North, South and other areas according to the region where speakers belong to the native language.

3.2 Speaker Demographic Information

3.2.1 Gender Balance

This database consists of 186 male speakers and 214 female speakers.(Char 3-2-1)

Gender	**#Male**	**#Female**	**Total**
Percentage	47%	53%	**100%**

Chart 3-2-1

3.2.2 Age Distribution

A (16-25 years old) 316 people; B (26-40 years old) 71 people; C (41 years old or above) 13 people. (Chart 3-2-2)

	Age Range	**#Speaker**	**Percentage**	**#Male**	**#Female**
A	**16-25 yrs**	316	79%	140	176
B	**26-40 yrs**	71	18%	36	35
C	**> 41 yrs**	13	3%	10	3
Total		**400**	**100%**	**186**	**214**

Chart 3-2-2

3.2.3 Dialectal Regions

331 people in the North, 60 in the South, 9 in other areas. (Chart 3-2-3)

Accent Area	**#Speaker**	**%Speaker**
North	331	83%
South	60	15%
Other Areas	9	2%

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

合计	**400**	**100%**

Chart 3-2-3

4 Recording Processing

4.1 Recording Environment

Quiet indoors, not including other people voice, and other noises without reverberation. The speaker reads recording text at regular speed.

4.2 Recording Equipment

Recording equipment includes high fidelity microphone (Audio Technica 2035) and recorder (Roland-44), iOS-system mobile phone, and Android-systems mobile phone.

4.3 Recording Method

The speaker is 20cm from the high-fidelity microphone and reads the recording text with the normal volume of normal speed. Android-system and iOS-system mobile phones respectively with microphone interval layout. (Chart 4-3)

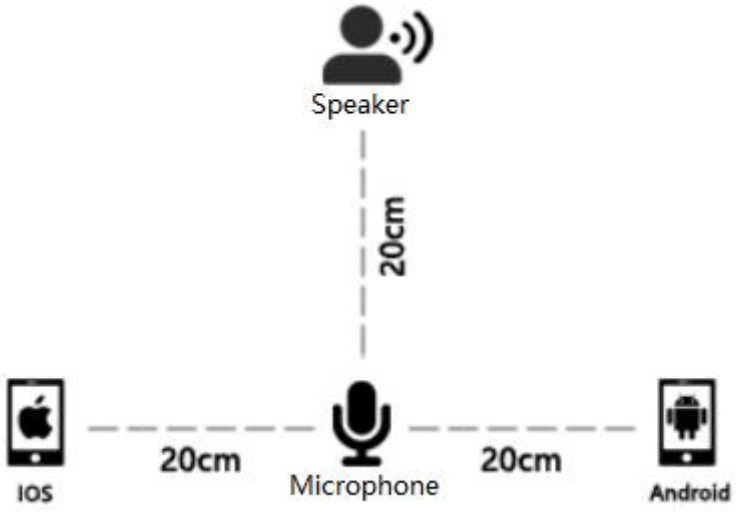

Chart 4-3

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

5 Speech Content Annotation

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 completely 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.

6 Corpus Catalog

6.1 Directory Structure

Directory tree		
Corpus Catalog		
| AISHELL-ASR0009-OS1.docx		**(Text instruction)**
└─DOC		**(Description file)**
||├─all_wav_list.txt		**(Audio list)**
||├─content.txt		**(Transcripted content)**
||├─spkrinfo.txt		**(Speaker information)**
|		
└─SPEECHDATA		
|├─C0001		
|| ||─MIC		**(Mic Audio File)**
|| ||	**BAC0009C0001W0001.wav**	**(Audio)**
|| ||	**BAC0009C0001W0001.txt**	**(Text)**
|| ||─IOS		**(IOS Audio File)**
|| ||─ANDROID		**(ANDROID Audio File)**

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail： bd@aishelldata.com

6.2 Naming Rule

6.2.1 Directory Naming Rules (Chart 6-2-1)

/<CORPUS>/<USAGE>/<FILE_ID>/<AUDIO_ID>/<SPEECH_ID>

e.g. ChineseMandarinSpeechRecognitionCorpus/SPEECHDATA/C0001/MIC/BAC0009C0001W0001.wav

Directory	Content	Note
CORPUS	Chinese Mandarin Speech Recognition Corpus	Database Name
USAGE	SPEECHDATA	Folder Name
FILE_ID	C0001	Speaker ID Folder
AUDIO_ID	MIC	MIC audio file
SENTENCE_ID	BAC0009C0001W0001.txt	TXT documents
SPEECH_ID	BAC0009C0001W0001.wav	WAV documents

Chart 6-2-1

6.2.2 File Naming Rules (Chart 6-2-2)

<CORPUS_ID><SPEAKER_IC><WAV_NUM>

e.g. BAC0009C0001W0001.wav

File	Content	Note
CORPUS_ID	BAC009	Database Name
SPEAKER_NUM	C0001	Speaker ID
WAV_NUM	W0001	WAV number

Chart 6-2-2

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com