AISHELL-4
多通道中文会议开源语音数据库

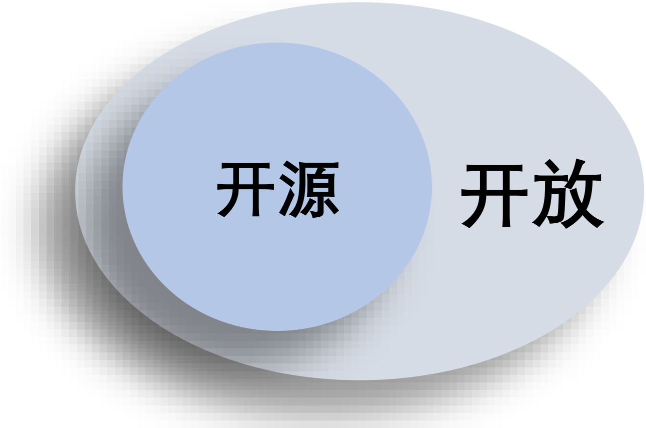
卜辉 CEO
北京希尔贝壳科技有限公司

Contents

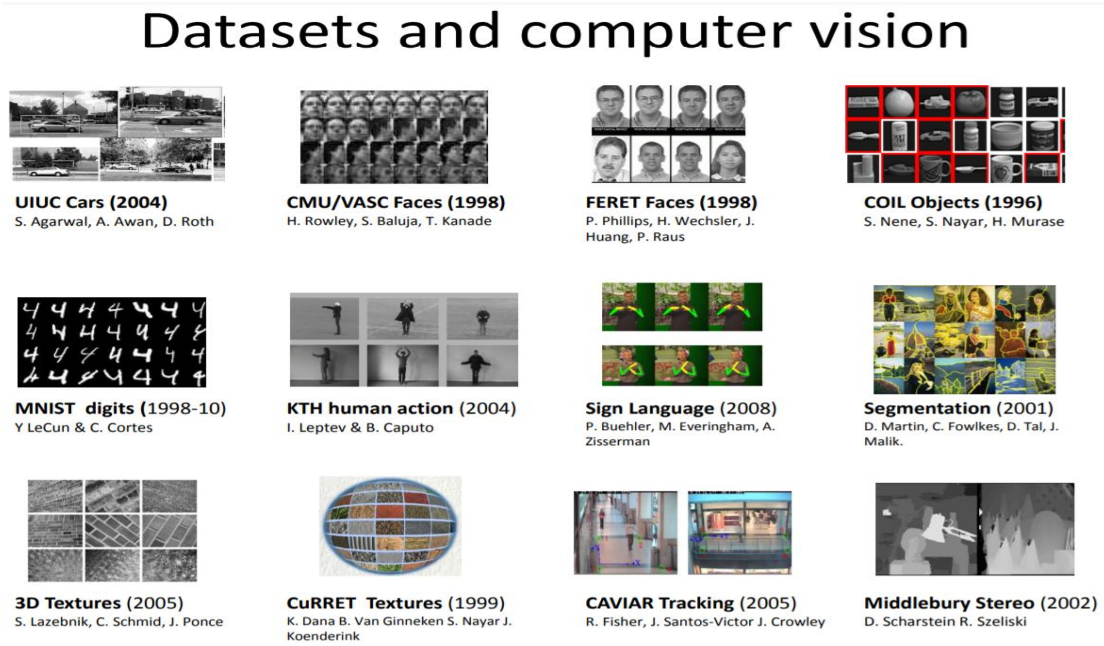
公共数据集开放提供燃料

ImageNet 中含有超过 1500 万由人手工注释的图片，包含超过 2.2 万个类别，从2010年ILSVRC第一届比赛开始，随着参赛队伍的增多，算法逐渐逼近人类识别水平

- 开放：开放不一定开源；
- 开源：开放的途径，指算法、工具等开源；



发布时间：2009 ～ 2010　　1400万的图像



OpenSLR

Home Resources

Resources

Resource	Name	Category	Summary
SLR1	Yesno	Speech	Sixty recordings of one individual saying yes or no in Hebrew; each recording is eight words long.
SLR2	OpenFST	Software	A mirror of the OpenFst toolkit
SLR3	sph2pipe	Software	A mirror of the sph2pipe software
SLR4	sctk	Software	A mirror of the sctk scoring software
SLR5	MSU Switchboard transcipts	Text	A mirror of the Mississippi State transcripts and lexicon for Switchboard.
SLR6	Vystadial	Speech	English and Czech data, mirrored from the Vystadial project
SLR7	TED-LIUM	Speech	English speech recognition training corpus from TED talks, created by Laboratoire d'Informatique de l'Université du Maine (LIUM) (mirrored here)
SLR8	Sprakbanken	Text	Danish pronunciation dictionary generated using eSpeak
SLR9	The AMI pack	Text	Some auxiliary non-speech data used to build AMI systems with Kaldi
SLR10	SRE Data	Misc	Various files from SRE data that NIST used to host online
SLR11	LibriSpeech language models, vocabulary and G2P models	Text	Language modelling resources, for use with the LibriSpeech ASR corpus
SLR12	LibriSpeech ASR corpus	Speech	Large-scale (1000 hours) corpus of read English speech
SLR13	RWCP Sound Scene Database	Speech + Software	A database of recordings of real-world sounds and measured room impulse responses
SLR14	BEEP Dictionary	Text	Phonemic transcriptions of over 250,000 English words. (British English pronunciations)
SLR15	SRE Speaker List	Misc	A list linking speakers across NIST SRE corpra
SLR16	The AMI Corpus	Speech	Acoustic speech data and meta-data from The AMI corpus.
SLR17	MUSAN	Audio	A corpus of music, speech, and noise
SLR18	THCHS-30	Speech	A Free Chinese Speech Corpus Released by CSLT@Tsinghua University
SLR19	TED-LIUMv2	Audio	TED-LIUM corpus release 2, English speech recognition training corpus from TED talks, created by Laboratoire d'Informatique de l'Université du Maine (LIUM) (mirrored here)
SLR20	Aachen Impulse Response Database	Audio	Aachen Impulse Response database (AIR); a database of room impulse responses (mirrored here)
SLR21	Spanish Word list	Text	A list of words in Spanish with frequency derived from a large corpus (Spanish Gigaword).
SLR22	THUYG-20	Speech	A free Uyghur speech database Released by CSLT@Tsinghua University & Xinjiang University
SLR23	NIST LRE 2007 Key	Misc	A file containing metadata for the utterances in the LRE 2007 evaluation
SLR24	Iban	Speech	Iban language text and speech corpora for ASR
SLR25	ALFFA (African Languages in the Field: speech Fundamentals and Automation)	Speech	Amharic, Swahili and Wolof data, mirrored from the ALFFA git repository
SLR26	Simulated Room Impulse Response Database	Audio	A database of simulated room impulse responses
SLR27	Cantab-TEDLIUM Release 1.1 (February 2015)	Text	Cantab Research Language models for the TEDLIUM database
SLR28	Room Impulse Response and Noise Database	Audio	A database of simulated and real room impulse responses, isotropic and point-source noises. The audio files in this data are all in 16k sampling rate and 16-bit precision.
SLR29	Sprakbanken_Swe	Text	Swedish pronunciation dictionary
SLR30	Sinhala TTS	Speech	Sinhalese multi-speaker TTS corpora
SLR31	Mini LibriSpeech ASR corpus	Speech	Subset of LibriSpeech corpus for purpose of regression testing
SLR32	High quality TTS data for four South African languages (af, st, tn, xh)	Speech	Multi-speaker TTS data for four South African languages, Afrikaans, Sesotho, Setswana and isiXhosa.
SLR33	Aishell	Speech	Mandarin data, provided by Beijing Shell Shell Technology Co.,Ltd
SLR34	Santiago Spanish Lexicon	Text	A pronouncing dictionary for the Spanish language.
SLR35	Large Javanese ASR training data set	Speech	Javanese ASR training data set containing ~185K utterances.
SLR36	Large Sundanese ASR training data set	Speech	Sundanese ASR training data set containing ~220K utterances.
SLR37	High quality TTS data for Bengali languages	Speech	Multi-speaker TTS data for Bangladesh Bengali (bn-BD) and Indian Bengali (bn-IN).
SLR38	Free ST Chinese Mandarin Corpus	Speech	A free Chinese Mandarin corpus by Surfingtech (www.surfing.ai), containing utterances from 855 speakers, 102600 utterances;
SLR39	Heroico	Speech	Spanish data, mirrored from the LDC
SLR40	Zeroth-Korean	Speech Corpus for Automatic Speech Recognition	Korean Open-source Speech Corpus for Speech Recognition by Zeroth Project (https://github.com/goodatlas/zeroth)
SLR41	High quality TTS data for Javanese.	Speech	Multi-speaker TTS data for Javanese (jv-ID)
SLR42	High quality TTS data for Khmer.	Speech	Multi-speaker TTS data for Khmer (km-KH)
SLR43	High quality TTS data for Nepali.	Speech	Multi-speaker TTS data for Nepali (ne-NP)
SLR44	High quality TTS data for Sundanese.	Speech	Multi-speaker TTS data for Sundanese (su-ID)
SLR45	Free ST American English Corpus	Speech	A free American English corpus by Surfingtech (www.surfing.ai), containing utterances from 10 speakers, Each speaker has about 350 utterances;
SLR46	Tunisian_MSA	Speech	Tunisian Modern Standard Arabic
SLR47	Primewords Chinese Corpus Set 1	Speech	Chinese Mandarin corpus released by Shanghai Primewords Co. Ltd. (www.primewords.cn), containing 100 hours of speech data.
SLR48	MADCAT Arabic data splits	Other	Unofficial data splits (dev/train/test) for the MADCAT Arabic LDC corpus
SLR49	VoxCeleb Data	Misc	Various files for the VoxCeleb datasets
SLR50	MADCAT Chinese data splits	Other	Unofficial data splits (dev/train/test) for the MADCAT Chinese LDC corpus
SLR51	TED-LIUM Release 3	Speech	TED-LIUM corpus release 3
SLR52	Large Sinhala ASR training data set	Speech	Sinhala ASR training data set containing ~185K utterances.

中文语音开源现状

2017	SLR33	Aishell	Speech	Mandarin data, provided by Beijing Shell Shell Technology Co.,Ltd
2019	SLR85	HI-MIA	Speech	A far-field text-dependent speaker verification database for AISHELL Speaker Verification Challenge 2019
2020	SLR93	AISHELL-3	Speech	Mandarin data, provided by Beijing Shell Shell Technology Co., Ltd.
2021	SLR111	AISHELL-4	Speech	A Free Mandarin Multi-channel Meeting Speech Corpus, provided by Beijing Shell Shell Technology Co.,Ltd

7
1
0

THCHS-30
aidatatang_200h

Primewords Chinese dataset
MAGICDATA Mandarin Dataset

10000H

GigaSpeech
是一个不断发展的、多领域英语语音识别语料库。

10000H

中文领域的大数据、大模型还有多远?

2 AISHELL-4 数据库介绍

AISHELL-1

Kaldi recipe

s5: a speech recognition recipe
v1: a speaker recognition recipe

http://www.openslr.org/33/

178H
400 Speakers

- Sampling Rate: 16KHz
  Sample Format: 16bit
  Environment: Indoor

Abstract:
An open-source Mandarin speech corpus called AISHELL-1 is released. It is by far the largest corpus which is suitable for conducting the speech recognition research and building speech recognition systems for Mandarin. The recording procedure, including audio capturing devices and environments are presented in details. The preparation of the related resources, including transcriptions and lexicon are described. The corpus is released with a Kaldi recipe. Experimental results implies that the quality of audio recordings and transcriptions are promising.

Published in: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)

Date of Conference: 1-3 Nov. 2017

INSPEC Accession Number: 17843434

Date Added to IEEE *Xplore*: 14 June 2018

DOI: 10.1109/ICSDA.2017.8384449

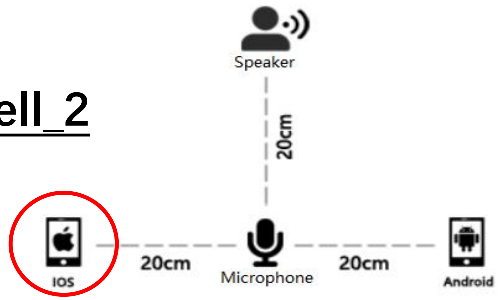
▸**ISBN Information:**

Publisher: IEEE

Kaldi recipe

AISHELL-1

speech recognition recipe
kaldi/egs/aishell2/



http://www.aishelltech.com/aishell_2

1000H
1991 Speakers

- Sampling Rate: 16KHz
  Sample Format: 16bit
  Environment: Indoor

arXiv.org > cs > arXiv:1808.10583

Search...
Help | Ad

Computer Science > Computation and Language

[Submitted on 31 Aug 2018 (v1), last revised 13 Sep 2018 (this version, v2)]

AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale

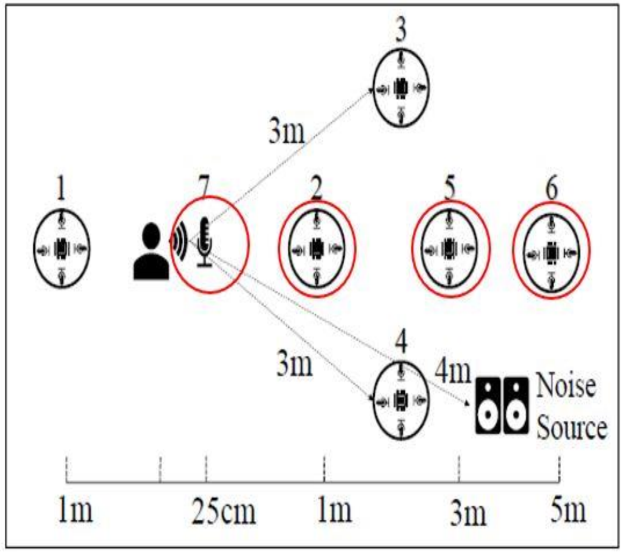
Jiayu Du, Xingyu Na, Xuechen Liu, Hui Bu

AISHELL-1 is by far the largest open-source speech corpus available for Mandarin speech recognition research. It was released with a baseline system containing solid training and testing pipelines for Mandarin ASR. In AISHELL-2, 1000 hours of clean read-speech data from iOS is published, which is free for academic usage. On top of AISHELL-2 corpus, an improved recipe is developed and released, containing key components for industrial applications, such as Chinese word segmentation, flexible vocabulary expension and phone set transformation etc. Pipelines support various state-of-the-art techniques, such as time-delayed neural networks and Lattic-Free MMI objective funciton. In addition, we also release dev and test data from other channels(Android and Mic). For research community, we hope that AISHELL-2 corpus can be a solid resource for topics like transfer learning and robust ASR. For industry, we hope AISHELL-2 recipe can be a helpful reference for building meaningful industrial systems and products.

AISHELL-WakeUp-1

HI-MIA

“你好，米雅” “Hi, Mia” Open Source



AISHELL Speaker Verification Challenge 2019

HI-MIA

http://www.openslr.org/85/

AISHELL-WakeUp-1

http://www.aishelltech.com/wakeup_data

1561H

340 Speakers

Sampling Rate: 44.1KHz & 16KHz

Sample Format: 16bit

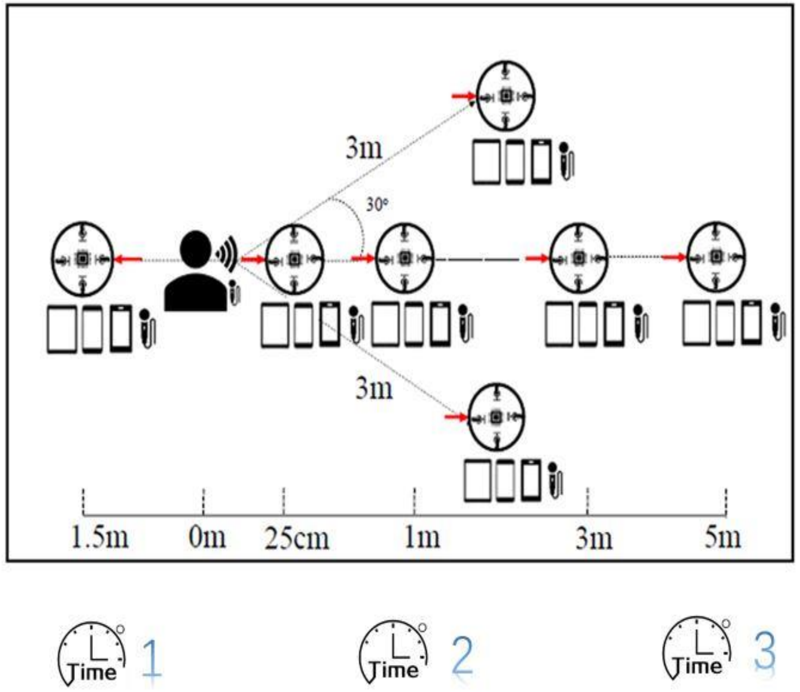
Environment: Indoor

2 AISHELL-4 数据库介绍

AISHELL-DMASH

DMASH

Distributed Microphone Arrays in Smart Home (DMASH) Dataset



500 Speakers

INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI, CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

Open Source

The INTERSPEECH 2020
Far-Field Speaker Verification Challenge

http://www.aishelltech.com/DMASH_Dataset

50000H

Sampling Rate: 44.1KHz & 16KHz

Sample Format: 16bit

Environment: Indoor

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

NUS National University of Singapore

USC University of Southern California

AISHELL

FFSVC 2020

Far-Field Speaker Verification Challenge

Introduction
Data Description
Data Download
Evaluation Plan
Baseline Paper
System Descriptions
Tasks
Task 1
Task 2
Task 3
Important Dates
Registration
Submission
FAQ
Leaderboards

Introduction

Welcome to the Interspeech 2020 Far-Field Speaker Verification Challenge (FFSVC 2020).

Speaker verification is a key technology in speech processing and biometric authentication, which has broad impact on our daily lives, e.g. security, customer service, mobile devices, smart speakers. Recently, speech based human computer interaction has become more and more popular in far-field smart home and smart city applications, e.g. mobile devices, smart speakers, smart TVs, automobiles. Due to the usage of deep learning methods, the performances of speaker verification in telephone channel and close-talking microphone channel have been enhanced dramatically. However, there are still some open research questions that can be further explored for speaker verification in the far-field and complex environments, including but not limited to

- Far-field text-dependent speaker verification for wake up control
- Far-field text-independent speaker verification with complex environments
- Far-field speaker verification with cross-channel enrollment and test
- Far-field speaker verification with single multi-channel microphone array
- Far-field speaker verification with multiple distributed microphone arrays
- Far-field speaker verification with front-end speech enhancement methods
- Far-field speaker verification with end-to-end modeling using data augmentation
- Far-field speaker verification with front-end and back-end joint modeling
- Far-field speaker verification with transfer learning and domain adaptation

AISHELL-3

- Sampling Rate: 44.1KHz
  Sample Format: 16bit
  Environment: Indoor

85H
218 Speakers

Download-Dataset
http://www.aishelltech.com/aishell_3

Download-Paper
https://arxiv.org/abs/2010.11567

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

AISHELL

multi-speaker Text-to-Speech (TTS) systems

AISHELL-3 Corpus: Baseline System Samples

System recipe

Arxiv: 2010.11567
Github Repo: sos1sos2Sixteen/aishell-3-baseline-fc
Dataset Download: www.aishelltech.com/aishell_3
For further questions regarding the dataset: tech@aishelldata.com

Authors

- SHI, Yao (Wuhan University, Duke-Kunshan University)
- BU, Hui (AISHELL)
- XU, Xin (AISHELL)
- ZHANG, Shaoji (AISHELL)
- LI, Ming (Duke-Kunshan University, Wuhan University) ······ ming.li369@dukekunshan.edu

AISHELL-3 V2 **INTERSPEECH 2021**

The Interspeech 2021 Program Committee are pleased to inform you that your paper

Paper ID: 755
Title: AISHELL-3: A Multi-Speaker Mandarin TTS Corpus

has been accepted for presentation at the conference. Please read through the rest of this email carefully

2 AISHELL-4 数据库介绍

AISHELL 4

一个通过麦克风阵列实录的八通道中文普通话会议场景语音数据集

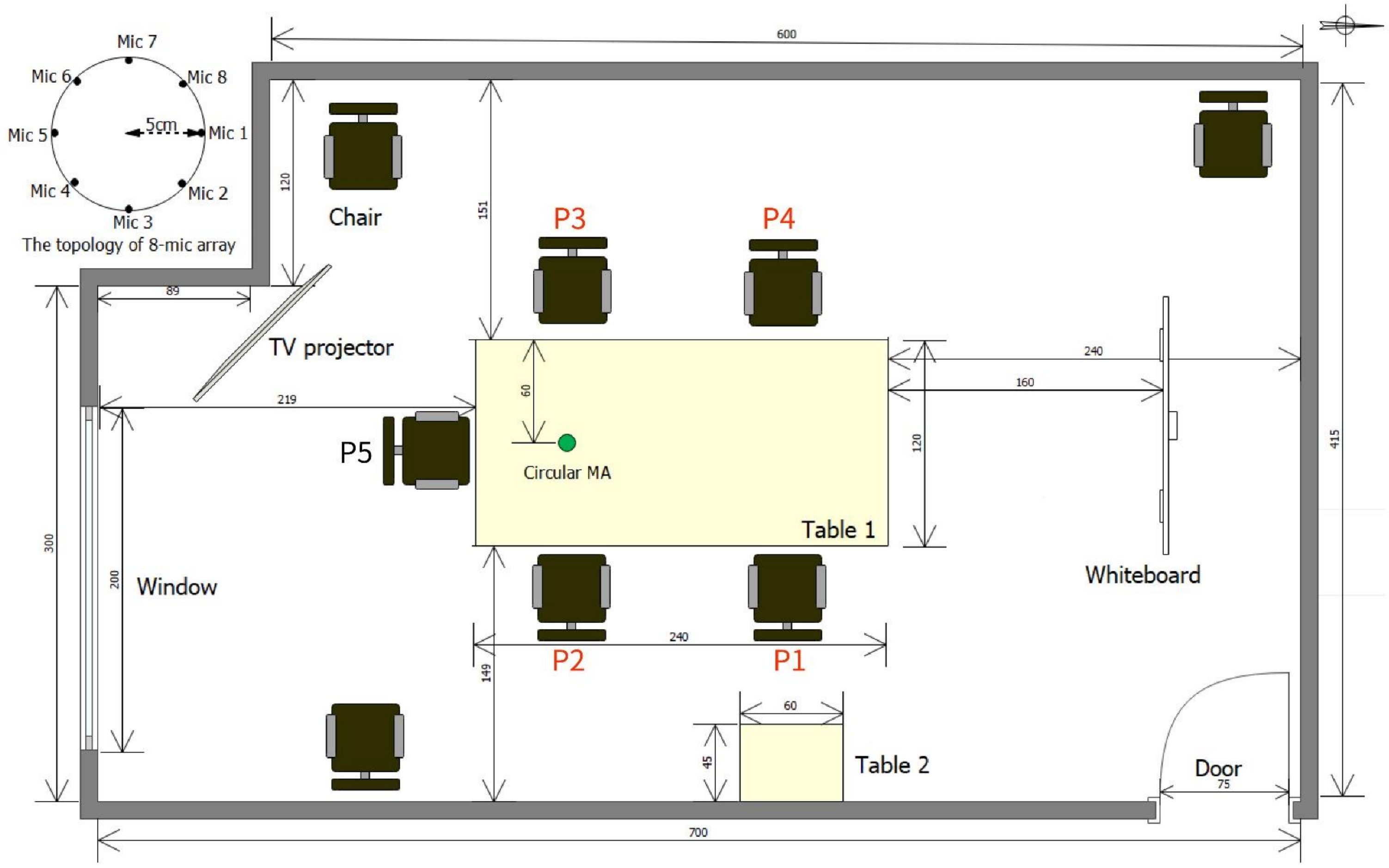
120 小时 | 120 Hours
211 场会议 | 211 Meeting Sessions
10个 会议室 | 10 Meeting Rooms
60 人 | 60 Speakers

Speech front-end processing
Speech recognition
Speaker diarization

开源系统
Open Source

该数据集共包含**211**场会议，每场会议4至8人，数据集共**120**小时左右。该数据集旨在促进实际应用场景下多说话人处理的研究。AISHELL-4数据包括了实际会议场景下各种重要特性，例如停顿、重叠、说话人轮转、噪声等。同时数据集提供了准确的音字转写文本及时间戳信息，方便研究者进行诸如前端处理、语音识别、说话人分割等单独任务，并可以进行联合优化。

2 AISHELL-4 数据库介绍



Theme

财经
房产
家居
教育
科技
时尚
时政
体育
游戏
娱乐

120H 60 Speaker

Meeting sessions

211

Meeting rooms

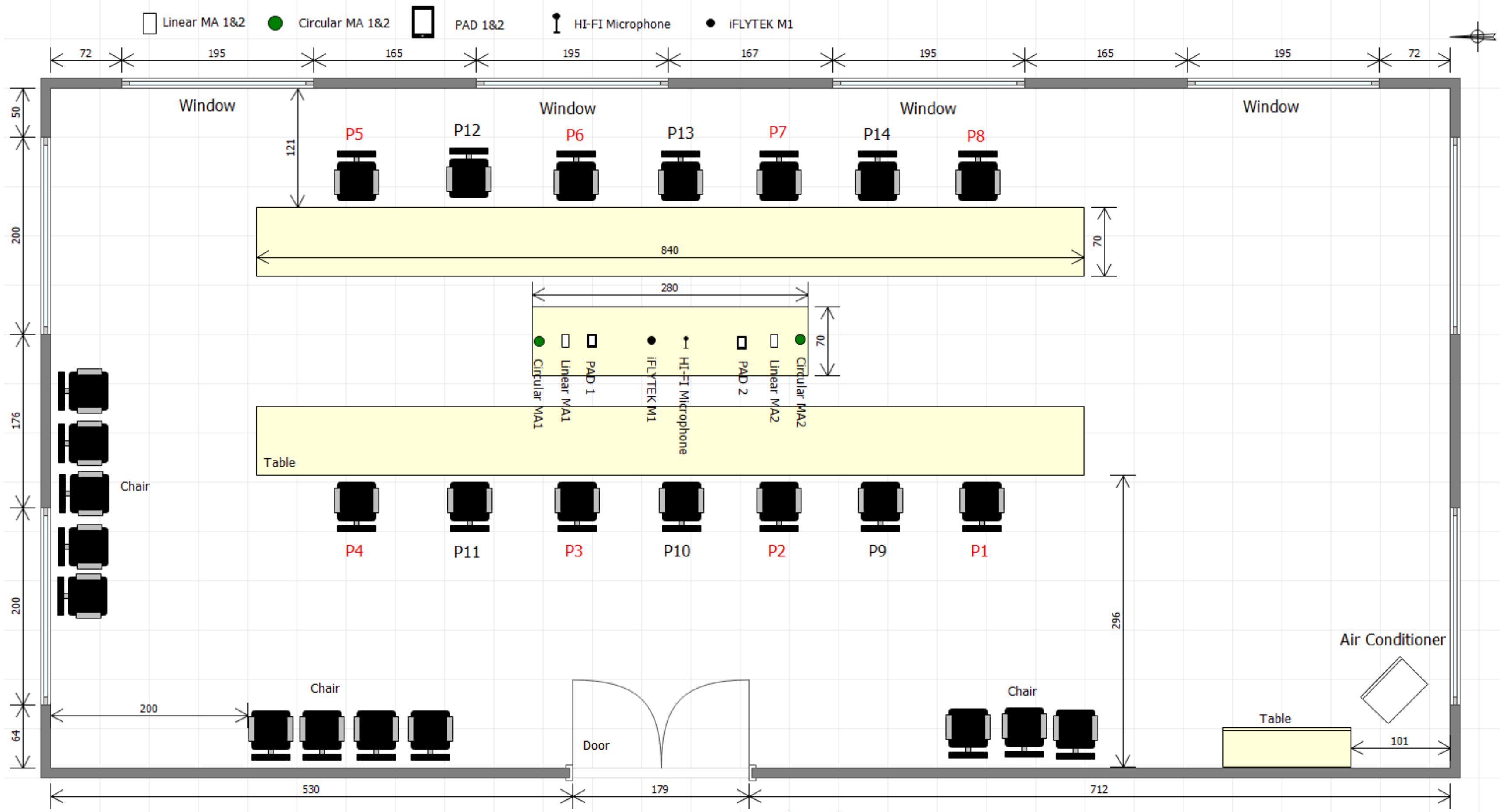
10

Recording Device

8-chnanel

circular microphone array (16kHz, 16-bit);

AISHELL-ASR0055



Meeting Room Schema [LR003]

370H 162 Speaker

Meeting sessions

639

Meeting rooms

20

Recording Device

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)

2 AISHELL-4 数据库介绍

http://www.openslr.org/111/

http://www.aishelltech.com/aishell_4

OpenSLR

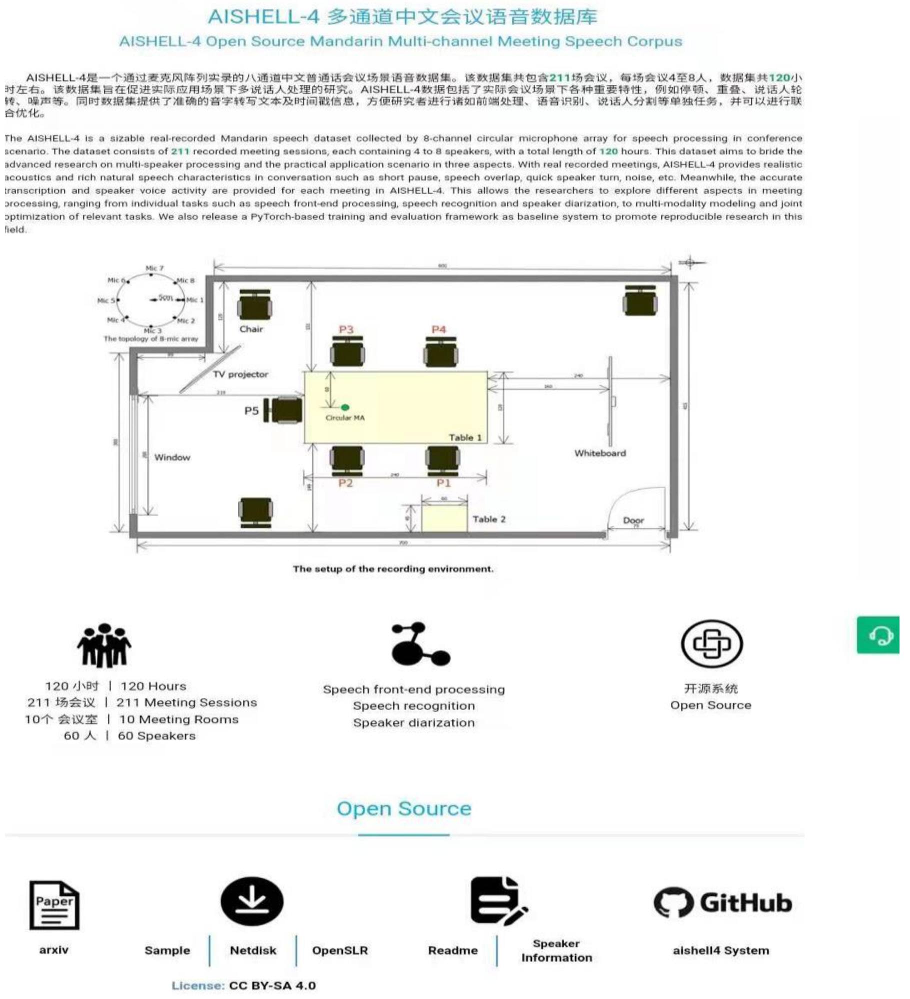
Home Resources

AISHELL-4

Identifier: SLR111
Summary: A Free Mandarin Multi-channel Meeting Speech Corpus, provided by Beijing Shell Shell Technology Co.,Ltd
Category: Speech
License: CC BY-SA 4.0
Downloads (use a mirror closer to you):
train_L.tar.gz [7.0G] (Training set of large room, 8-channel microphone array speech) Mirrors: [China]
train_M.tar.gz [25G] (Training set of medium room, 8-channel microphone array speech) Mirrors: [China]
train_S.tar.gz [14G] (Training set of small room, 8-channel microphone array speech) Mirrors: [China]
test.tar.gz [5.2G] (Test set) Mirrors: [China]
About this resource:

The AISHELL-4 is a sizable real-recorded Mandarin speech dataset collected by 8-channel circular microphone array for speech process dataset consists of 211 recorded meeting sessions, each containing 4 to 8 speakers, with a total length of 120 hours. This dataset aims research on multi-speaker processing and the practical application scenario in three aspects. With real recorded meetings, AISHELL-4 p rich natural speech characteristics in conversation such as short pause, speech overlap, quick speaker turn, noise, etc. Meanwhile, the speaker voice activity are provided for each meeting in AISHELL-4. This allows the researchers to explore different aspects in meeting p individual tasks such as speech front-end processing, speech recognition and speaker diarization, to multi-modality modeling and joint We also release a PyTorch-based training and evaluation framework as a baseline system to promote reproducible research in this fiel generated samples are available here.

You can cite the data using the following BibTeX entry:



2 AISHELL-4 数据库介绍

https://github.com/felixfuyihui/AISHELL-4

AISHELL-4

This project is associated with the recently-released AIHSHELL-4 dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. The project, served as baseline, is divided into five parts, named ***data_preparation***, ***front_end***, ***asr*** and ***sd***. The Speaker Independent (SI) task only evaluates the ability of front end (FE) and ASR models, while the Speaker Dependent (SD) task evaluates the joint ability of speaker diarization, front end and ASR models. The goal of this project is to simplify the training and evaluation procedure and make it easy and flexible for researchers to carry out experiments and verify neural network based methods.

Setup

```
git clone https://github.com/felixfuyihui/AISHELL-4.git
pip install -r requirements.txt
```

Introduction

- Data Preparation: Prepare the training and evaluation data.
- Front End: Train and evaluate the front end model.
- ASR: Train and evaluate the asr model.
- Speaker Diarization: Generate the speaker diarization results.
- Evaluation: Evaluate the results of models above and generate the CERs for Speaker Independent and Speaker Dependent tasks respectively.

General steps

1. Generate training data for fe and asr model and evaluation data for Speaker Independent task.
2. Do speaker diarization to generate rttm which includes vad and speaker diarization information.
3. Generate evaluation data for Speaker Dependent task with the results from step 2.
4. Train FE and ASR model respectively.
5. Generate the FE results of evaluation data for Speaker Independent and Speaker Dependent tasks respectively.
6. Generate the ASR results of evaluation data for Speaker Independent and Speaker Dependent tasks respectively with the results from step 2 and 3 for No FE results.
7. Generate the ASR results of evaluation data for Speaker Independent and Speaker Dependent tasks respectively with the results from step 5 for FE results.
8. Generate CER results for Speaker Independent and Speaker Dependent tasks of (No) FE with the results from step 6 and 7 respectively.

Contributors

音频语音与语言处理研究组 ASLP Audio, Speech and Language Processing Group NPU
希尔贝壳 AISHELL

AISHELL联合西北工业大学、中国科学技术大学、微软合著的论文

INTERSPEECH 2021接收

Cornell University
We ... the Simons

arXiv.org > cs > arXiv:2104.03603

Search...
Help | Advanced S

Computer Science > Sound

[Submitted on 8 Apr 2021 (v1), last revised 18 Jun 2021 (this version, v2)]

AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, Jingdong Chen

In this paper, we present AISHELL-4, a sizable real-recorded Mandarin speech dataset collected by 8-channel circular microphone array for speech processing in conference scenario. The dataset consists of 211 recorded meeting sessions, each containing 4 to 8 speakers, with a total length of 118 hours. This dataset aims to bride the advanced research on multi-speaker processing and the practical application scenario in three aspects. With real recorded meetings, AISHELL-4 provides realistic acoustics and rich natural speech characteristics in conversation such as short pause, speech overlap, quick speaker turn, noise, etc. Meanwhile, the accurate transcription and speaker voice activity are provided for each meeting in AISHELL-4. This allows the researchers to explore different aspects in meeting processing, ranging from individual tasks such as speech front-end processing, speech recognition and speaker diarization, to multi-modality modeling and joint optimization of relevant tasks. Given most open source dataset for multi-speaker tasks are in English, AISHELL-4 is the only Mandarin dataset for conversation speech, providing additional value for data diversity in speech community.

Comments: Accepted by Interspeech 2021
Subjects: **Sound (cs.SD)**; Audio and Speech Processing (eess.AS)
Cite as: arXiv:2104.03603 **[cs.SD]**
(or arXiv:2104.03603v2 **[cs.SD]** for this version)

AISHELL-4 模型框架综述

1. 前端模型
2. ASR模型
3. 说话人分割模型

付艺辉，硕士研究生
西北工业大学音频语音与语言处理研究组
导师为谢磊教授
主要研究方向为语音前端处理及语音前后端结合

前端模型

1. 模型介绍

基于深度学习的MVDR模型，通过LSTM估计两个目标语音的mask，随后进行自适应波束形成(MVDR)，得到增强分离后的语音

2. 训练数据

语音：Librispeech

噪声：MUSAN

RIR: 镜像法仿真，RT60 [0.6,1.2]s

语音重叠：训练数据分为三等份：1)无重叠，2)重叠率[0,0.2], 3)重叠率[0.2,0.8]

数据量：364h

ASR模型

1. 模型介绍

   基于Transformer的端到端语音识别模型，使用CTC和CE loss联合训练

2. 训练数据

   语音：AISHELL-1, aidatatang_200zh, Primewords,AISHELL-4实录数据

   噪声：MUSAN

   RIR: 镜像法仿真，RT60 [0.6,1.2]s

   数据量：768h

说话人分割模型

模型介绍

1. VAD模型使用KALDI开源的Chime6的SAD模型

2. 说话人切分模型是基于BUT的VBx系统：ResNet101 + PLDA + AHC + VBx 的方案。

Speaker Embedding的训练数据是Voxceleb + CN-Celeb

http://kaldi-asr.org/models.html

https://github.com/BUTSpeechFIT/VBx/tree/master/VBx

模型开源 https://github.com/felixfuyihui/AISHELL-4

felixfuyihui Update Readme.md 598bbc1 yesterday 71 commits

asr	Update Readme.md	3 days ago
data_preparation	Update Readme.md	yesterday
eval	Update Readme.md	3 days ago
front_end	Update Readme.md	4 days ago
sd	Delete requirements.txt	4 days ago
LICENSE	Create LICENSE	9 days ago
README.md	Update README.md	4 days ago
fig_aishell.jpg	fug	5 days ago
fig_aslp.jpg	fug	5 days ago
requirements.txt	Update requirements.txt	4 days ago

Introduction

- Data Preparation: Prepare the training and evaluation data.
- Front End: Train and evaluate the front end model.
- ASR: Train and evaluate the asr model.
- Speaker Diarization: Generate the speaker diarization results.
- Evaluation: Evaluate the results of models above and generate the CERs for Speaker Independent and Speaker Dependent tasks respectively.

General steps

1. Generate training data for fe and asr model and evaluation data for Speaker Independent task.
2. Do speaker diarization to generate rttm which includes vad and speaker diarization information.
3. Generate evaluation data for Speaker Dependent task with the results from step 2.
4. Train FE and ASR model respectively.
5. Generate the FE results of evaluation data for Speaker Independent and Speaker Dependent tasks respectively.
6. Generate the ASR results of evaluation data for Speaker Independent and Speaker Dependent tasks respectively with the results from step 2 and 3 for No FE results.
7. Generate the ASR results of evaluation data for Speaker Independent and Speaker Dependent tasks respectively with the results from step 5 for FE results.
8. Generate CER results for Speaker Independent and Speaker Dependent tasks of (No) FE with the results from step 6 and 7 respectively.

模型训练及测试流程

1. 数据准备，包括准备RIR, 各向同性噪声，前端及ASR的训练数据，说话人无关及说话人相关的测试数据
2. 通过说话人分割模型获取说话人分割及VAD模型
3. 分别训练前端及ASR模型
4. 获取测试数据的前端模型推理结果
5. 获取步骤4的ASR推理结果
6. 通过Asclite2工具包获取说话人无关及说话人相关CER结果

4 未来展望

研发建设不完善的语言数据

方言

民族语言

多语并行

等等

结合图像、感知等的数据来形成多模态智能语音数据

图像识别的痛点：光线、动作、属性等

语音识别的痛点：语速、口音、噪音等

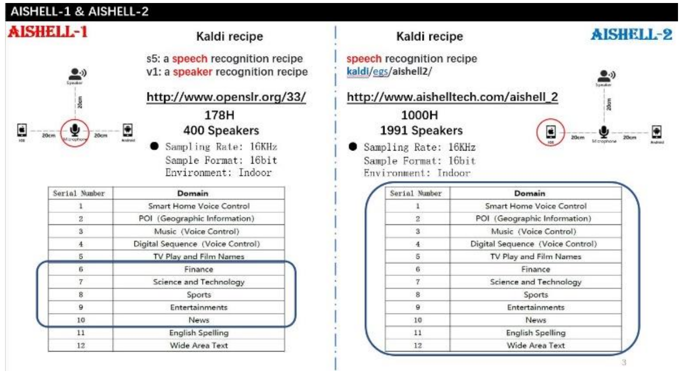
语义理解的痛点：知识不足、模糊处理等

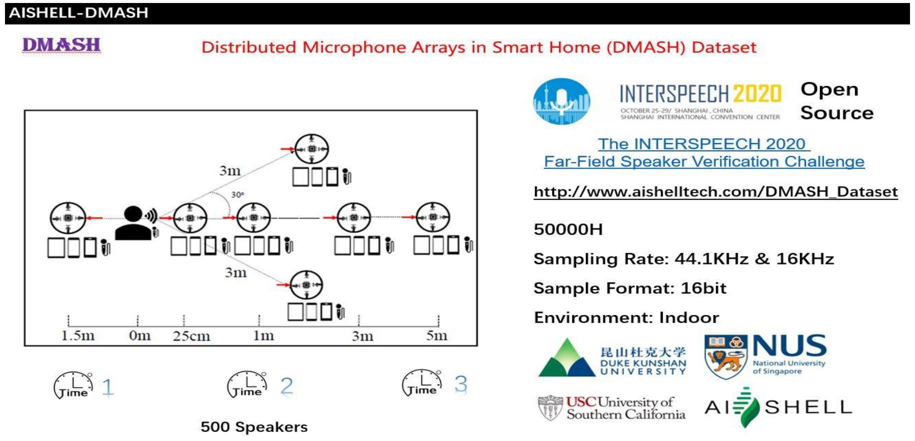
单模态向多模态转变，优点互补优化高智能系统

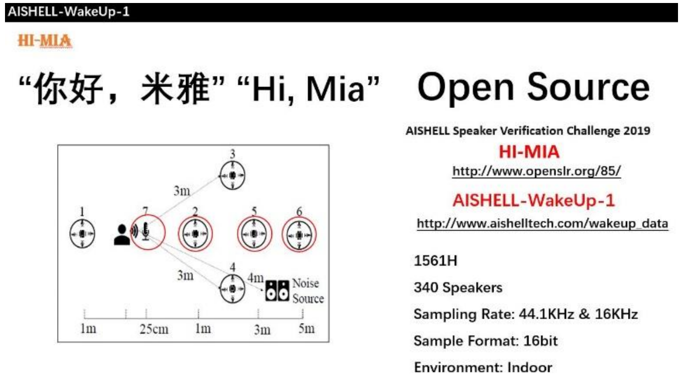
数据的开源、开放

从用国外的数据到国外用我们的数据

558 **200**







AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario

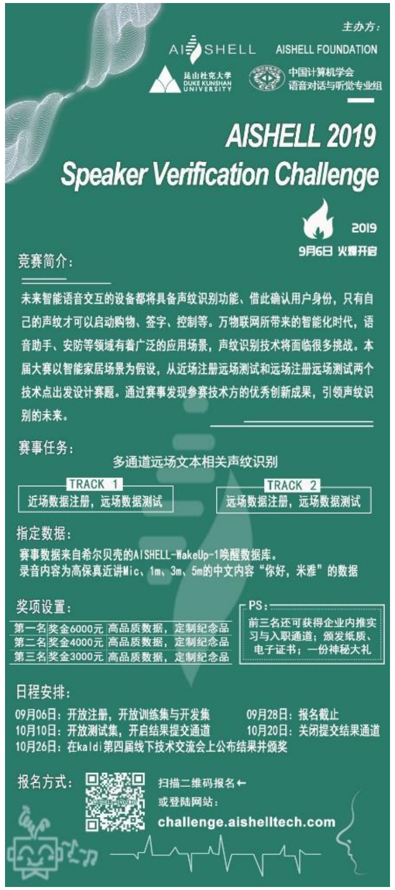
Yihui Fu[1,], Luyao Cheng[1,*], Shubo Lv[1], Yukai Jv[1], Yuxiang Kong[1], Zhuo Chen[3]*
Yanxin Hu[1], Lei Xie[1], Jian Wu[3], Hui Bu[2], Xin Xu[2], Jun Du[4], Jingdong Chen[1]

[1]Northwestern Polytechnical University, Xi'an, China
[2]Beijing Shell Shell Technology Co., Ltd., Beijing, China
[3]Microsoft Corporation
[4]University of Science and Technology of China, Hefei, China

{yhfu, lycheng, shblv, ykjv, yxkong, yxhu}@npu-aslp.org, lxie@nwpu.edu.cn, {zhuc, wujian}@microsoft.com, {buhui, xuxin}@aishelldata.com, jundu@ustc.edu.cn, jingdongchen@ieee.org

4 未来展望

高价值、前沿的赛事
AI语音技术的第一靶场



INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI , CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

FFSVC 2020
Far-Field Speaker Verification Challenge

昆山杜克大学 DUKE KUNSHAN UNIVERSITY
NUS National University of Singapore
USC University of Southern California
AISHELL

Distributed Microphone Arrays in Smart Home Database. (DMASH)

Team	Paper
184	5+

http://2020.ffsvc.org/

Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing (ConferencingSpeech 2021)

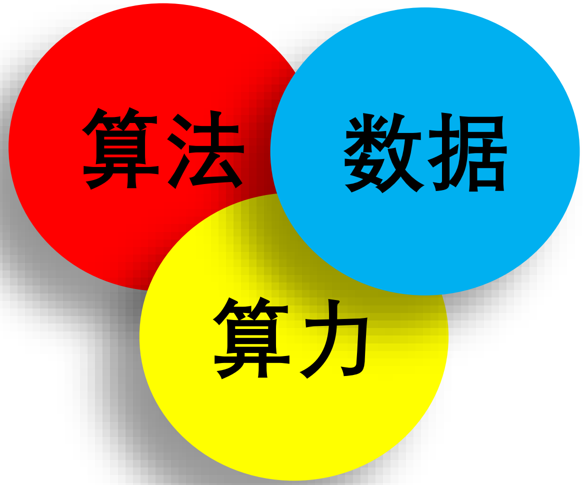
The ConferencingSpeech 2021 challenge is proposed to stimulate research in multi-channel speech enhancement and aims for processing the far-field speech from microphone arrays in the video conferencing rooms. Targeting the real video conferencing room application, the ConferencingSpeech 2021 challenge database is recorded from real speakers. The number of speakers and distances between speakers and microphone arrays vary according to the sizes of meeting rooms. Multiple microphone arrays from three different types of geometric topology are allocated in each recording environment.

Multi-Speaker Multi-Style Voice Cloning Challenge (M2VoC)

Text-to-speech (TTS) or speech synthesis has witnessed significant performance improvement with the help of deep learning. The latest advances in end-to-end text-to-speech paradigm and neural vocoder have enabled us to produce very realistic and natural-sounding synthetic speech reaching almost human-parity performance. But this amazing ability is still limited to the ideal scenarios with a large single-speaker less-expressive training set. The speech quality, target similarity, expressiveness and robustness are still not satisfied for synthetic speech with different speakers and various styles, especially in real-world low-resourced conditions, e.g., each speaker only has a few samples at hand. The current open solutions are also not robust enough to unseen speakers. We call this challenging task as **multi-speaker multi-style voice cloning (M2VoC).**

人才社区的开源、开放

AI人才的建设应该是AI新基建最应该夯实的



AISHELL会持续投入做开源，为人工智能民主化。

感谢一路合作过的伙伴：**AISHELL Foundation、KALDI社区、昆山杜克大学语音与多模式智能信息处理实验（DUK SMIIP Lab）、西北工业大学音频语音与语言处理研究组（ASLP@NPU）、清华大学语音和语言技术中心（CSLT@Tsinghua University）、中国科学技术大学、新加坡国立大学、微软、小米、腾讯天籁实验室**等。

THANKS

感谢一直支持AISHELL的开发者