HI-MIA声纹识别实战

第1节-说话人识别简介

讲师：覃晓逸

课程目录：

1.1 课程目标

课程目标：深入了解 HIMIA 数据库并能独立实现声纹模型的训练与推理

HI-MIA

Identifier: SLR85
Summary: A far-field text-dependent speaker verification database for AISHELL Speaker Verification Challenge 2019
Category: Speech
License: Apache License v.2.0
Downloads (use a mirror closer to you):
train.tar.gz [36G] (Training set with speaker dependent sub folders) Mirrors: [US] [EU] [CN]
dev.tar.gz [5.1G] (Dev set with speaker dependent sub folders) Mirrors: [US] [EU] [CN]
test.tar.gz [4.7G] (Test set with target/non-target answer) Mirrors: [US] [EU] [CN]
test_v2.tar.gz [4.7G] (Updated test set fixing corrupted audio files) Mirrors: [US] [EU] [CN]
filename_mapping.tar.gz [5.9M] (Filename mapping rules for multi-channel information) Mirrors: [US] [EU] [CN]

About this resource:
The data is used in AISHELL Speaker Verification Challenge 2019. It is extracted from a larger database called AISHELL-WakeUp-1.

The contents are wake-up words "Hi, Mia" in both Chinese and English. The data is collected in real home environment using microphone arrays and Hi-Fi microphone. The collection process and development of a baseline system was described in the paper below. The data used in the challenge is extracted from 1 Hi-Fi microphone and 16-channel circular microphone arrays for 1/3/5 meters. And the contents are the Chinese wake-up words. The whole set is divided into train (254 people), dev (42 people) and test (44 people) subsets. Test subset is provided with paired target/non-target answer to evaluate verification results.

You can cite the data using the following BibTeX entry:

```
@misc{himia,
    title={HI-MIA : A Far-field Text-Dependent Speaker Verification Database and the Baselines},
    author={Xiaoyi Qin and Hui Bu and Ming Li},
    year={2019},
    eprint={1912.01231},
    archivePrefix={arXiv},
    primaryClass={cs.SD}
}
```

External URL: http://aishelltech.com/wakeup_data

HI-MIA-CW

Identifier: SLR120
Summary: A Free Mandarin Supplemental Speech Corpus to HI-MIA Database, whose contents are negative samples for wake-up words "Hi, Mia".
Category: Speech
License: CC BY-SA 4.0
Downloads (use a mirror closer to you):
data.tgz [550M] (Speech files) Mirrors: [US] [EU] [CN]
resource.tgz [55K] (Speaker info and transcriptions) Mirrors: [US] [EU] [CN]

About this resource:
The HI-MIA-CW is a supplemental database to the HI-MIA wakeup database, and we used the same setup of HI-MIA database to further record 16434 audios.

The specific text of the audios is the HI-MIA confusion words in Chinese, which are the negative samples for wake-up words "hi, Mia" (ni hao mi ya). The text details can be found in the paper and the transcription file in resources. Each audio sample was recorded in real home environment using high fidelity microphone (48kHz,16-bit). Then we re-sampled to 16kHz to build the database. It contains 35 speakers. There is no overlap between these 35 speakers and the speakers who are in the previous HI-MIA database. This dataset aims to promote the advanced research on wakeup words detection. It serves as negative samples for the wakeup words detection system. It helps researchers test the performance when encountering the confusing words.

You can cite the data using the following BibTeX entry:

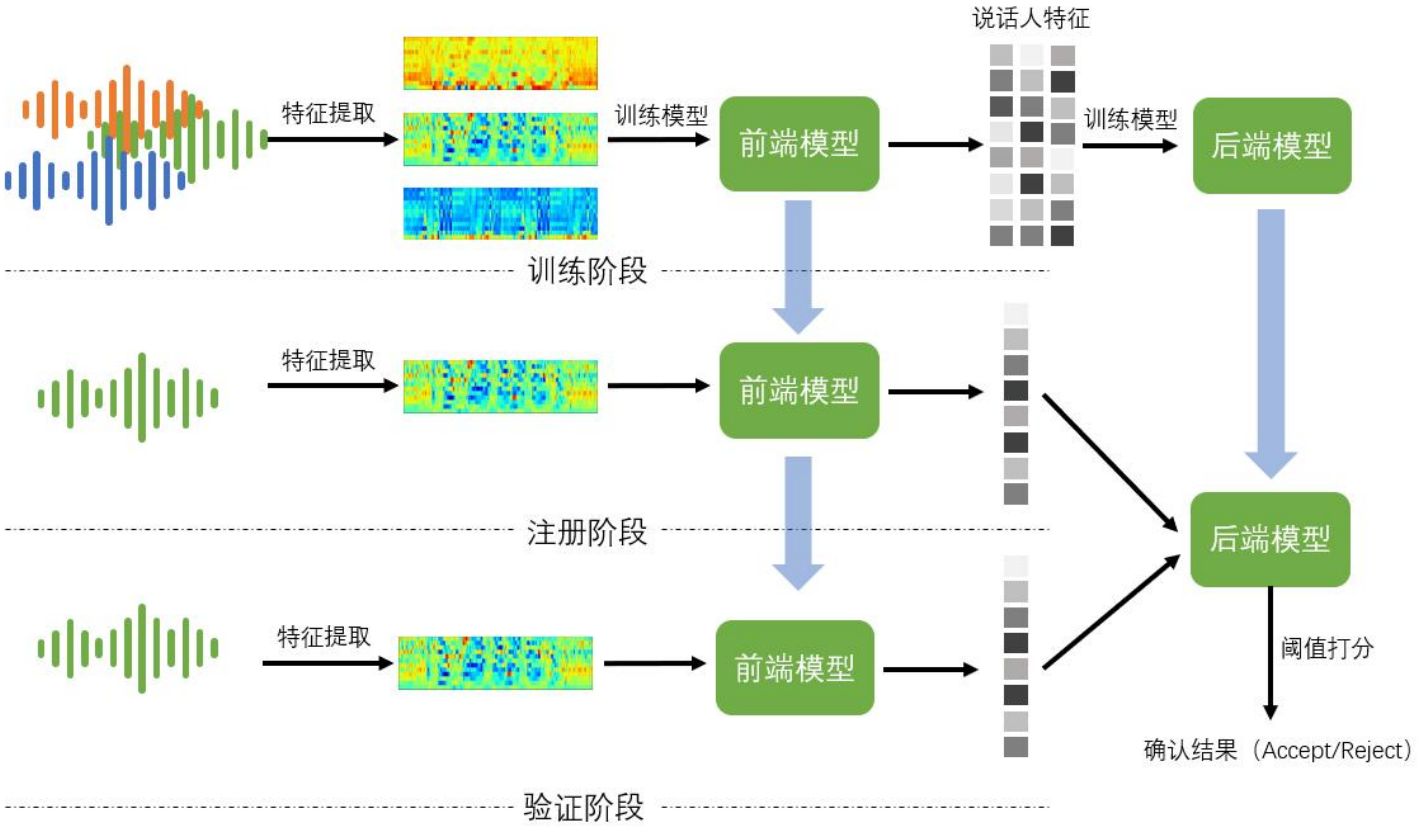
https://www.openslr.org/85/

https://www.openslr.org/120/

https://www.aishelltech.com/SVC_2019
http://www.aishelltech.com/wakeup_data

1.1 课程目标

课程目标：深入了解 HIMIA 数据库并能独立实现声纹模型的训练与推理



1.1 课程目标

课程目标：深入了解 HIMIA 数据库并能独立实现声纹模型的训练与推理

针对人群：有一定的深度学习基础和编程能力

1.1 课程目标

A.1说话人识别简介

A.2 基于端到端的说话人识别

B.1 模型实现

B.2 HIMIA库数据训练

1.2 说话人识别简介

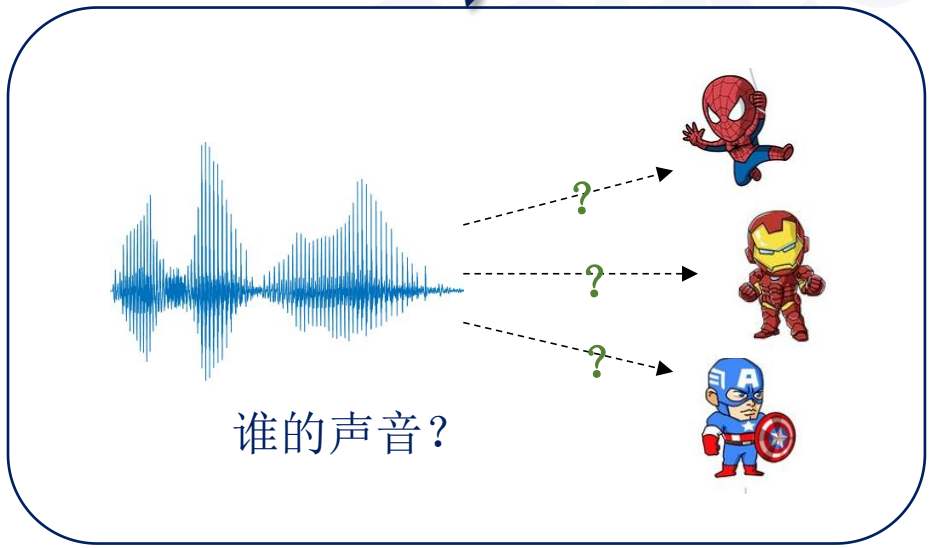
自动说话人识别（Automatic Speaker Verification，ASV），又称声纹识别，是一种通过语音信号来辨识和确认说话人身份的生物识别技术。

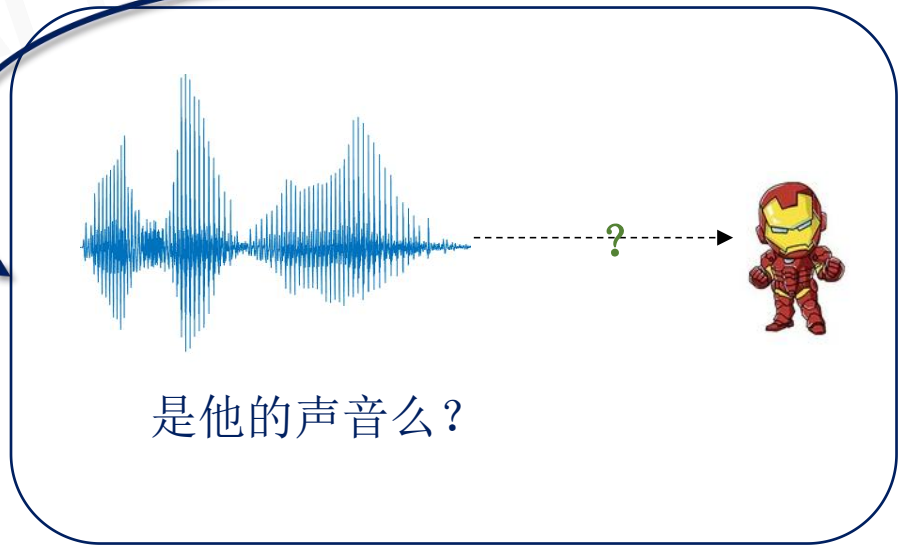
按内容分类

- 文本相关（text-dependent）
- 文本无关（text-independent）

按任务分类

- 说话人辨别（Speaker Recognition）
- 说话人确认（Speaker Verification）





1.2 说话人识别简介

自动说话人识别（Automatic Speaker Verification，ASV），又称声纹识别，是一种通过语音信号来辨识和确认说话人身份的生物识别技术。

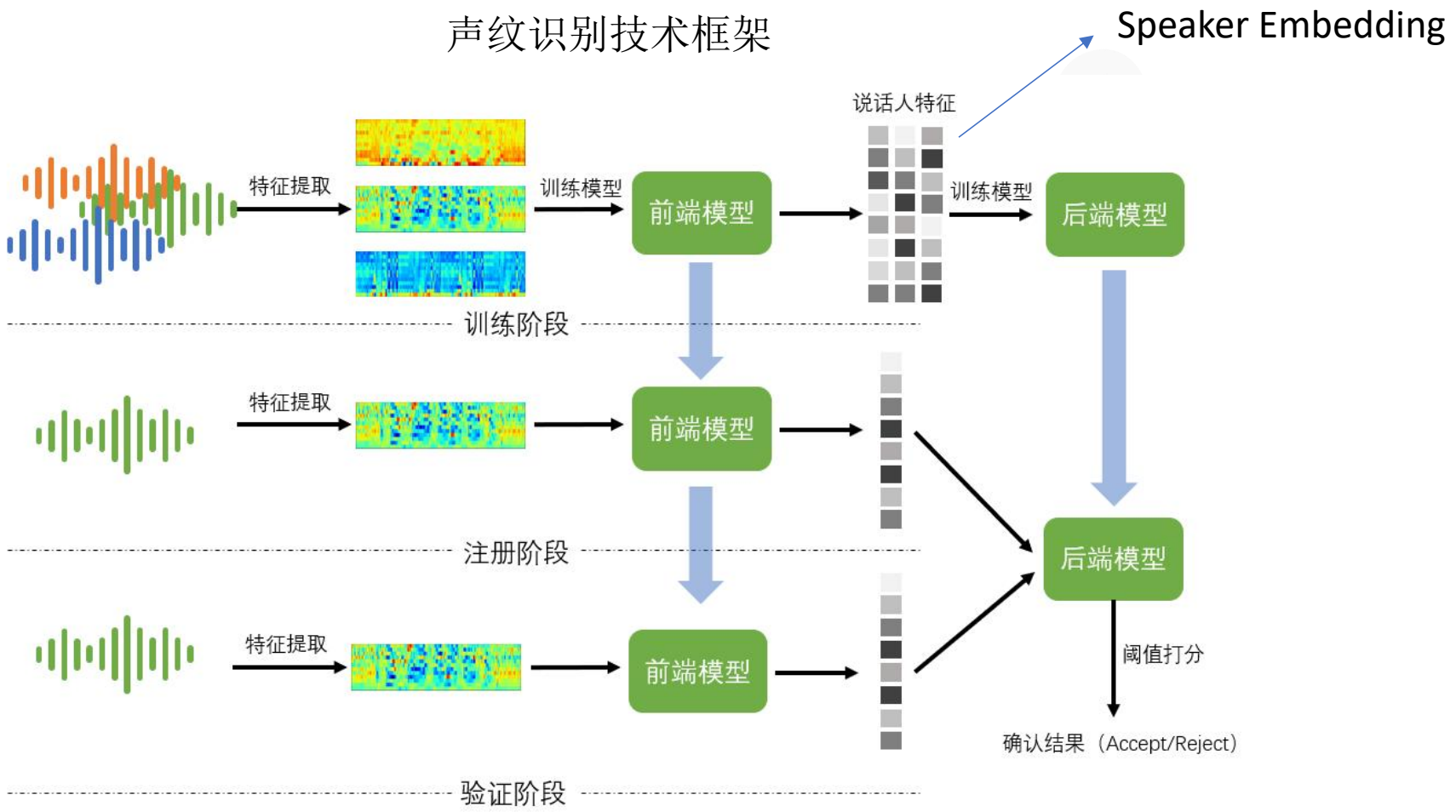
说话人识别应用

- 公安司法鉴定，辅助举证手段
- 智能家居，智能音箱的控制，智能座舱
- 安全防卫，微信的声音锁、银行客户、小区门禁
- 军事国防：特定人监控，命令保护
- 智能客服：身份确认，VIP客服

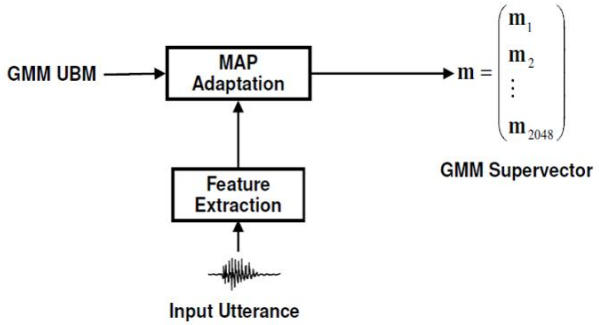
关联任务

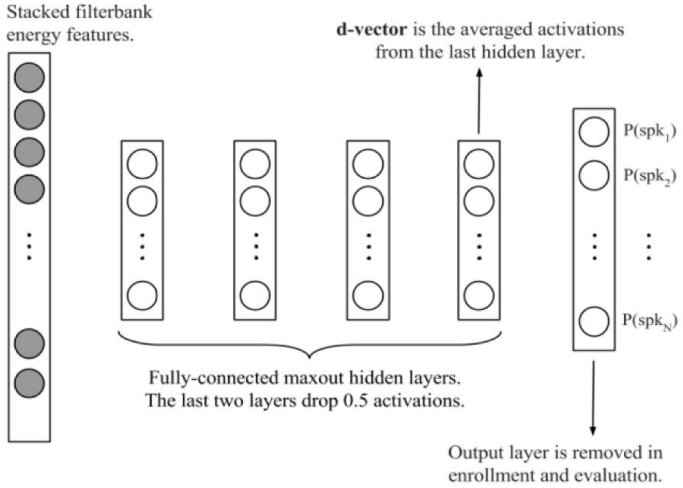
- 说话人日志（Speaker diarisation）
- 特定说话人分离（Target speaker separation）
- 特定说话人合成/变声（Target speaker synthesis/voice conversion）
- 目标说话人语音识别（Target speaker ASR）
- 语音分类任务（Speech Classification）

1.2 说话人识别简介



1.3 发展及研究现状





ResNet
F-TDNN
D-TDNN
ECAPA-TDNN

GMM-UBM (Reynolds et al., 2000)

D-vector (Variani et al. 2014)

End-to-End

1964年
Talker recognition

GMM-UBM/i-vector with PLDA
(Dehak et al., 2010)

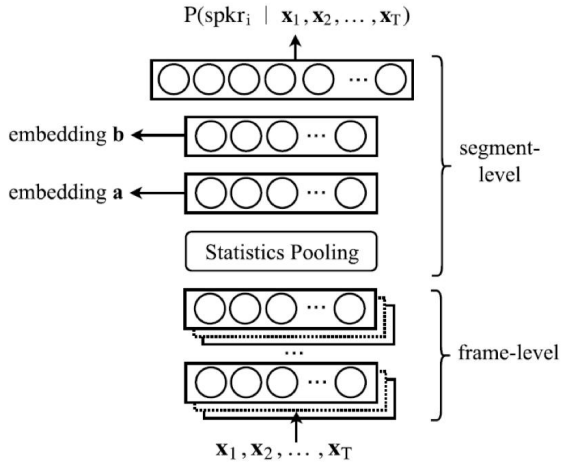
由单因子分析模型，给定说话人s一段语音k，那么这段语音的Supervector定义为：

$$M_{s,k} = m_u + T\omega_{s,k}$$

- m_u：UBM的均值
- T：全局差异空间矩阵
- $\omega_{s,k}$：低维隐变量，称为i-vector

Baum-Welch统计量 → EM（E步：S 均值方差；M步：最大似然估计）

X-vector (Snyder et al., 2017, 2018)



GMM-UBM/i-vector PLDA说话人识别系统

局部优化

多步解决问题

无监督学习

GMM-UBM

MFCC

端到端说话人识别系统

全局优化

端到端“一步完成”

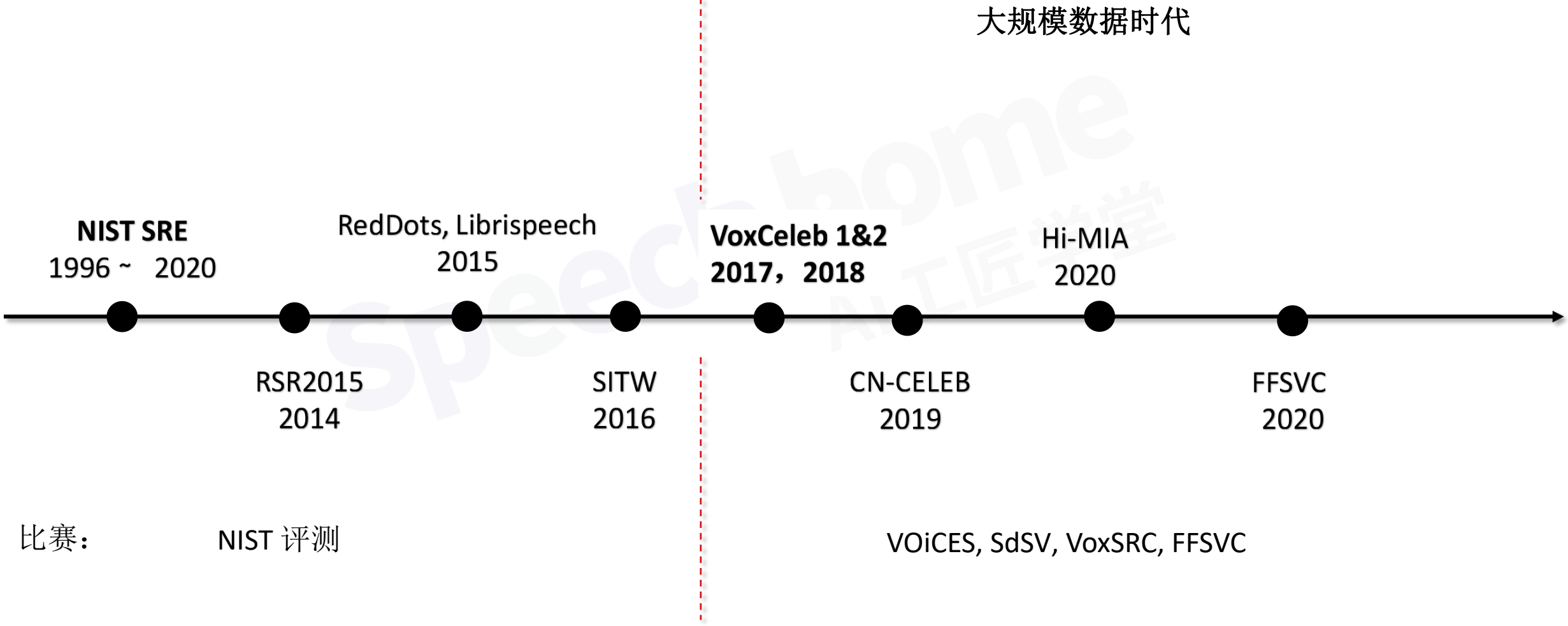
有监督学习

GAP/GSP/ASP/LDE

TDNN/CNN/RNN/Transformer

Waveform/spectrum/MFbank/MFCC

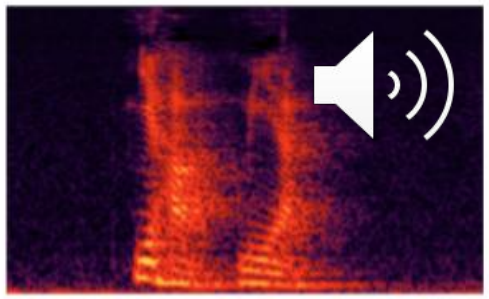
1.3 发展及研究现状



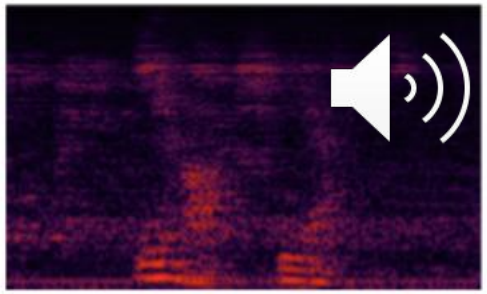
1.3 发展及研究现状

1.4 面临的挑战

远场



(a) iPhone 0.25M



(d) Mic Array 5M

跨语种

- 训练英文
  测试中文
- 注册中文
  测试英文

波斯语

英语

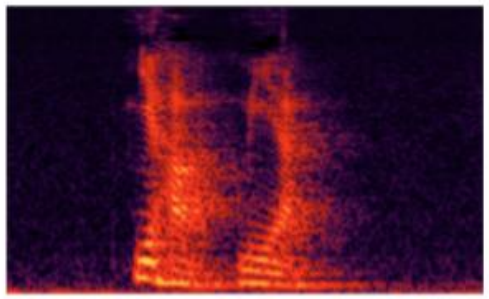
短语音

1s左右的语音

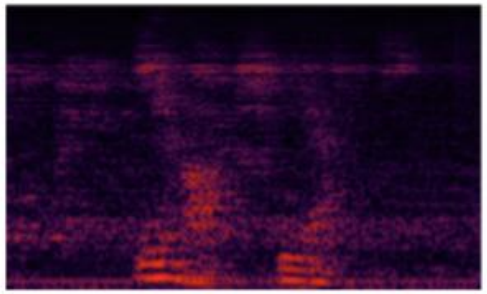
长时间跨度

1.4 面临的挑战

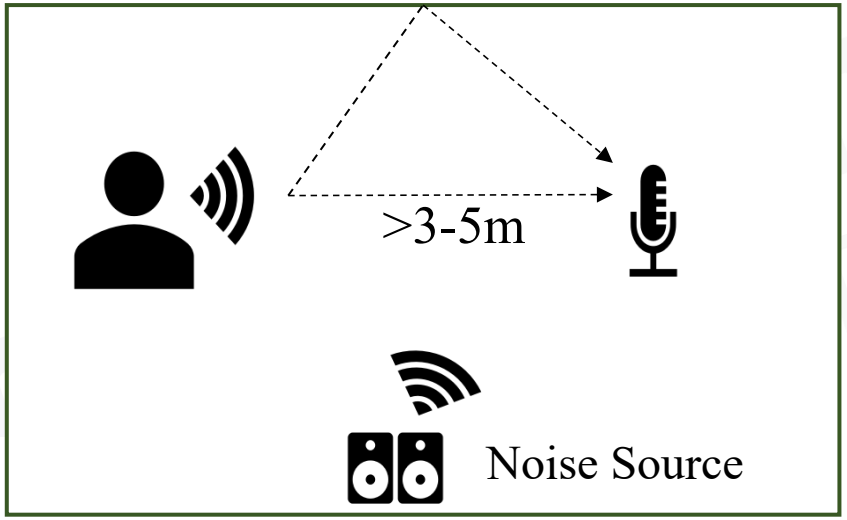
远场

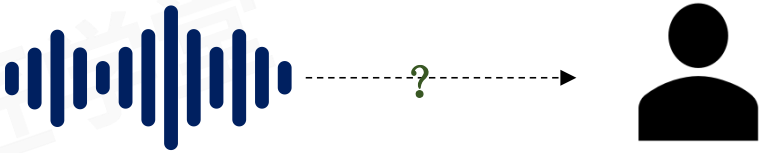


(a) iPhone 0.25M



(d) Mic Array 5M





远场说话人确认任务

课程问题可随时联系班主任