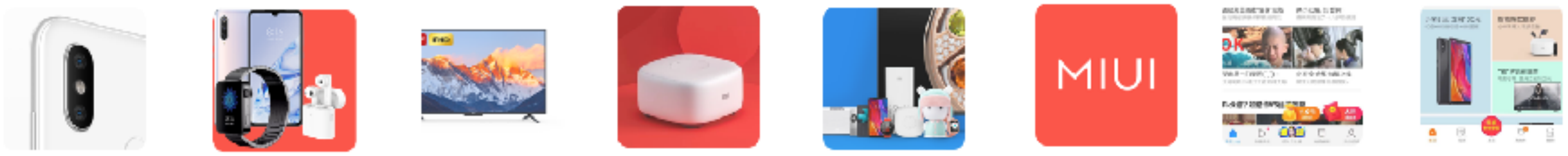
声音感知技术的无障碍协同应用

张俊博

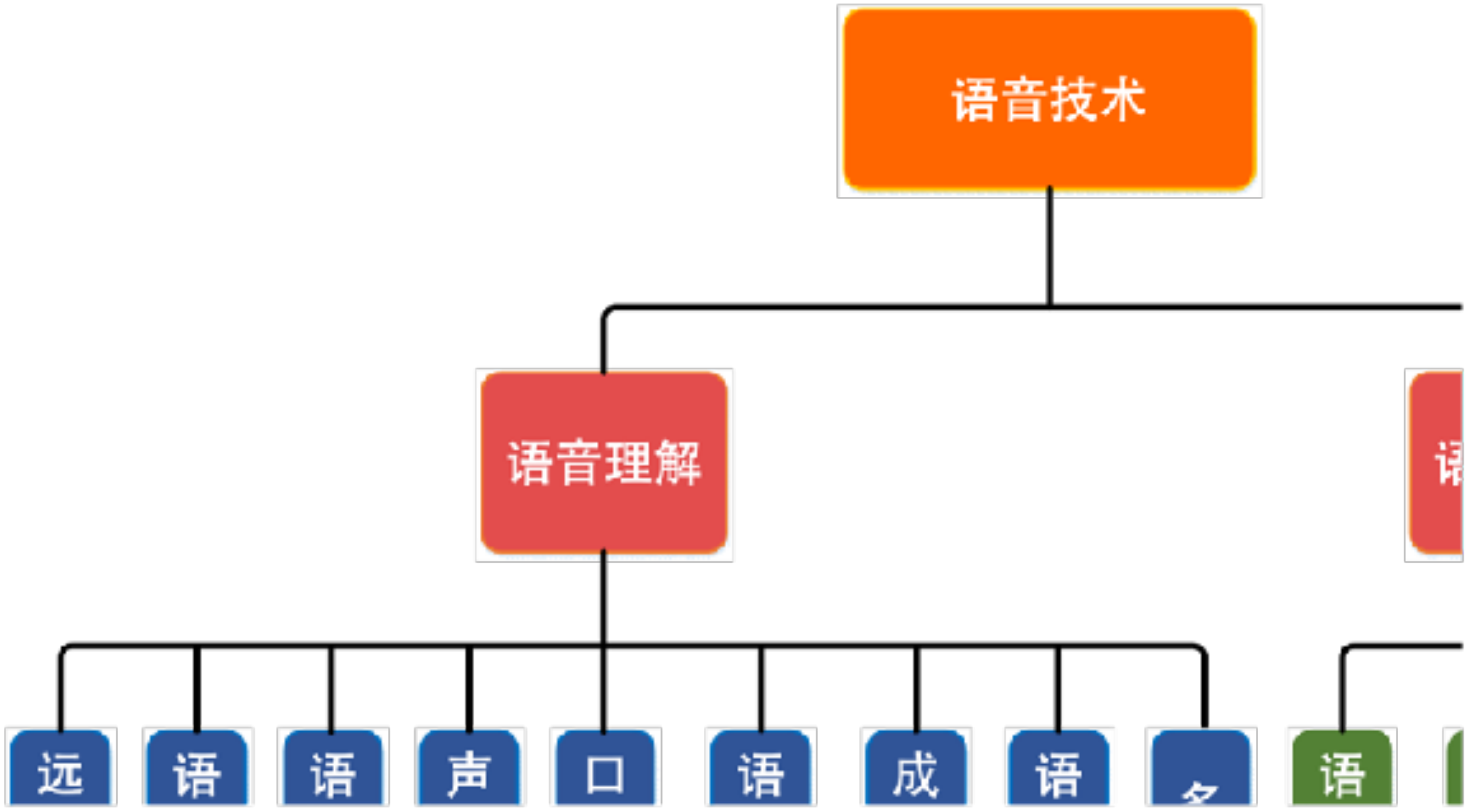
小米集团技术委员会
Xiaomi Technical Committee

小米 AI 实验室 - 声学语音组



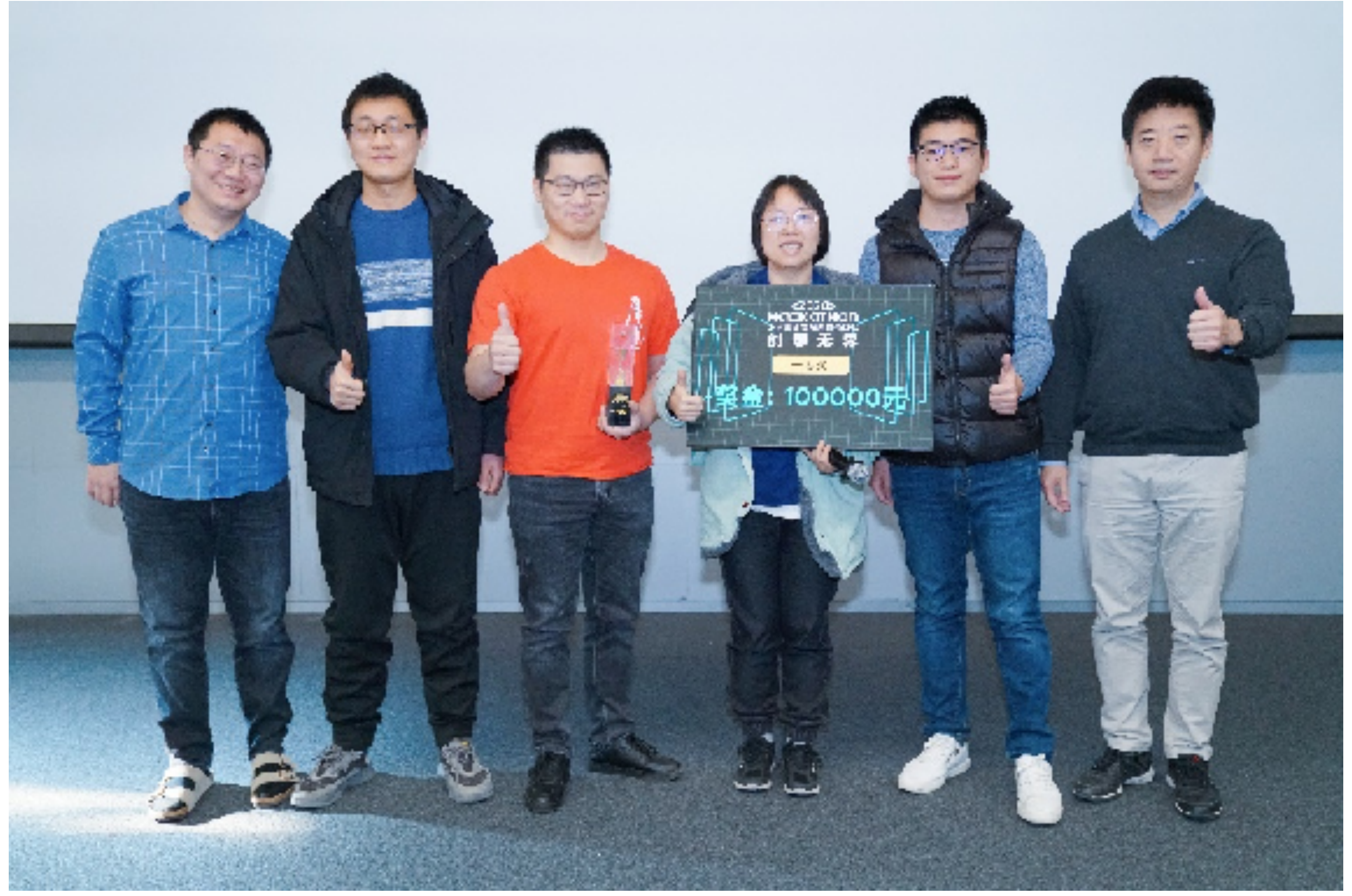
手机与手机周边 AIoT MIUI 销服

图像	声学语音	NLP	知识图谱	深度学习平台工具



Mi-Speech

从一届黑客马拉松比赛说起



作品 初衷

<2020> HACKATHON

我们的队伍叫DAKUI，是因为一位叫张大奎的朋友。他出生于河南焦作农村，博士毕业于北京理工大学计算机专业。因自幼患脑性瘫痪，走路、说话都比一般人更加困难。脑瘫会让肌肉协调能力受损，大奎说话比较费劲，口齿也不太清楚，和人交流的时候，听的人也需要全神贯注，配合推测和确认，才能理解，双方都挺费劲，这也阻碍了大奎和朋友们的交流。

除了大奎，还有很多人，都需要AI辅助与替代沟通技术。据统计推测，中国有600多万脑瘫患者，他们在努力地工作，认真地生活。于是我们做了这款名为“聆听”的软件，希望让大家能更好的地聆听脑瘫患者的声音，让交流变得更加轻松。

让全球每个人，都能享受科技带来的美好生活。

计算机博士张大奎

外卖员许龙庆

诗人余秀华


Interspeech 2018
2-6 September 2018, Hyderabad

Empirical Evaluation of Speaker Adaptation on DNN based Acoustic Model

Ke Wang[1,2], Junbo Zhang[2], Yujun Wang[2], Lei Xie[1]*

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Xiaomi, Beijing, China

Abstract

1. Introduction

2429 10.21437/Interspeech.2018-1897

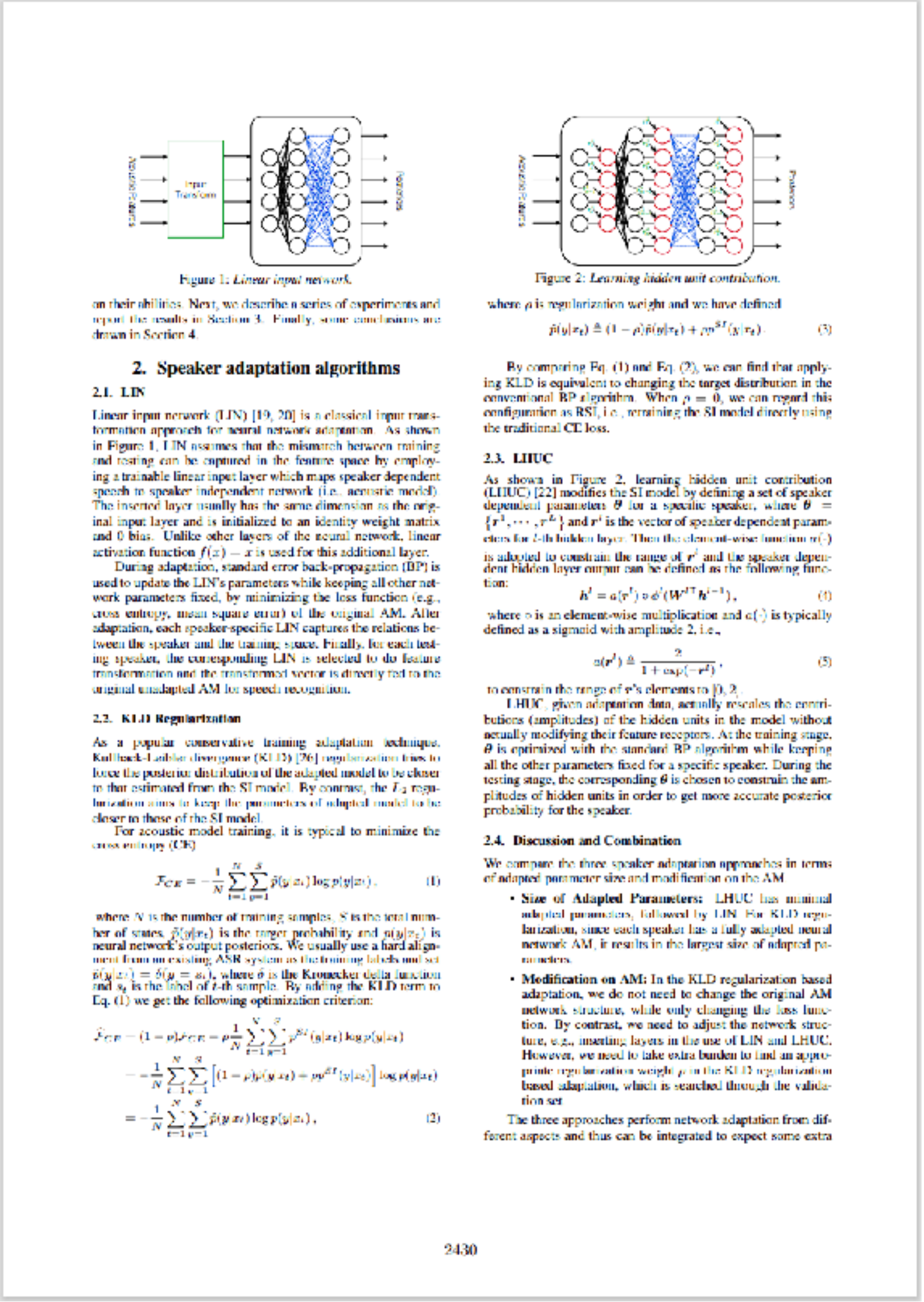


Figure 1: *Linear input network.*

Figure 2: *Learning hidden unit contribution.*

2. Speaker adaptation algorithms

2.1. LIN

2.2. KLD Regularization

2.3. LHUC

2.4. Discussion and Combination

2430

小爱语音识别服务 识别率不足 10%

使用 5 分钟的演讲数据做模型自适应

自适应后的模型 识别率大于 95%

声音不止语音

- **Human sounds**
  - *Human voice*
  - *Whistling*
  - *Respiratory sounds*
  - *Human locomotion*
  - *Digestive*
  - *Hands*
  - *Heart sounds, heartbeat*
  - *Otoacoustic emission*
  - *Human group actions*
- **Animal sounds**
  - *Domestic animals, pets*
  - *Livestock, farm animals, working animals*
  - *Wild animals*
- **Natural sounds**
  - *Wind*
  - *Thunderstorm*
  - *Water*
  - *Fire*
- **Music**
  - *Musical instrument*
  - *Music genre*
  - *Musical concepts*
  - *Music role*
  - *Music mood*
- **Sounds of things**
  - *Vehicle*
  - *Engine*
  - *Domestic sounds, home sounds*
  - *Bell*
  - *Alarm*
  - *Mechanisms*
  - *Tools*
  - *Explosion*
  - *Wood*
  - *Glass*
  - *Liquid*
  - *Miscellaneous sources*
  - *Specific impact sounds*
- **Source-ambiguous sounds**
  - *Generic impact sounds*
  - *Surface contact*
  - *Deformable shell*
  - *Onomatopoeia*
  - *Silence*
  - *Other sourceless*
- **Channel, environment and background**
  - *Acoustic environment*
  - *Noise*
  - *Sound reproduction*

2780万

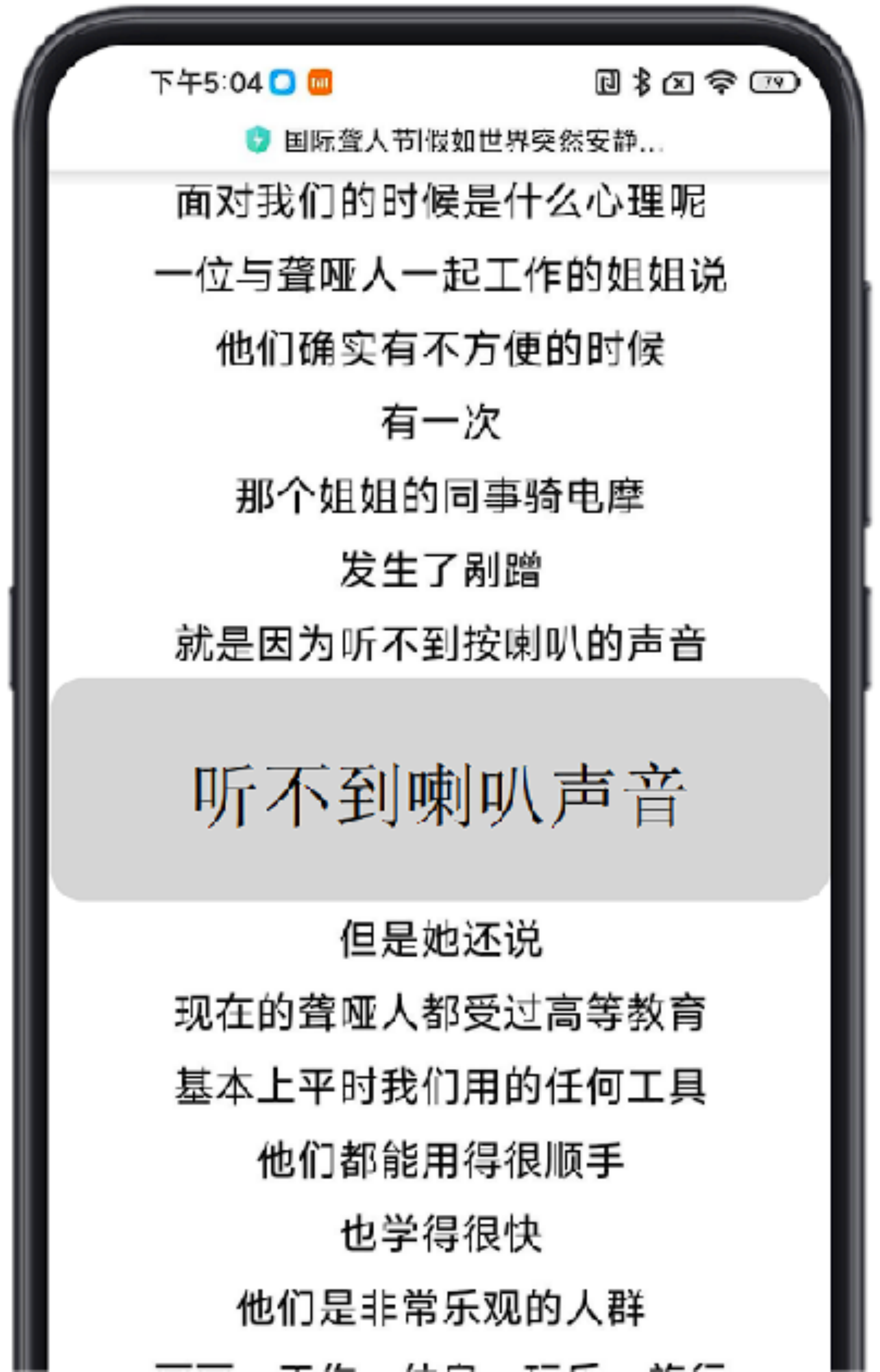
听力障碍人群

1.18亿

独居老人

在某种程度上

每个人都是“听障人士”

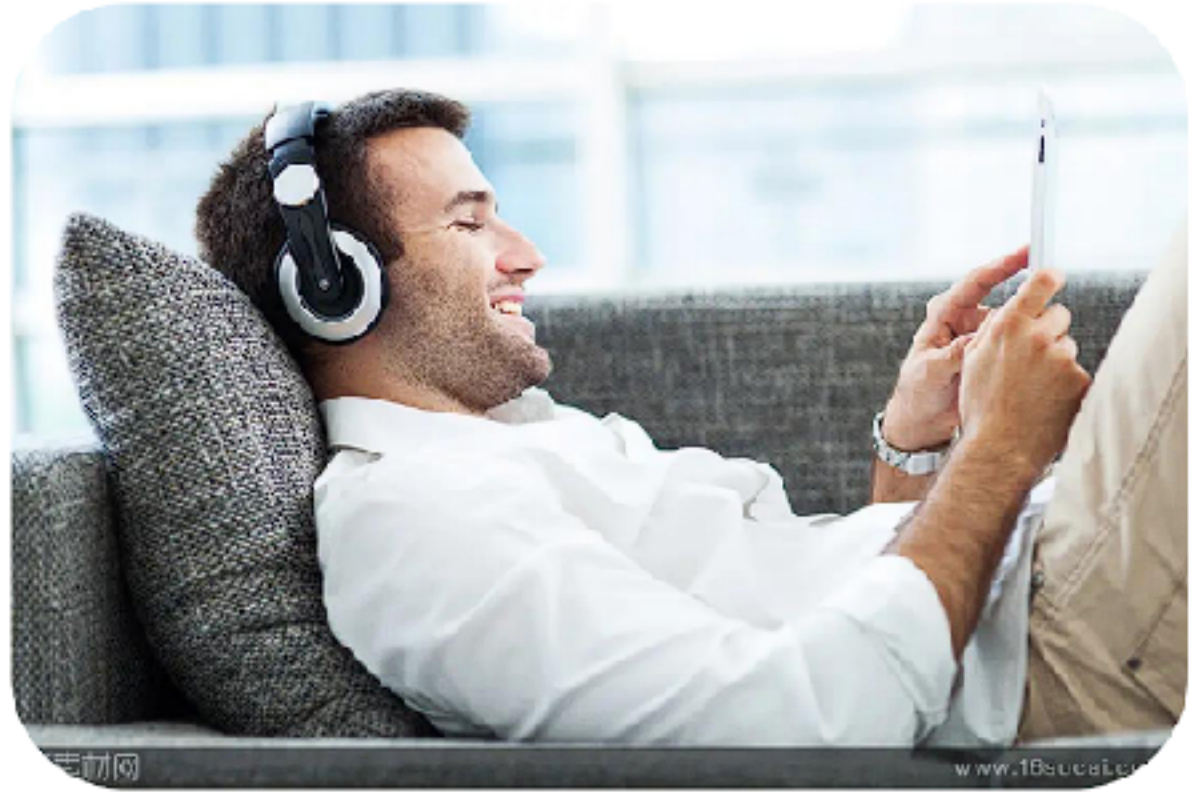


家中摔倒四天全靠敲盆求救 独居老人生命安全该如何保证

中国商报
原创 · 2020-08-05 17:04 · 中国商报官方搜狐号

中国商报/中国商网（记者 王彤旭）四天滴水未进，屋内饭已发霉，家属近20天没来看望，获救后一口气喝下一瓶水，安徽蚌埠一独居老人在家中摔倒四天靠敲盆求救的消息引起网友关注。事实上，该老人的经历绝非个例。独居老人辞世，两个月后遗体才被发现；独居老人家中猝死，尸体疑遭宠物狗啃食……随着我国步入老龄化社会，类似的新闻越来越多地出现在人们视野中，如何保证独居老人生命安全成为社会关注并亟待解决的问题。

在采访过程中，记者了解到，不少受访者认为，如果家中只剩下一位老人独居，安



让环境中的每种声音，可以被智能设备感知、识别，并反馈给用户

让听障人士“看到”周围的声音情况

让独居老人“看到”家中的声音

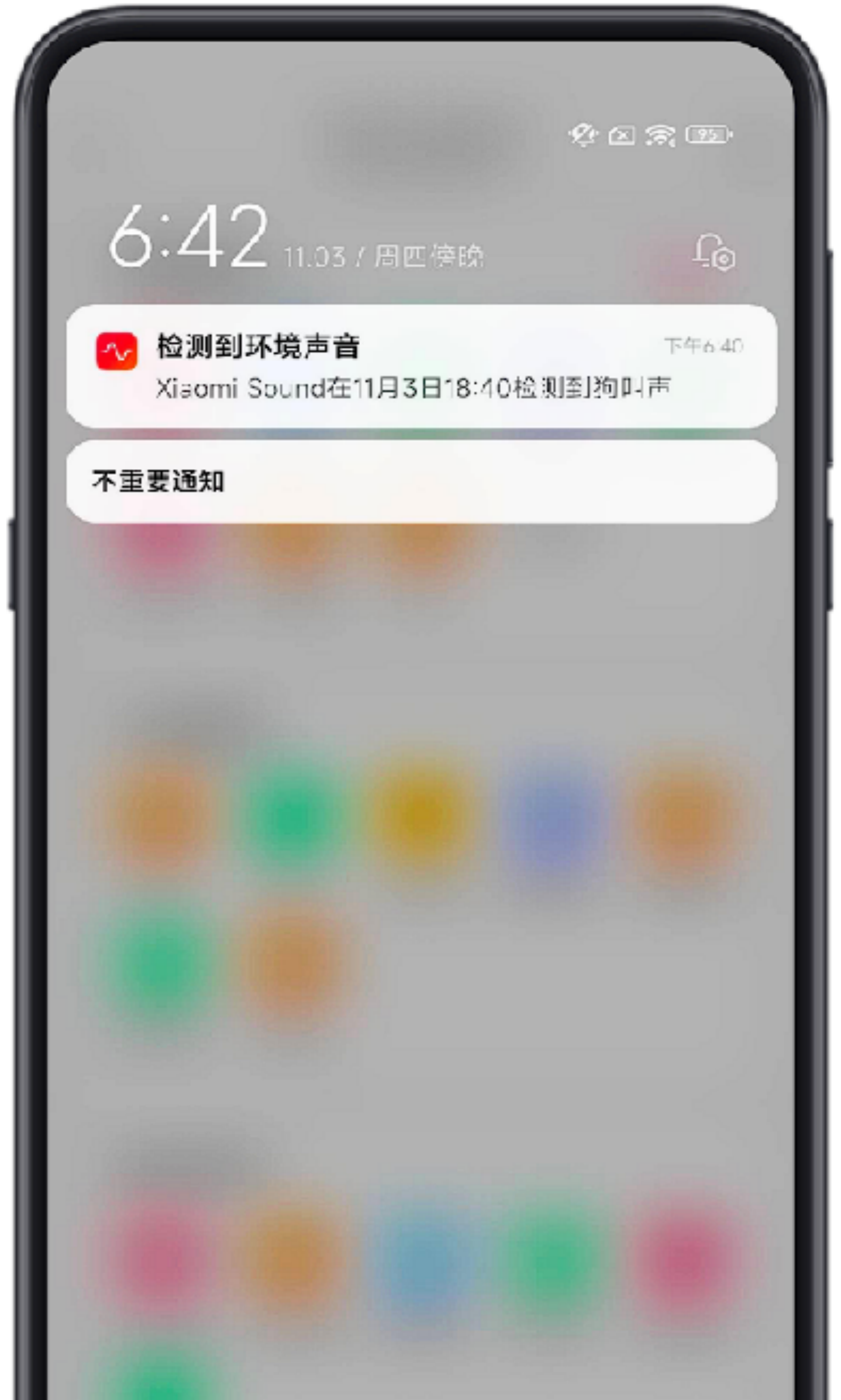
让婴儿的啼哭被爸爸妈妈及时“看到”

让宠物的叫声被主人“看到”

让敲门的声音被里屋的主人“看到”

……

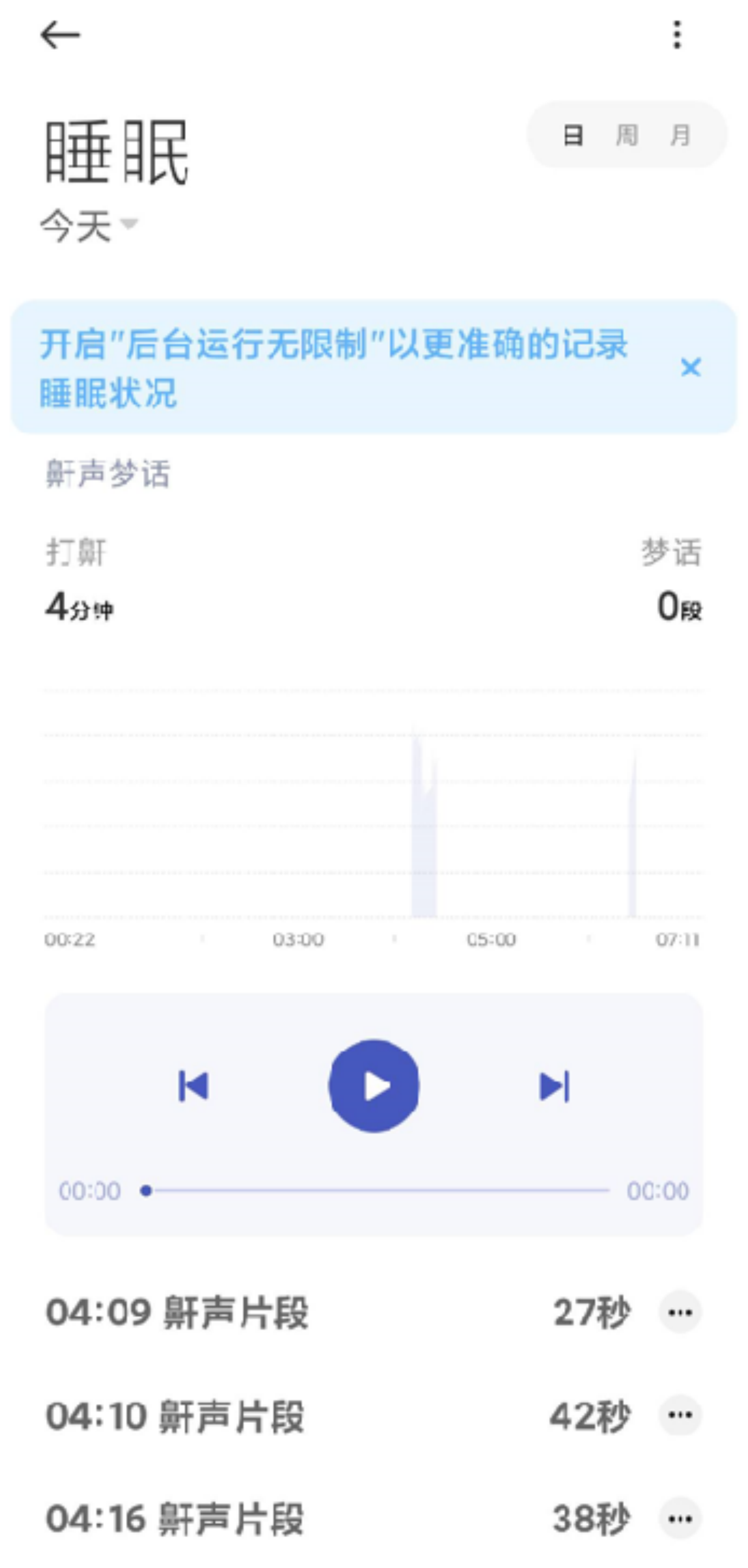
让声音，被“看到”




小米闻声

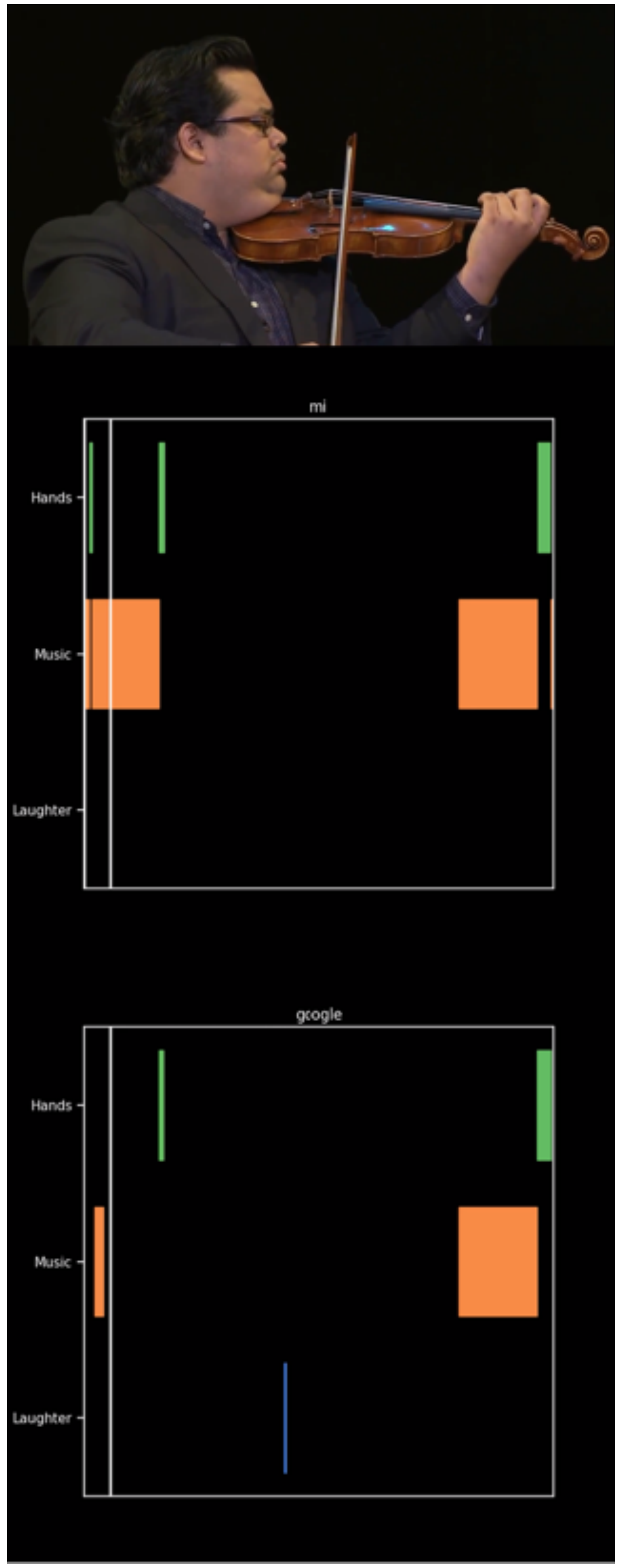

声音类别
报警
家用报警器（烟雾、燃气）
警笛
火警
住宅
婴儿啼哭
敲门
门铃
流水
趣味
猫叫
狗叫


10:52
环境音
您的音箱可识别特定类别的声音并通过小爱音箱App推送通知您，以此满足家庭异常情况监控、老年人安全守护、特殊人群无障碍辅助的需求。
Xiaomi Sound
家用报警器
婴儿啼哭
火警
流水
猫叫
狗叫
请注意，在可能导致您受伤或位于高风险环境时，不应完全依赖此功能。
注："家用报警器"声音包含烟雾报警器、燃气报警器及其他家电运行中可能发出的哔哔声，音箱检测到此类声音，会按"家用报警器"类别进行通知推送。

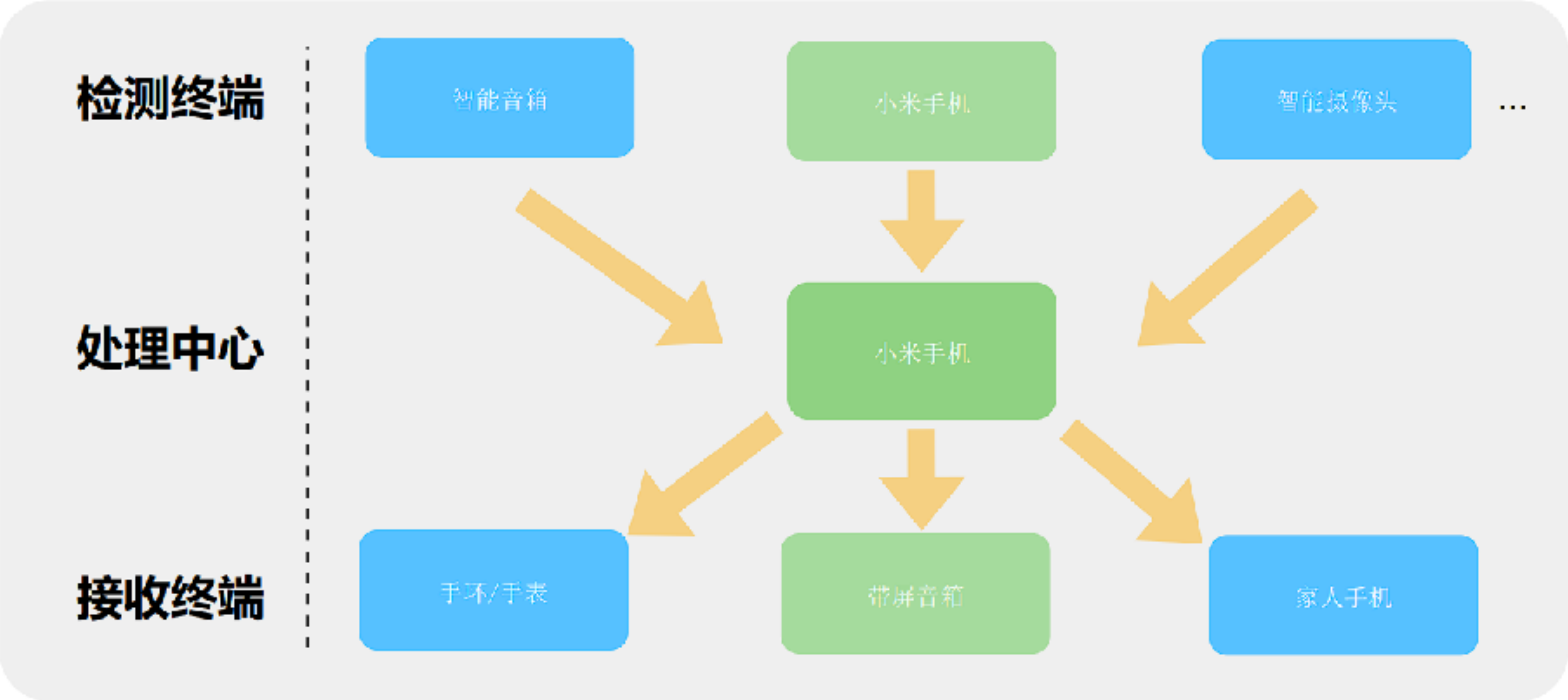

睡眠
日 周 月
今天
开启"后台运行无限制"以更准确的记录睡眠状况
鼾声梦话
打鼾
4分钟
梦话
0段
00:22
03:00
05:00
07:11
00:00
00:00
04:09 鼾声片段 27秒
04:10 鼾声片段 42秒
04:16 鼾声片段 38秒

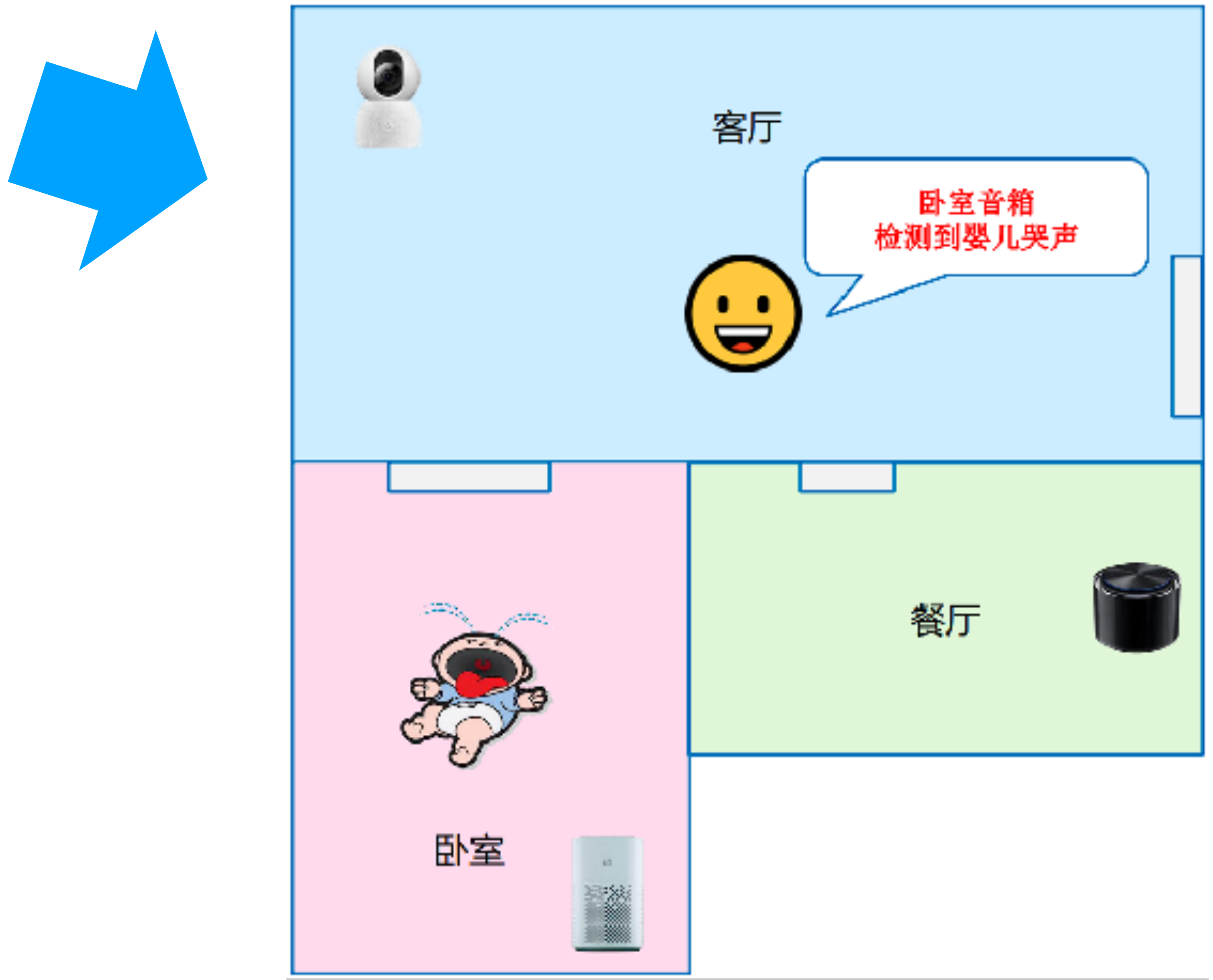

健康


(Laughter)
1:32 / 19:04 • Why were we so rowdy
The power of introverts | Susan Cain
TED
23M subscribers
Subscribe
393K
Share
Download
Clip
Save


mi
Hands
Music
Laughter
google
Hands
Music
Laughter

多端协同





分层通知，减少打扰

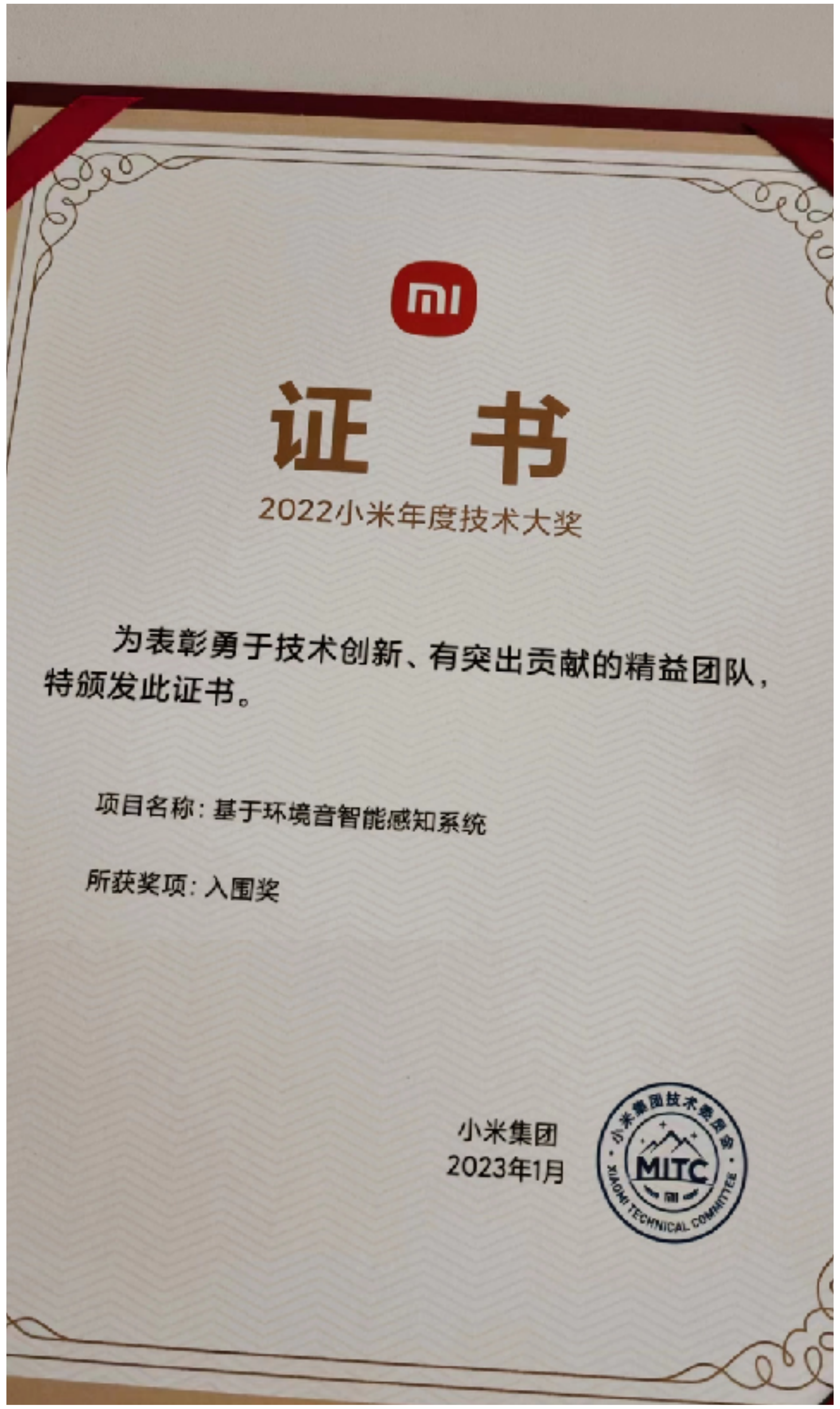
当同一房间内多设备检测到相同声音时，则只下发一条通知

当不同房间多设备检测到多次相同声音时，则多次下发多条通知

重点声音，高优通知

婴儿啼哭、烟雾警报等重要声音，迅速、多次、重点通知

宠物叫声、水流声，可在发生多次后再通知


证 书

2022小米年度技术大奖

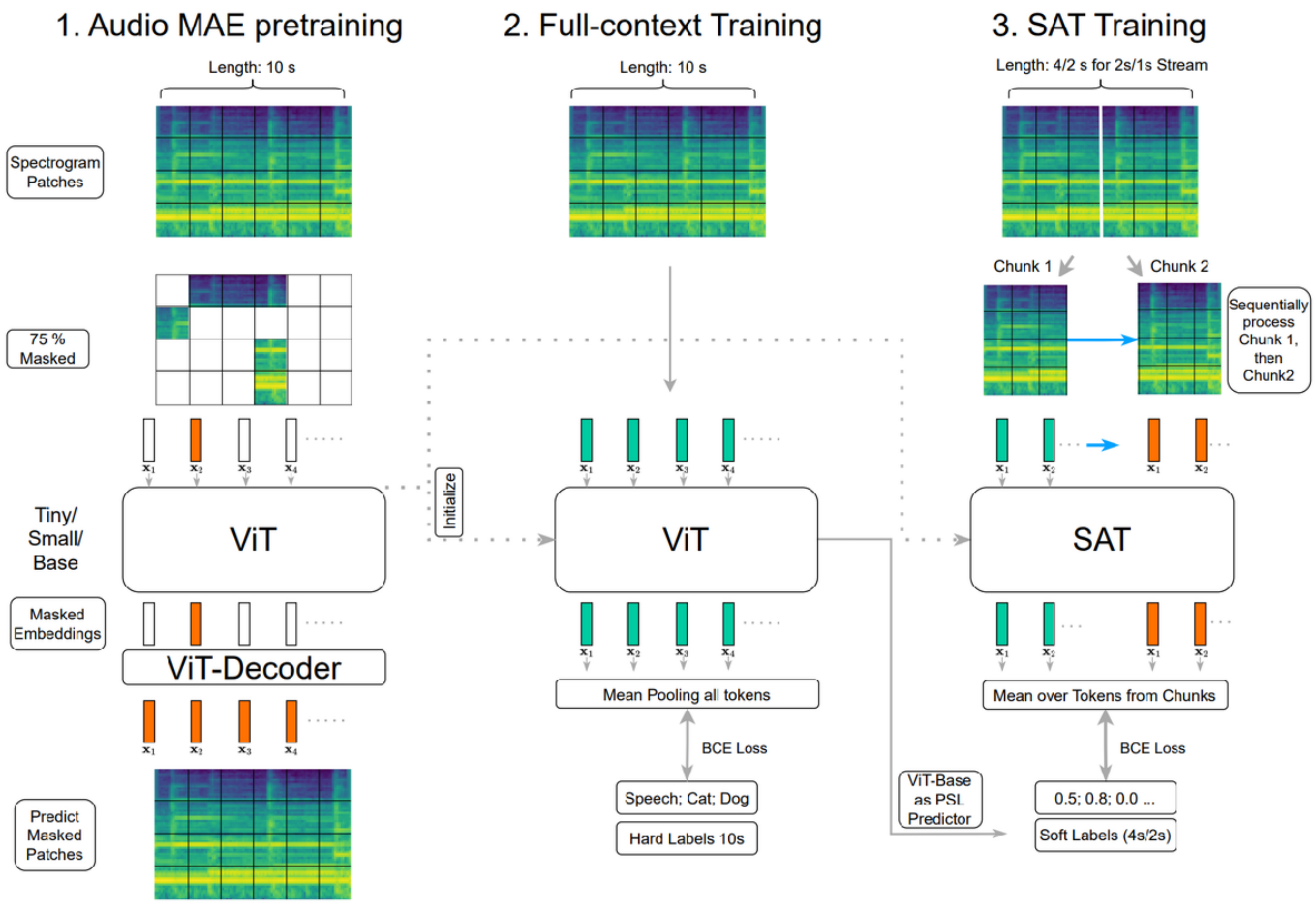
为表彰勇于技术创新、有突出贡献的精益团队，特颁发此证书。

项目名称：基于环境音智能感知系统

所获奖项：入围奖

小米集团
2023年1月

模型训练流程



Streaming Audio Transformers for Online Audio Tagging

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang

Xiaomi Corporation, Beijing, China

arXiv:2305.17834v1 [cs.SD] 29 May 2023

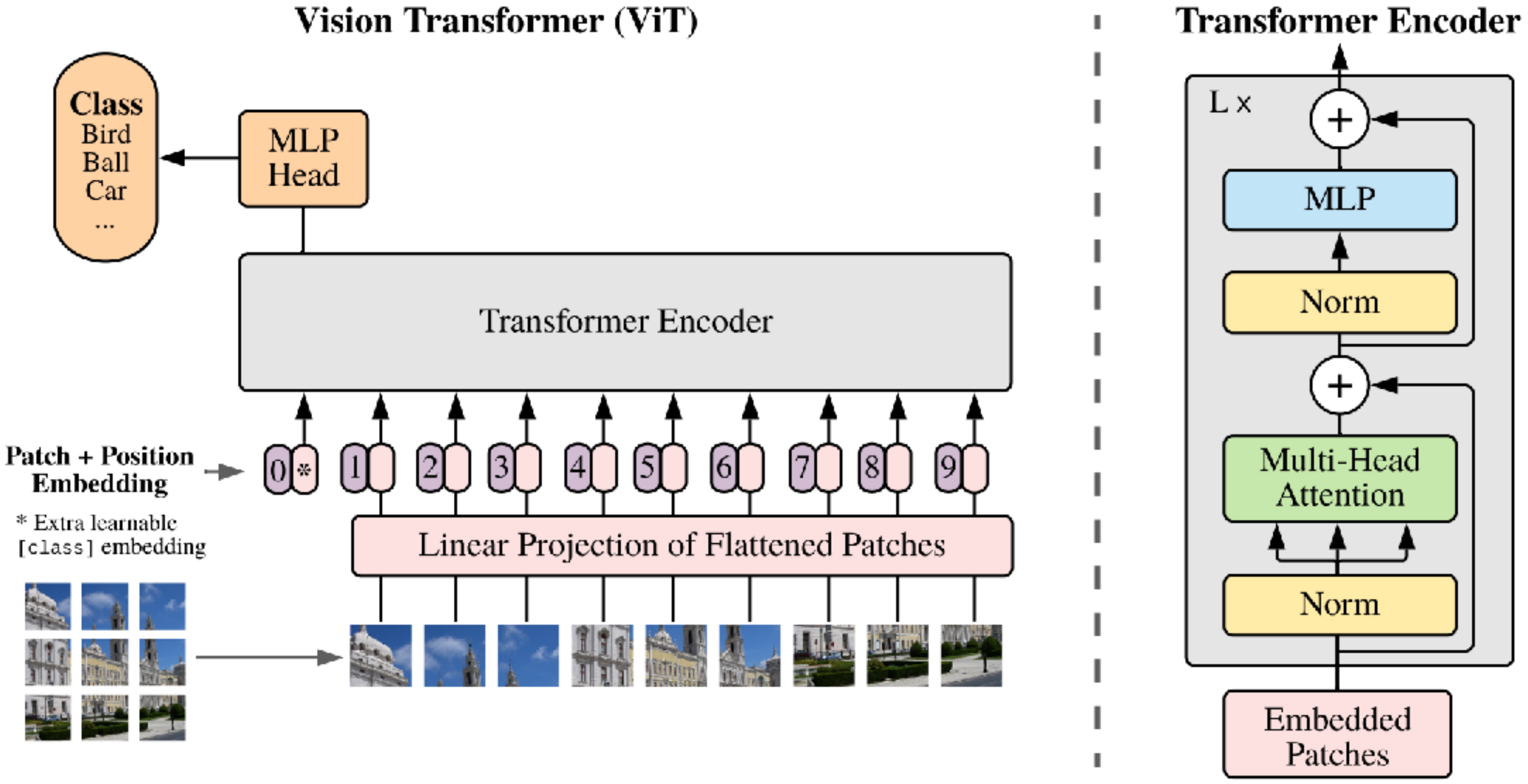
Abstract

1. Introduction

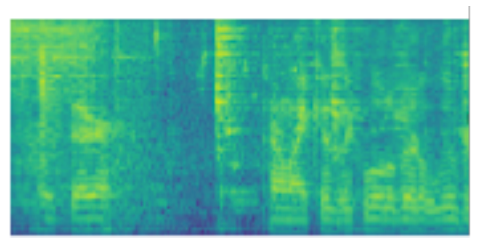
2. Vision Transformers for Audio Tagging

https://arxiv.org/abs/2305.17834

模型结构 - Vision Transformer (ViT)



图片

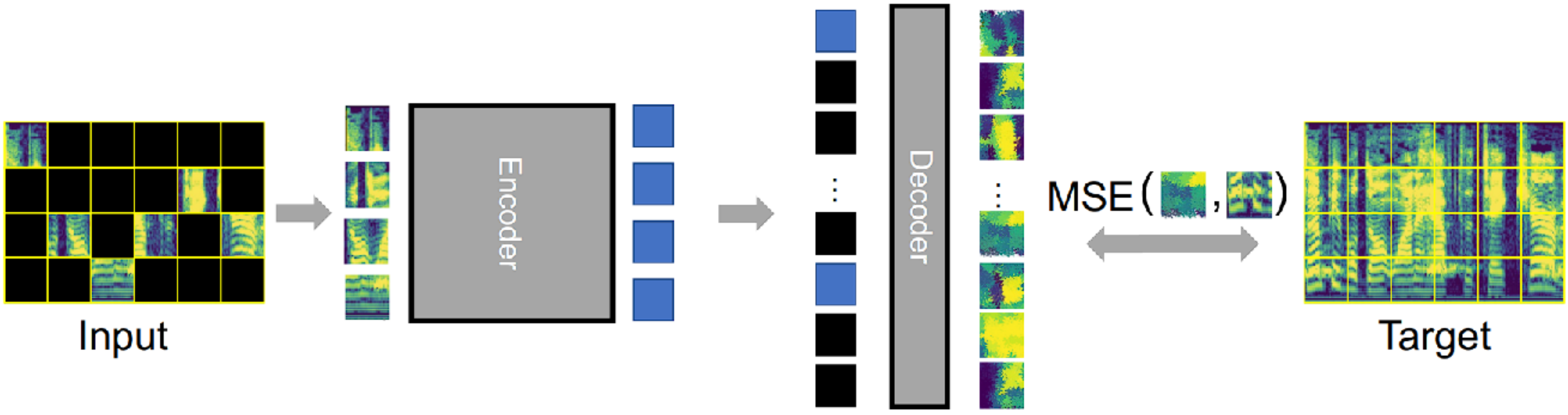


声音

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arxiv.org)

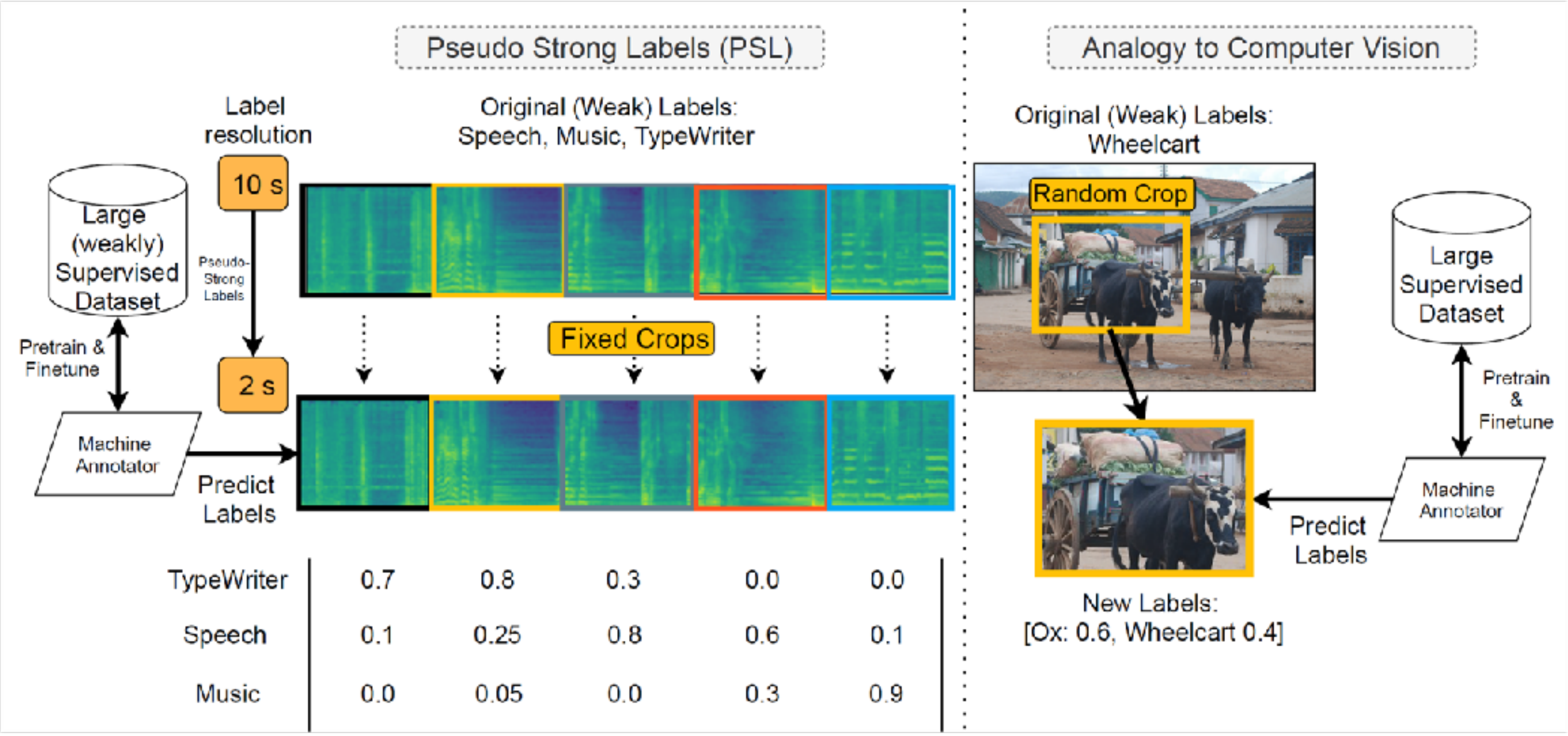
自监督训练 – Audio MAE



已收集时长约 31 年的训练数据

[2207.06405] Masked Autoencoders that Listen (arxiv.org)

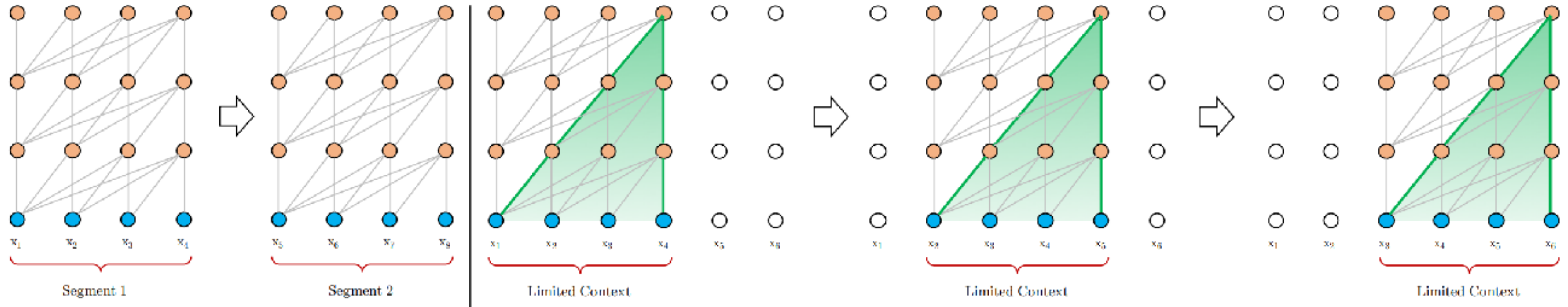
伪强标签生成 - PSL



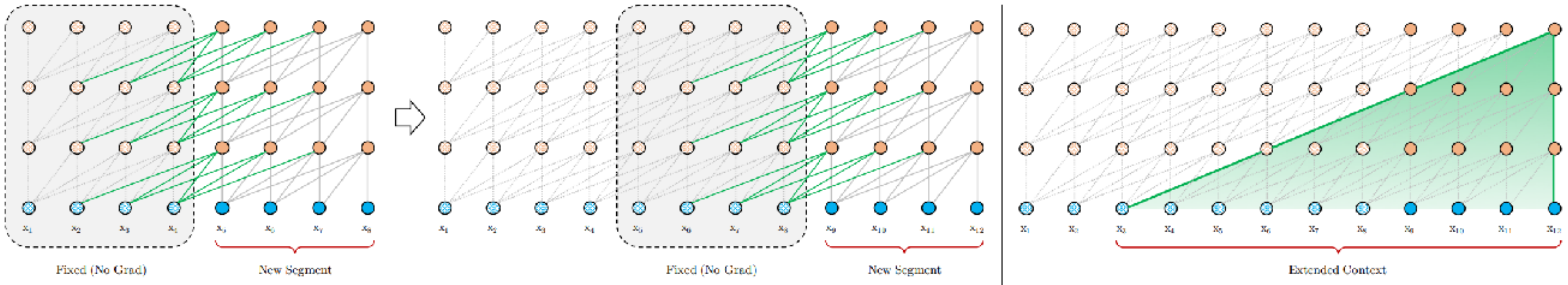
	AST-10s	AST-2s	(Ours)
mAP	**45.9**	39.7	44.2
参数量	86 M	86 M	**< 6 M**
峰值内存	2.2 G	2.2 G	**52 M**
延迟	10 s	**2 s**	**2 s**
推理速度 (小米10至尊版)	> 200 ms	> 200 ms	**~ 2 ms**

* [2104.01778] AST: Audio Spectrogram Transformer (arxiv.org)

流式训练



(a) Train phase. (b) Evaluation phase.



(a) Training phase. (b) Evaluation phase.

[1901.02860] Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context (arxiv.org)

自定义声音识别

使用 iPhone 识别声音

iPhone 可以持续听取某些声音（如婴儿哭声、门铃或汽笛声）并在识别出这些声音时通知你。

【注】当你可能受到伤害或受伤、在高风险或紧急情况下或者导航时，不要依赖 iPhone 识别声音。

设置声音识别

1. 前往"设置" >"辅助功能">"声音识别"，然后打开"声音识别"。
2. 轻点"声音"，然后打开想要 iPhone 识别的声音。

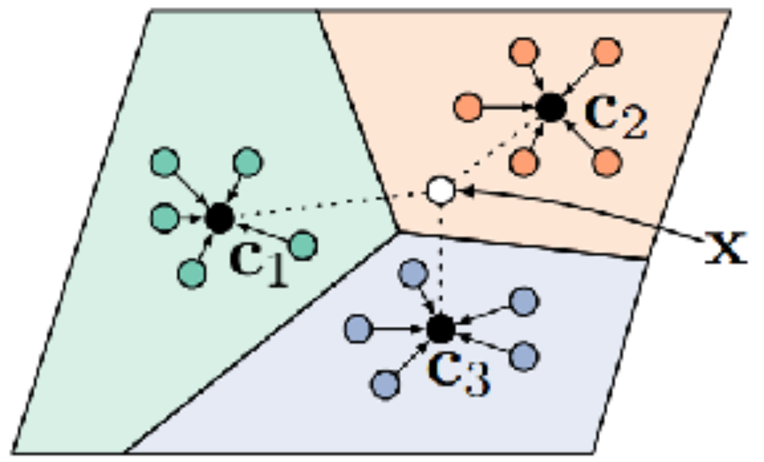
【提示】若要快速打开或关闭"声音识别"，请使用控制中心。

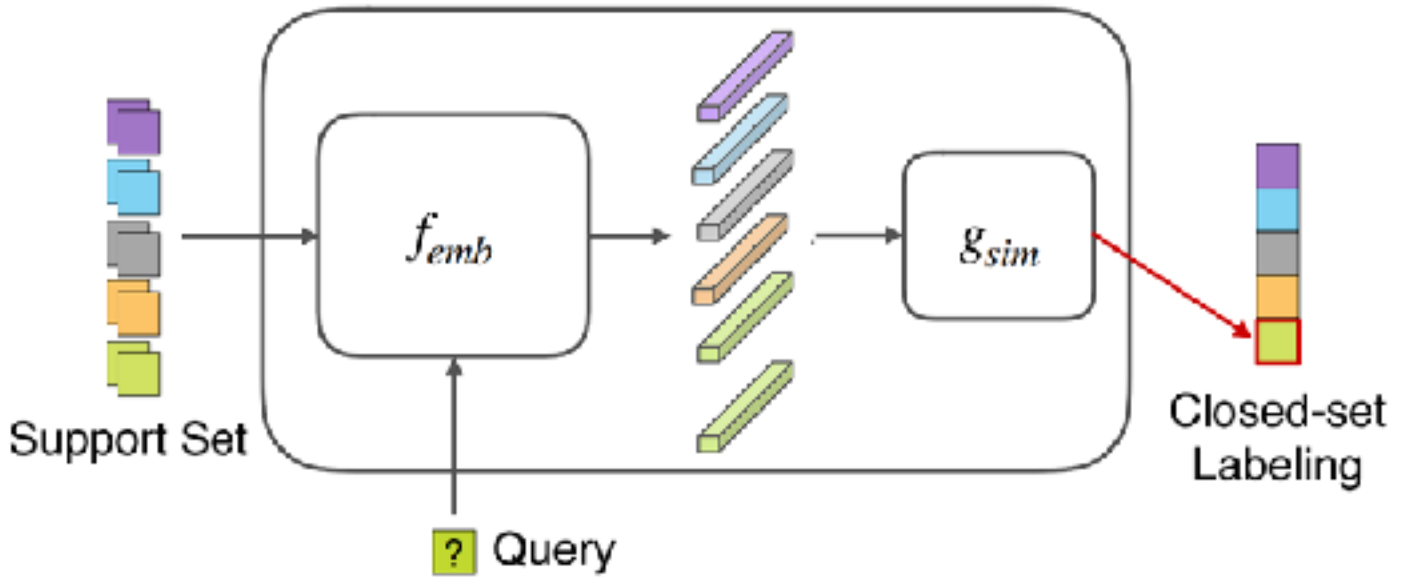
添加自定警报器、家电声或门铃声

如果声音未自动识别，你还可以设置 iPhone 以识别自定的警报器、家电声或门铃声。

1. 前往"设置" >"辅助功能">"声音识别">"声音"。
2. 轻点"自定警报器"或"自定家电声或门铃声"，然后输入一个名称。
3. 警报器、家电或门铃准备就绪后，将 iPhone 靠近声源并尽量减少背景噪声。
4. 轻点"开始听取"，然后按照屏幕指示操作。

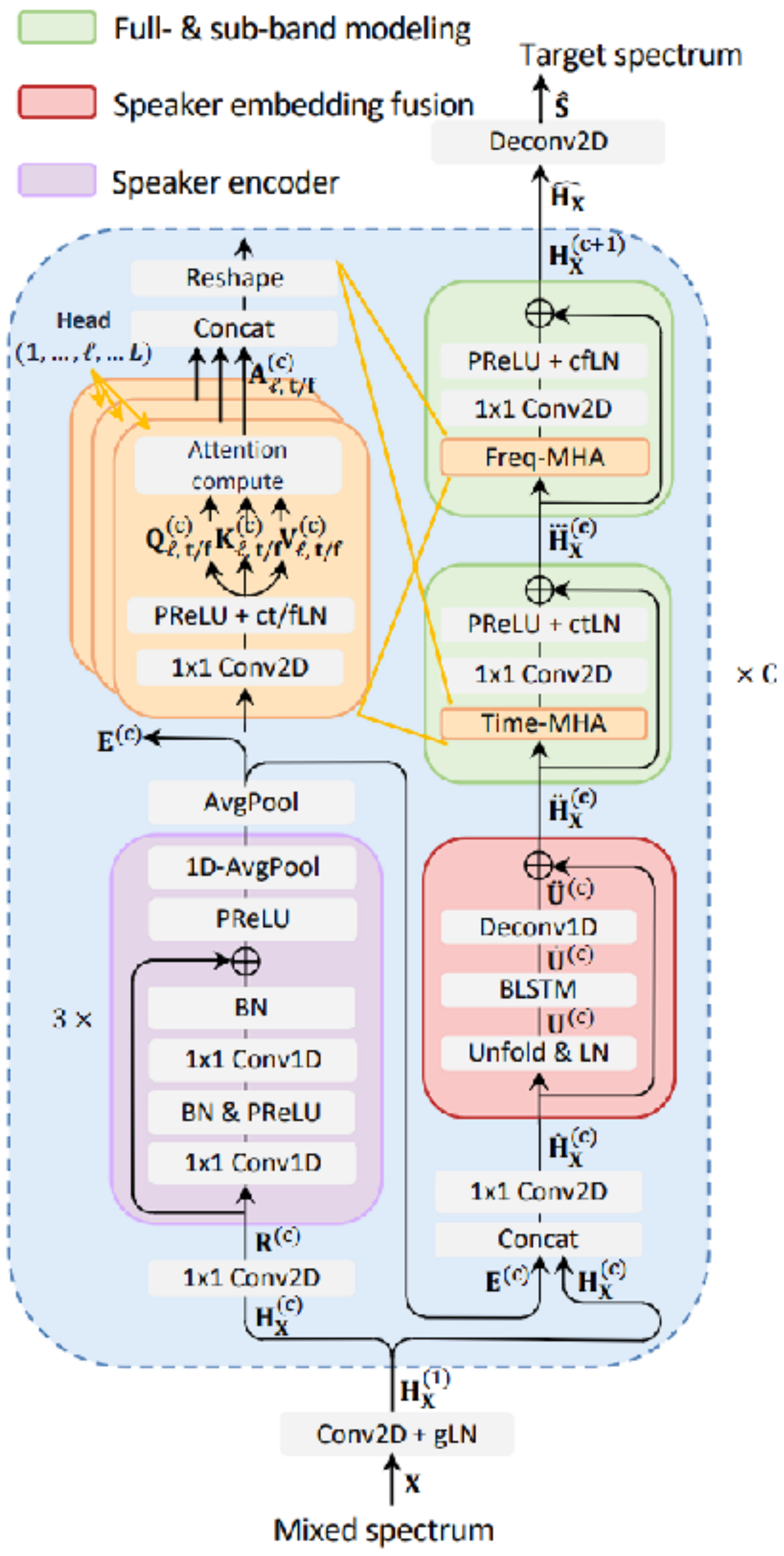
Few-shot Learning





Few-Shot Sound Event Detection | IEEE Conference Publication | IEEE Xplore

近距离干扰下的目标语音提取



INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

Focus the Sound around You: Monaural Target Speaker Extraction via Distance and Speaker Information

Jiaxin Lin[1,*], Peng Wang[2,*], Heinrich Dinkel[2], Jun Chen[1], Zhiyong Wu[1,†], Yongqing Wang[2], Zhiyong Yan[2], Junbo Zhang[2], Yujun Wang[2]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Xiaomi Inc., Beijing, China

Abstract

1. Introduction

近期发表论文

- A Lightweight Approach for Semi-supervised Sound Event Detection with Unsupervised Data Augmentation
- A Contrastive Semi-Supervised Learning Framework For Anomaly Sound Detection

- An Empirical Study of Weakly Supervised Audio Tagging Embeddings for General Audio Representations

- Pseudo Strong Labels For Large Scale Weakly Supervised Audio Tagging

- UniKW-AT: Unified Keyword Spotting and Audio Tagging

- Unified Keyword Spotting and Audio Tagging on Mobile Devices with Transformers

- Focus the Sound around You: Monaural Target Speaker Extraction via Distance and Speaker Information

思考和结语

- 技术去帮助障碍人群的同时，无障碍也提供了对技术的一个极致测试场景
- 解决障碍人士日常的需求的同时，也解决了普通人类似不便场景的需求
- 不便场景往往给预研探索提供了落地空间

谢谢！


小米集团技术委员会
MITC
MI
XIAOMI TECHNICAL COMMITTEE