北京
2019

遨游“视”界 做你所想
Explore World, Do What You Want

面向家居场景的远场语音增强技术

小米AI实验室　相非

出品：LiveVideoStack 音视频技术社区　CSDN

相非

小米AI实验室技术总监

中国声学学会理事兼声频工程分会委员

中国计算机学会语音对话与听觉专业组委员

北京邮电大学硕士，从事语音、音频领域研发和管理工作19年

曾在摩托罗拉任职9年，负责语音增强算法、3D音效引擎、语音对讲、VoIP等项目的研发工作，完成多款上市手机的音频架构设计

2017年2月从零组建小米远场声学团队，致力于人工智能声学和语音增强领域的技术研究，六麦阵列、就近唤醒等算法已上线小爱同学系列产品

语音技术发展历史

语音增强技术演进

家居场景技术挑战

远场增强技术展望

语音技术发展历史

语音增强技术演进

家居场景技术挑战

远场增强技术展望

语音技术发展历史

贝尔发明电话机

1876

1877

爱迪生发明留声机

无线电话问世

1943

1973

摩托罗拉推出第一部民用手机

苹果发布Siri

2007

2011

微信语音对讲上线

亚马逊推出ECHO

2014

语音技术发展历史

语音增强技术演进

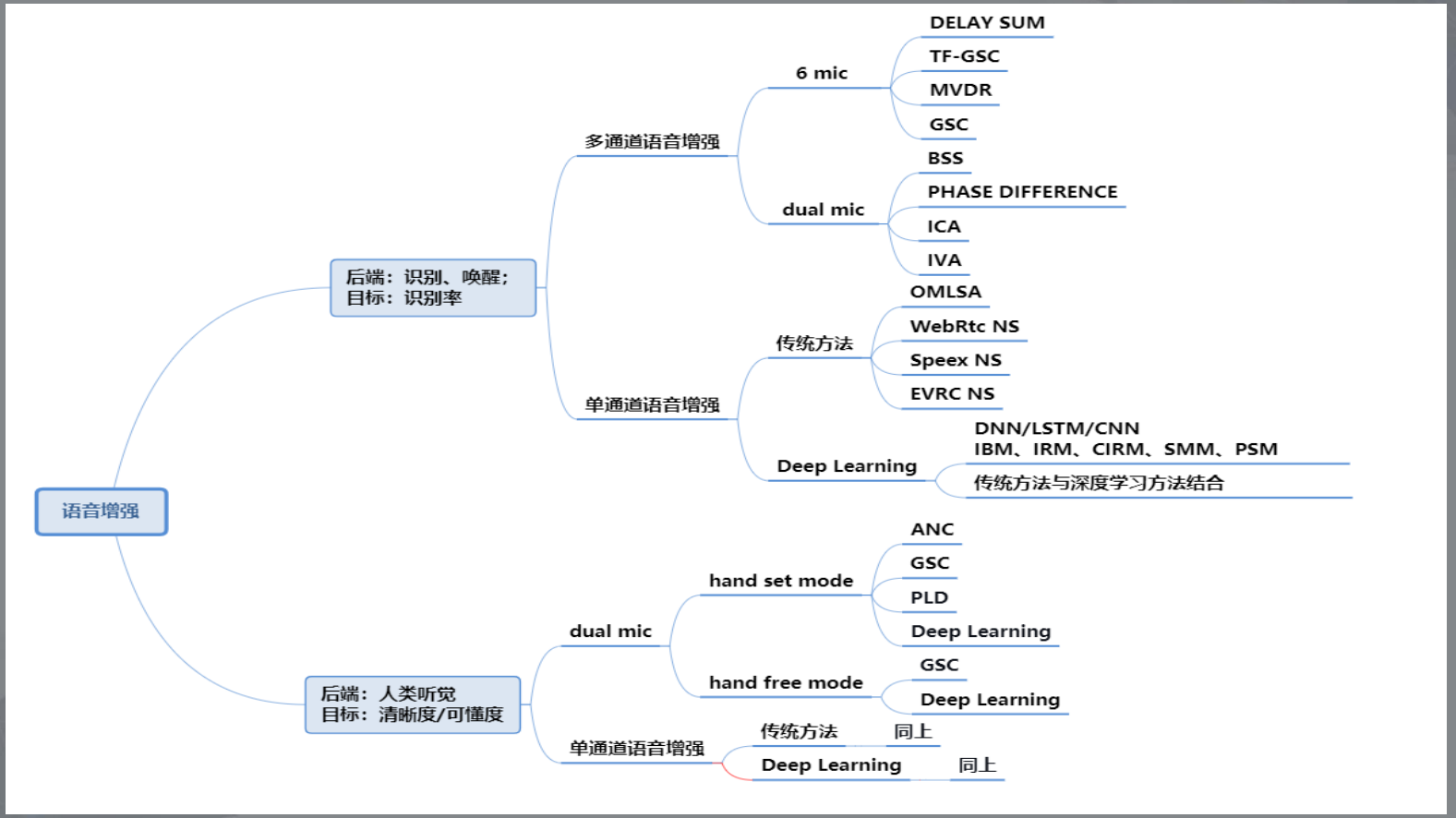
家居场景技术挑战

远场增强技术展望

语音增强技术演进

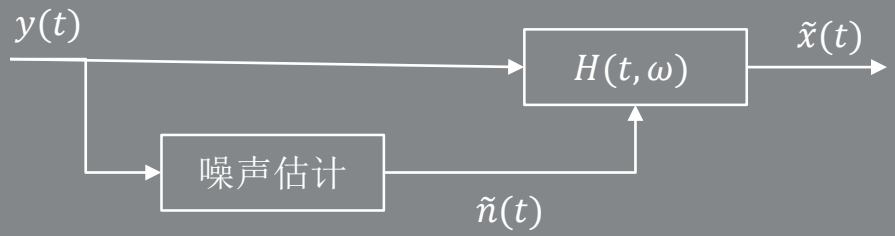
应用场景	语音通话	→	人机交互	
拾音距离	近场主导	→	远场+近场	
技术演进	单通道	→	多通道	→ 深度学习

语音增强算法概述


语音增强
后端：识别、唤醒；
目标：识别率
多通道语音增强
6 mic
DELAY SUM
TF-GSC
MVDR
GSC
dual mic
BSS
PHASE DIFFERENCE
ICA
IVA
单通道语音增强
传统方法
OMLSA
WebRtc NS
Speex NS
EVRC NS
Deep Learning
DNN/LSTM/CNN
IBM、IRM、CIRM、SMM、PSM
传统方法与深度学习方法结合
后端：人类听觉
目标：清晰度/可懂度
dual mic
hand set mode
ANC
GSC
PLD
Deep Learning
hand free mode
GSC
Deep Learning
单通道语音增强
传统方法
同上
Deep Learning
同上

- 基于信号的时/频域特征，估计带噪信号中的噪声特征，从而获得增强语音；

$$y(t) = s(t) + n(t) \quad \longrightarrow \quad \tilde{x}(t) = y(t) - \tilde{n}(t) = x(t) + e$$

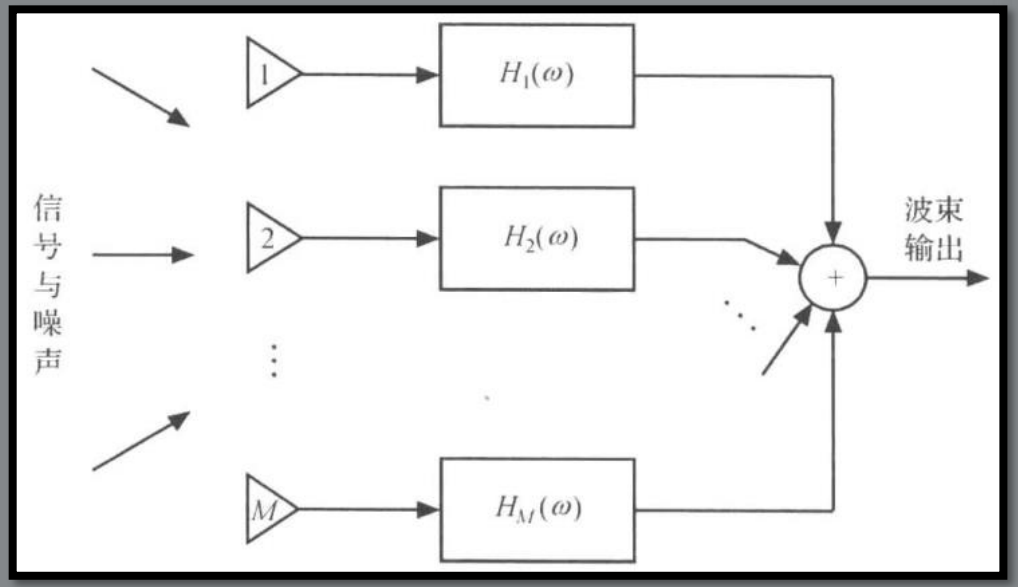


- 经典算法

OM-LSA / WebRtc NS / Speex NS / EVRC NS

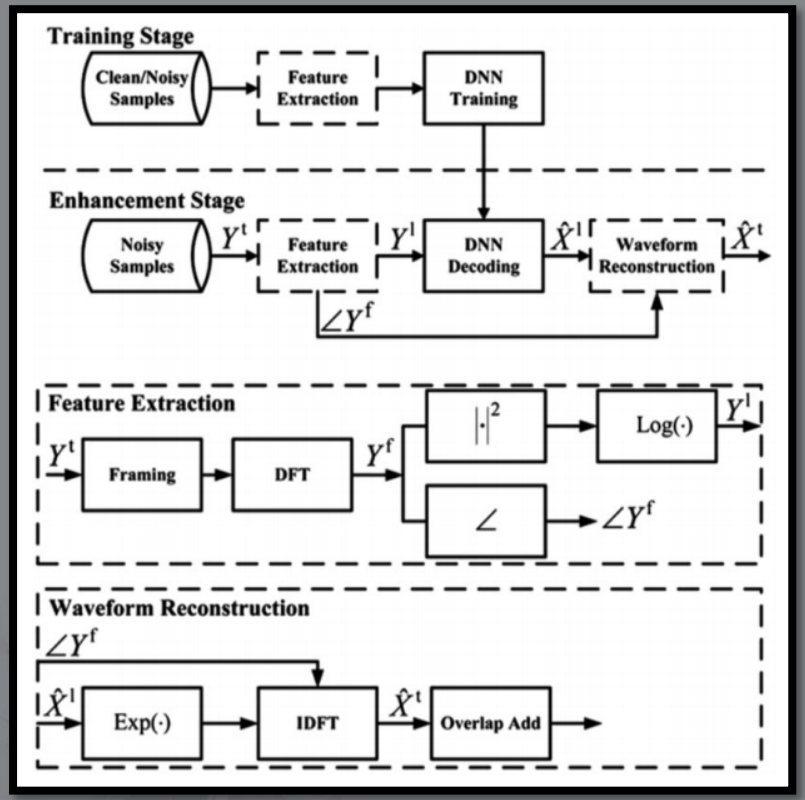
传统多通道语音增强算法

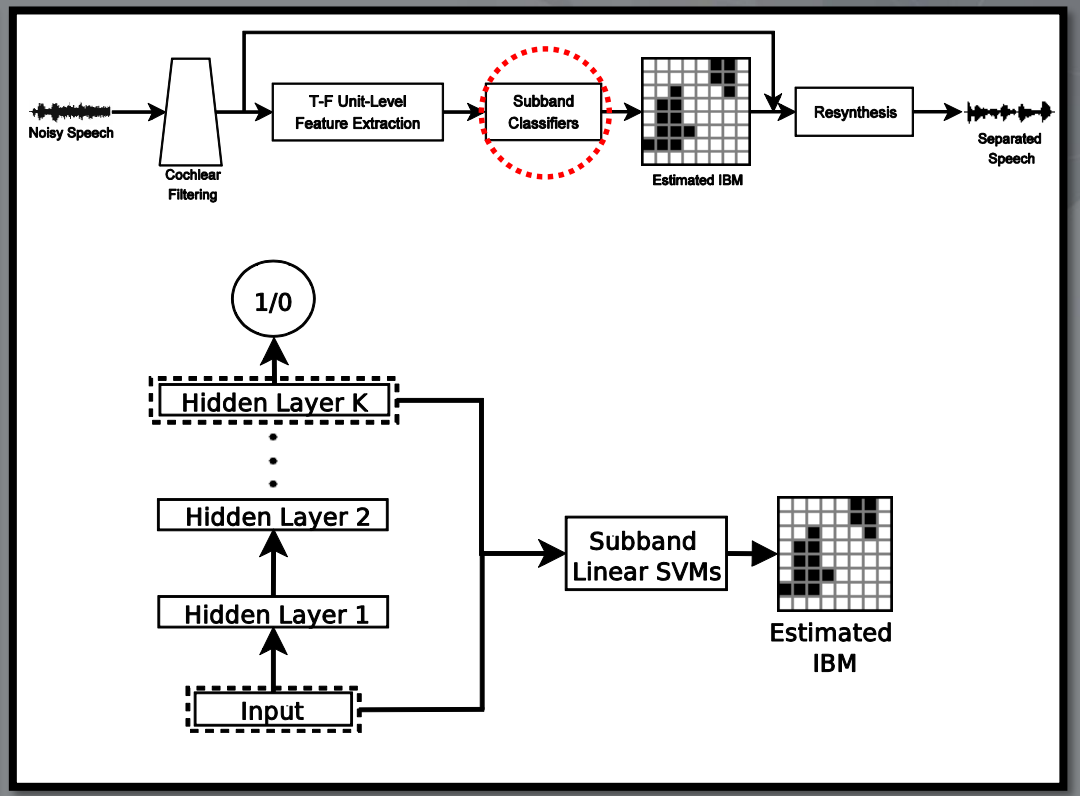
- 通过对各阵元采集数据进行线性时不变滤波再求和，得到波束输出：



- 经典算法： Delay-Sum MVDR

基于深度学习的单通道增强





谱映射

MASK

基于深度学习的多通道增强

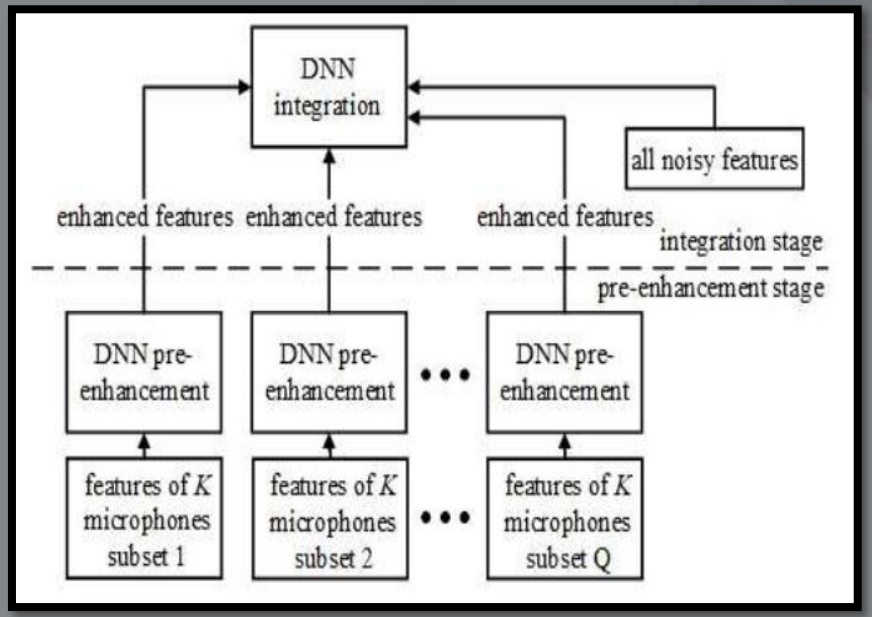
基于二阶模型的多通道语音增强

Stage 1: 预增强

- 多个独立的DNN模型；
- 利用K个麦克风信号获得增强信号；

Stage 2: 特征融合

- 预增强特征+带噪特征；



主流语音增强技术总结

	传统单通道增强	传统多通道增强	深度学习增强
优势	平稳噪声性能好 鲁棒性高 不依赖数据 运算量低	空间信息 线性滤波	性能优于传统算法 非平稳噪声性能好
劣势	非平稳噪声性能差	依赖阵列拓扑结构 同向噪声抑制弱 运算量略大	数据依赖 运算量大 模型大
典型算法	OMLSA WebRTC_NS Speex_NS EVRC_NS	Delay-Sum GSC MVDR	DNN/CNN/RNN 谱映射/MASK

语音样例

带噪语音

WebRTC

OMLSA

RNN

DNN

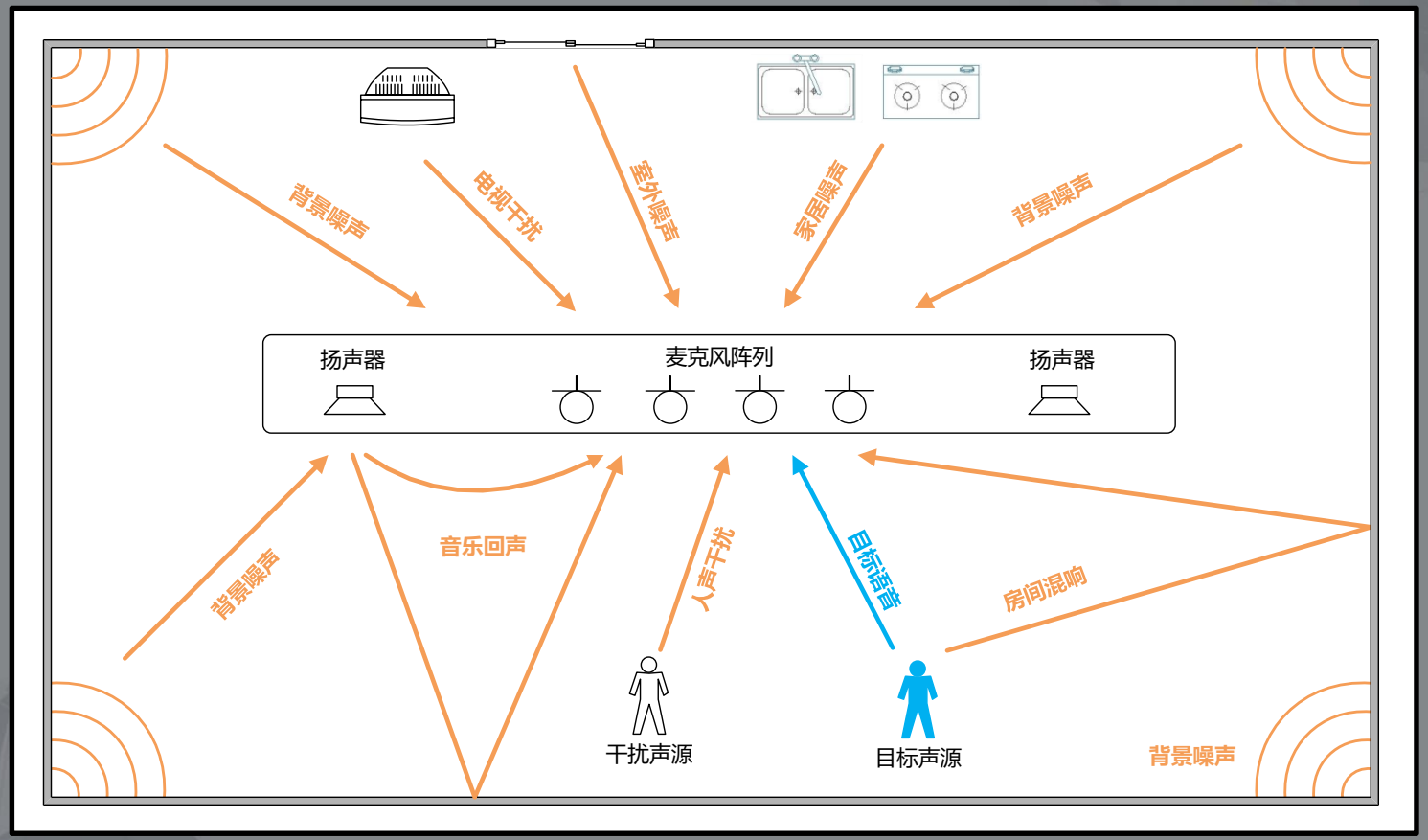
语音技术发展历史

语音增强技术演进

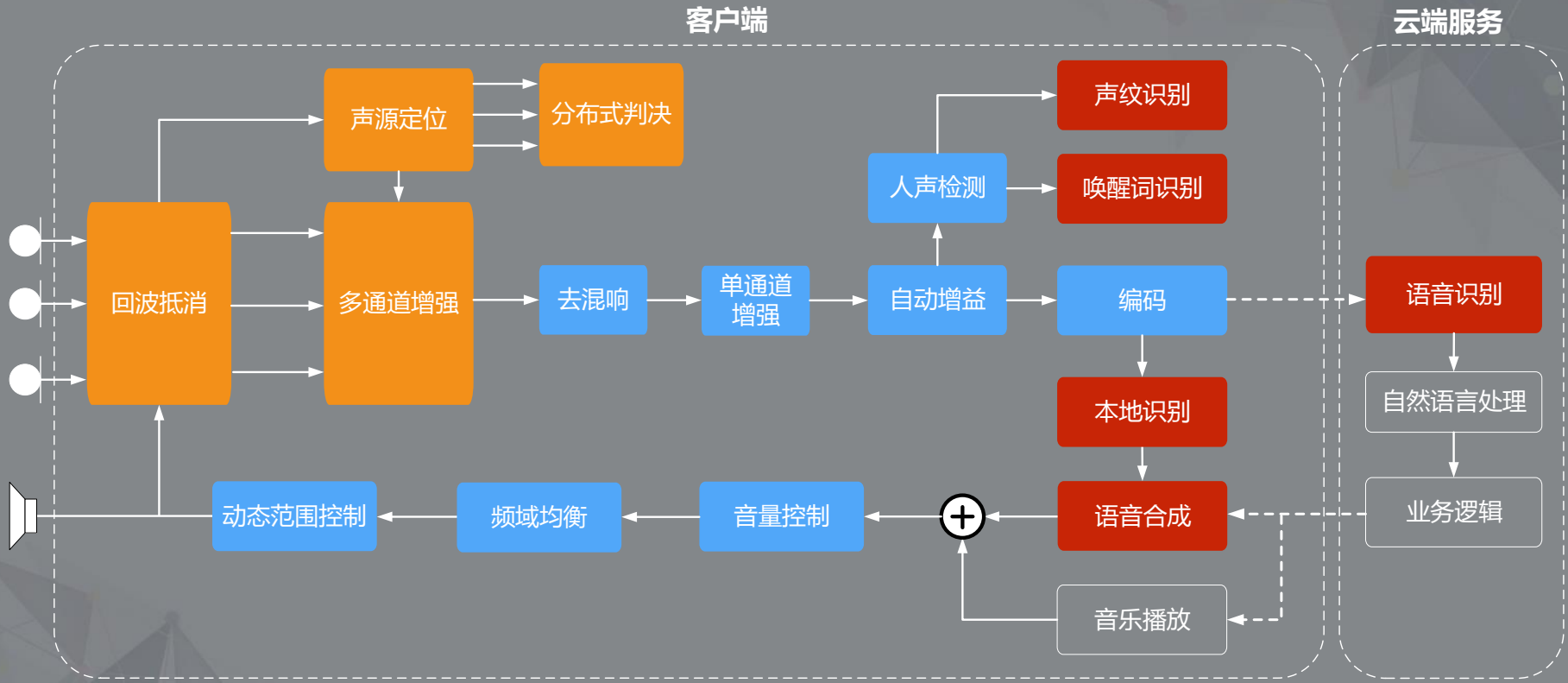
家居场景技术挑战

远场增强技术展望

家居场景的复杂声场环境



智能语音技术架构


客户端
云端服务
声源定位
分布式判决
回波抵消
多通道增强
去混响
单通道增强
自动增益
人声检测
声纹识别
唤醒词识别
编码
本地识别
语音合成
音乐播放
语音识别
自然语言处理
业务逻辑
动态范围控制
频域均衡
音量控制
多通道前端算法
单通道前端算法
识别及合成算法

远场增强落地难点

家居场景下的问题：

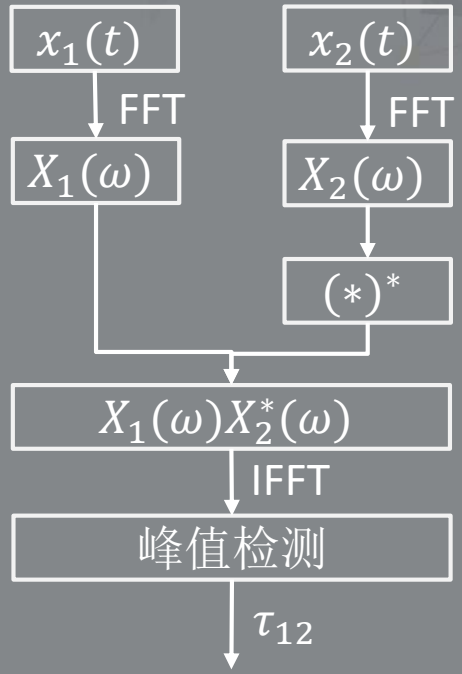
- 强干扰下，波达方向和噪声阵估计失准；
- 语音与干扰同向，增强性能下降；
- 麦克风信号失真，不满足阵元一致性假设；

常规声源定位方法

- 采用基于广义互相关方法的SRP获得。

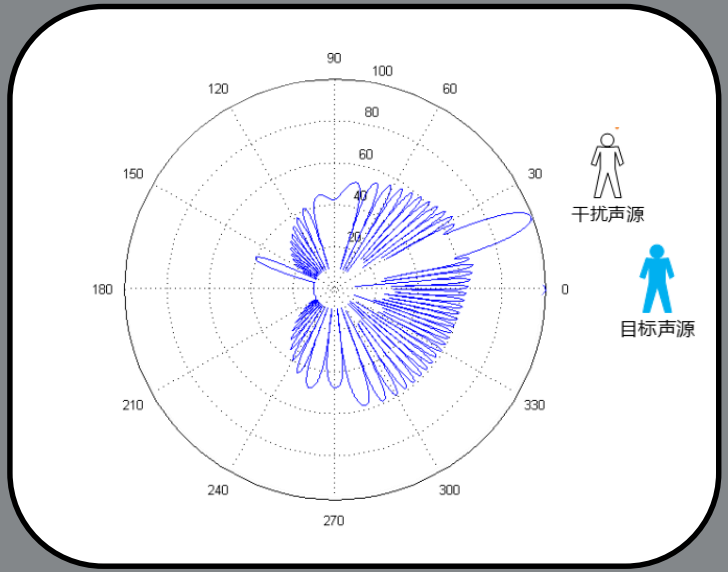
$$R^{l}_{m_1,m_2}(\tau)=\sum_{k=0}^{N-1}\frac{X^{l}_{m_1}[k]}{\left|X^{l}_{m_1}[k]\right|}\frac{X^{l}_{m_2}[k]^{*}}{\left|X^{l}_{m_2}[k]\right|}\,e^{\left(\frac{j2\pi k\tau}{N}\right)}$$

- ✓ 假想声源位置；
- ✓ 所有麦克风对接收信号的GCC-PHAT之和；
- ✓ SRP峰值点为估计位置；



波达方向失准的问题

- 强干扰源（电视）存在，会导致波达方向估计失准；
- 固定波束，导向矢量失准会导致阵列增益下降明显；



提升阵列增益

- 同时，常规波束形成，因旁瓣作用，会导致阵列增益有限；
- 采用信号自适应方法，根据噪声信号特征在干扰方向形成零陷；
- 保证期望方向增益不变的情况下，使得输出SINR最大。

$$SINR \triangleq \frac{E[|\boldsymbol{w}^H \boldsymbol{s}|^2]}{E[\boldsymbol{w}^H(\boldsymbol{i}+\boldsymbol{n})^2]} = \frac{\sigma_s^2 |\boldsymbol{w}^H \boldsymbol{\alpha}(\theta_s)|^2}{\boldsymbol{w}^H \boldsymbol{R}_{i+n} \boldsymbol{w}}$$

最优解:

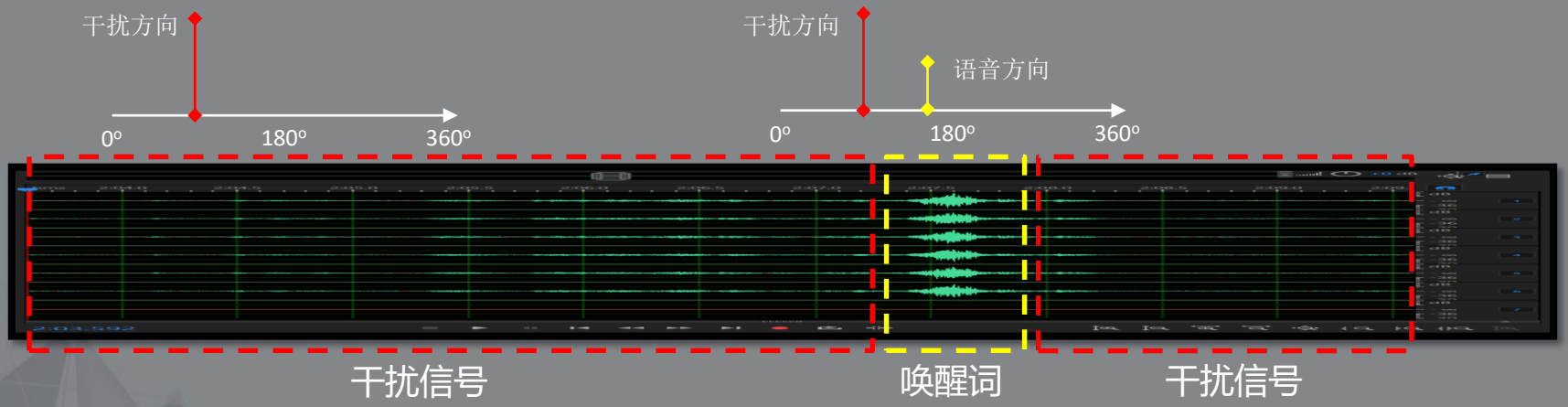
$$\boldsymbol{w}_{MVDR} = \frac{\boldsymbol{R}_{i+n}^{-1} \boldsymbol{\alpha}(\theta_s)}{\boldsymbol{\alpha}^H(\theta_s) \boldsymbol{R}_{i+n}^{-1} \boldsymbol{\alpha}(\theta_s)}$$

- 低信噪比下语音被干扰淹没；
- R_{nn}估计失准会导致语音损伤；

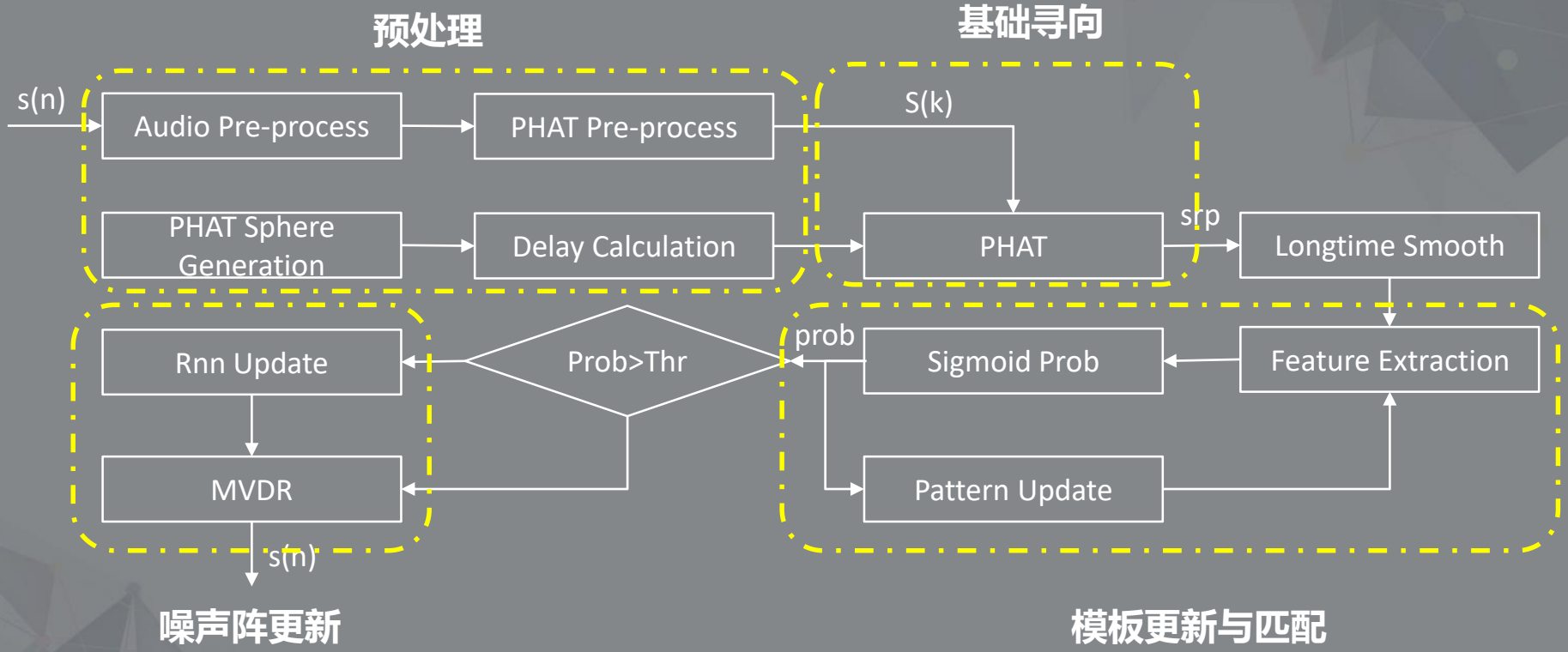
解决方案

- 基于唤醒词偶现和干扰源持续存在的假设；
- 检测唤醒词方向与干扰方向差异变化，修正波达方向和R_{nn}估计；

$$\rho^2 = cov^2(X,Y)/DX \cdot DY$$



基于SRP模板匹配追踪机制MVDR的前端增强算法



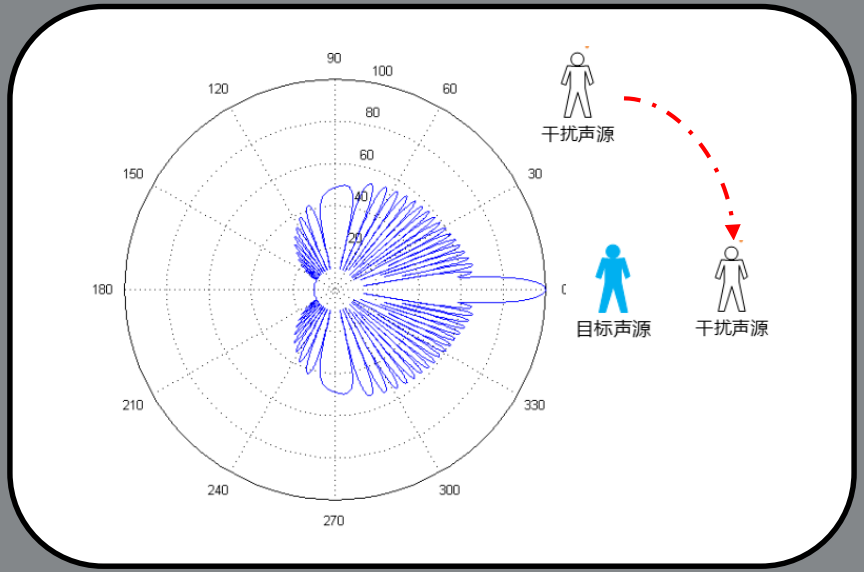
远场增强落地难点

家居场景下的问题：

- 强干扰下，波达方向和噪声阵估计失准；
- 语音与干扰同向，增强性能下降；
- 麦克风信号失真，不满足阵元一致性假设；

目标/干扰同向问题

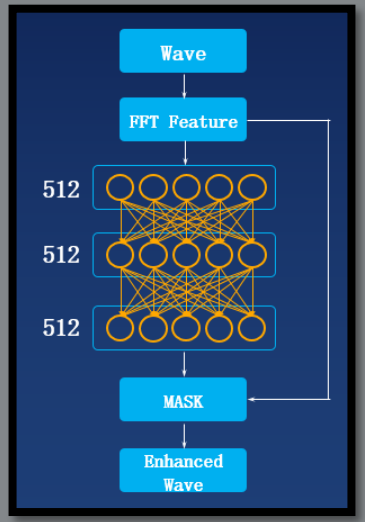
- 目标与干扰同向时，在空间上不具有区分度；
- 阵列增益对于目标和干扰源相同；

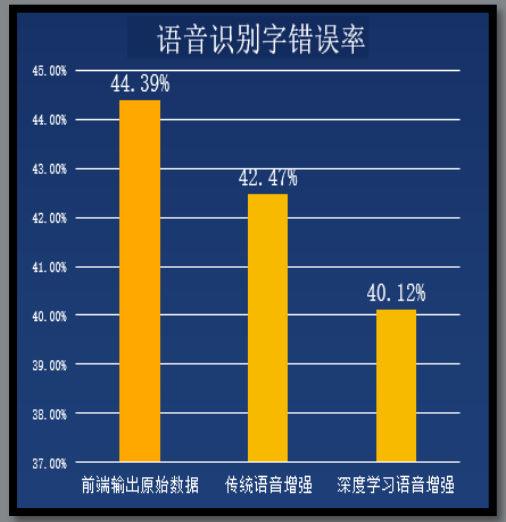


基于DNN的语音增强

- DNN增强能有效应对非平稳噪声；
- 采用单通道DNN增强算法，弥补目标与干扰同向时，阵列性能的下降。

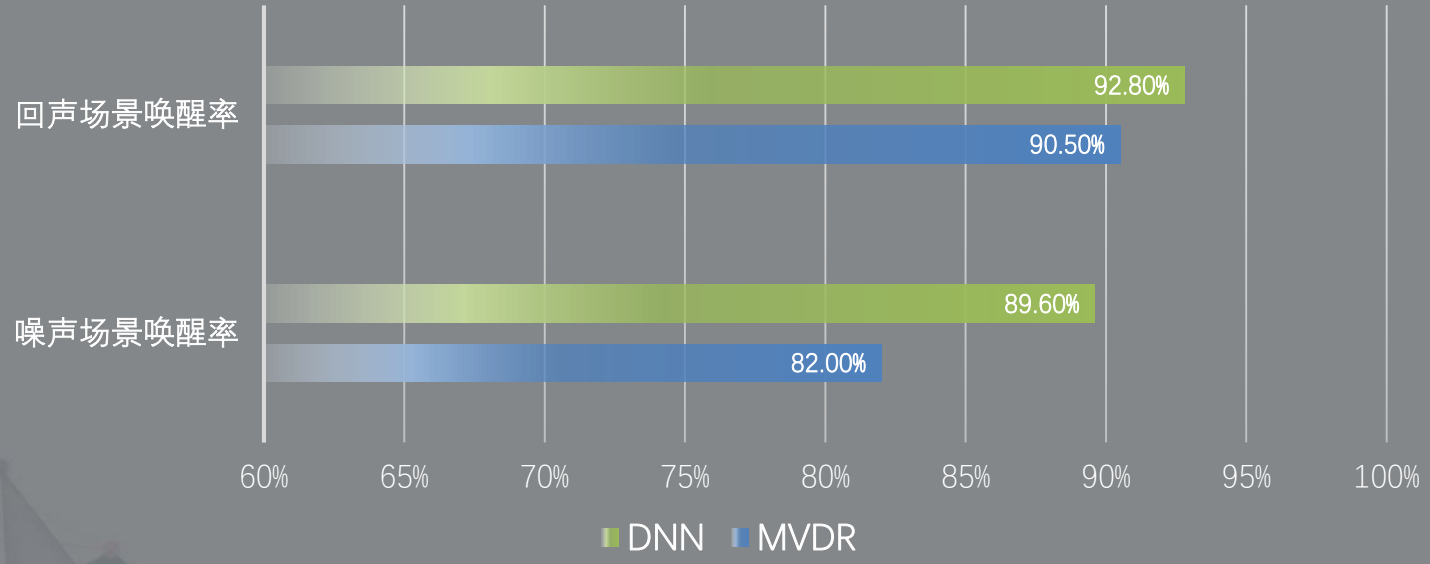
- 3层DNN神经网络模型；
- 单通道语音增强；





现阶段算法性能对比

- MVDR：MVDR算法+SRP模板匹配；
- DNN：MVDR+SRP模板匹配+DNN;



远场增强落地难点

家居场景下的问题：

- 强干扰下，波达方向和噪声阵估计失准；
- 语音与干扰同向，增强性能下降；
- 麦克风信号失真，不满足阵元一致性假设；

麦克风信号失真问题

问题：

- 麦克风被遗物遮挡；
- 器件老化/跌落导致内部密封性被破坏；

影响：

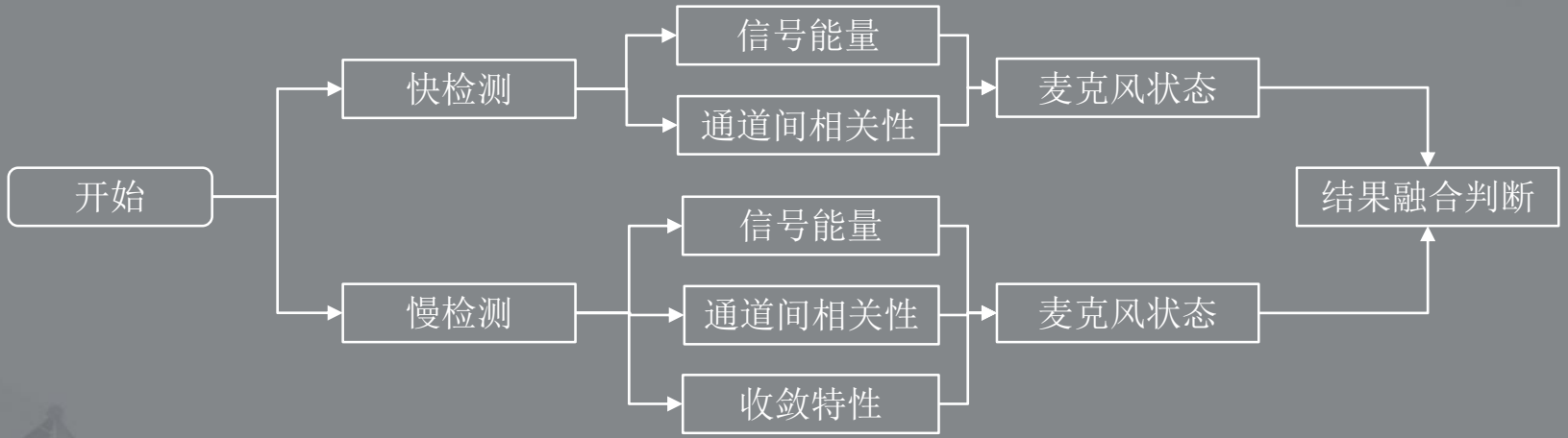
- 各麦克风之间拾音性能差异较大；
- 不满足阵列算法各阵元性能一致性假设；
- 阵列增强算法性能下降明显；

方案：

- 基于拾音信号，检测异常存在概率；
- 确认异常麦克风位置；
- 阵列算法降级，弥补性能劣势，提升用户体验。

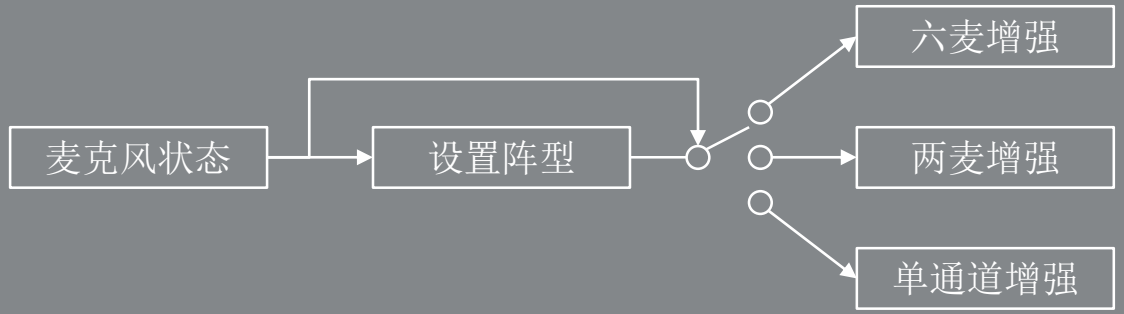
麦克风异常信号检测与补偿

信号异常检测



麦克风异常信号检测与补偿

依据检测结果的补偿策略



异常信号检测算法性能	
异常信号检测正确率	68.42%
正常信号检测正确率	100.00%

语音技术发展历史

语音增强技术演进

家居场景技术挑战

远场增强技术展望

远场语音增强技术展望

多声源

技术融合

端到端

智能语音应用场景


落地场景
智能家居
电视
音箱
净化器
儿童
插座
扫地机
移动互联网
手机
MIUI
金融
电商
广告
游戏
智能户外
耳机
平衡车
手环
无人机
智能车载
记录仪
净化器
核心技术
感知智能
认知智能
决策智能
服务聚合
音频
视频
设备控制
导航出行
资讯
医疗健康
O2O
金融
电商

北京
2019

遨游“视”界 做你所想
Explore World, Do What You Want

Thank you

出品：LiveVideoStack 音视频技术社区 CSDN