语音交互前端处理技术概览

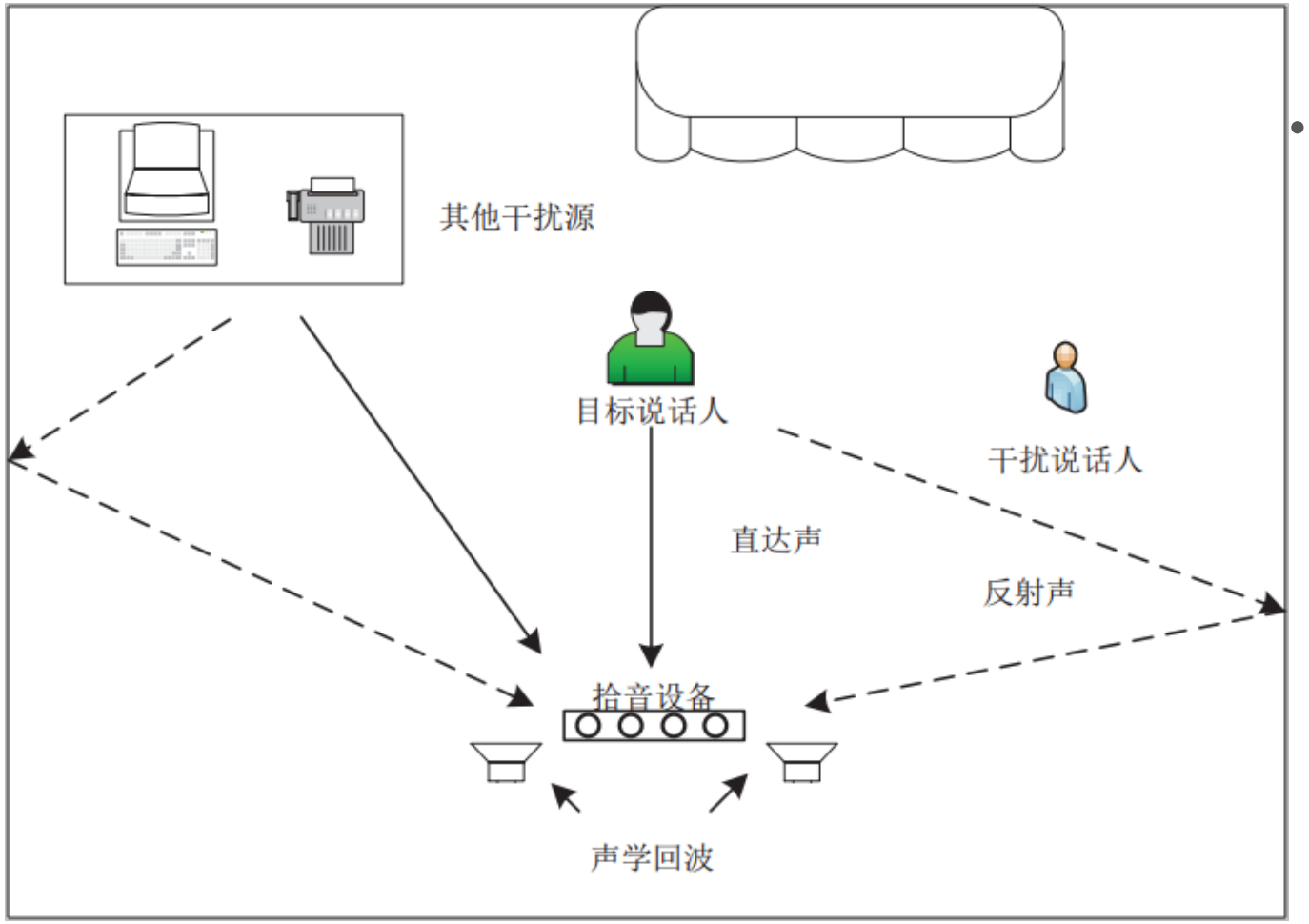
阿里巴巴达摩院机器智能语音实验室 付强 纳跃跃 田彪
Johns Hopkins University 王晓飞

什么是前端处理

服务于自然人机语音交互

“自然”意味着对语音交互的场合、使用模式等无约束！

痛点问题



- 远讲（场）交互，目标声源距离拾音设备较远，更易受到**声学回声、干扰声源、背景噪声、房间混响**等各种不利因素的影响

听不清。。。

听清世界的声音

人类需要听清——语音通信

- 更低的处理延时
- 更高的主观听感和可懂度

机器需要听清——语音识别

- 更高的信噪比
- 更好的声学模型适配

- 面对回声、干扰、噪声和混响等各种不利因素的挑战；
- 综合运用信号处理、机器学习手段以及融合语义层面的信息，提高目标语音的信噪比，增强后续处理的声环境稳健性。

一言以蔽之，前端处理是为了让获取的语音更加清晰自然，“听清世界的声音”！

场景碎片化

家

远讲、方向性干扰、房间混响较高

车

扩散场噪声强、混响小

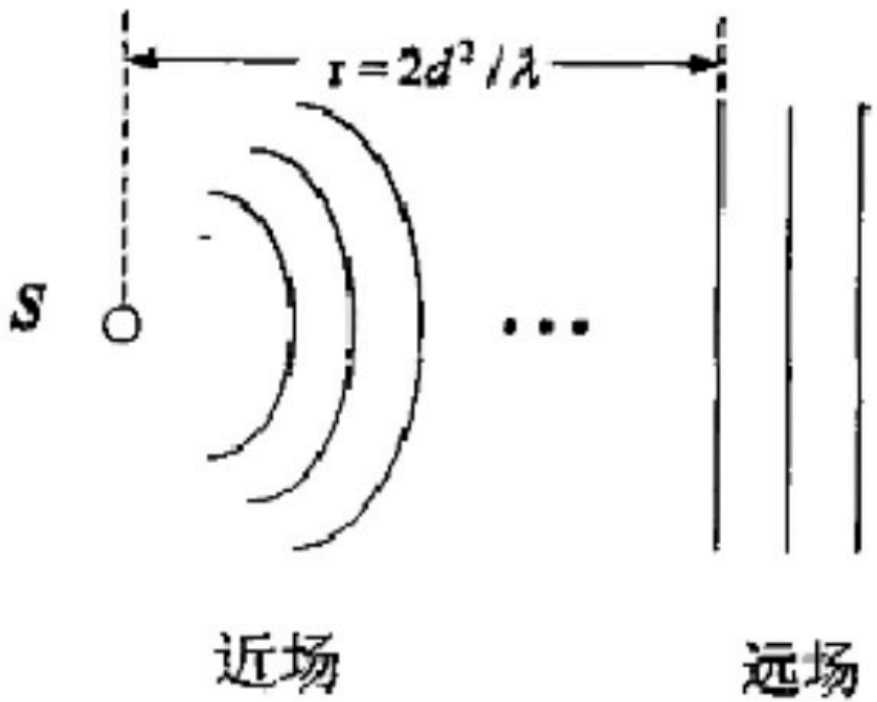
机场、展会

房间混响高、扩散场噪声强

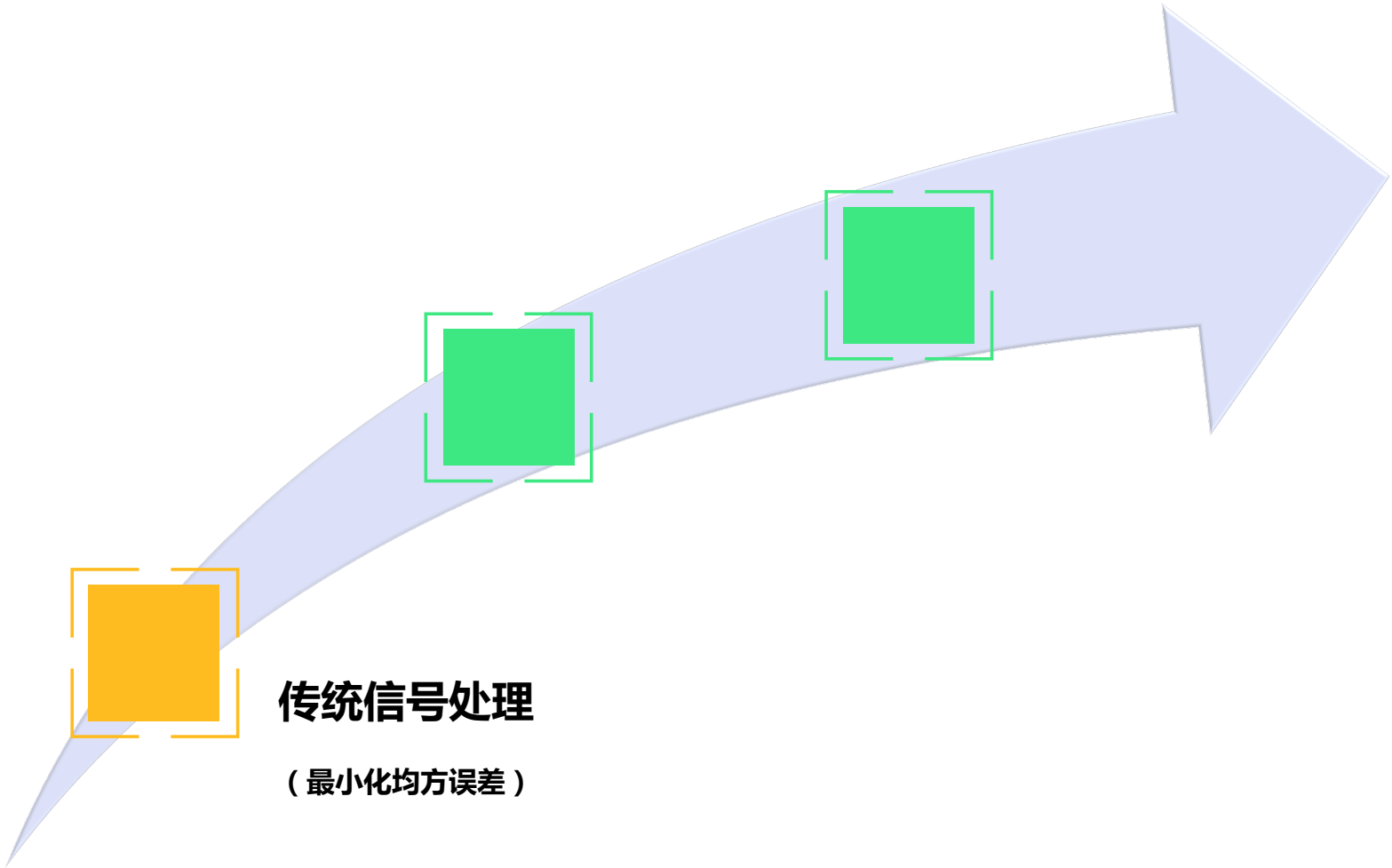
近场与远场

- 界限不绝对：远场声源距麦克风阵列中心远大于信号波长，反之是近场

- 近场：球面波假设- 幅度差；远场：平面波假设-延迟/相位差



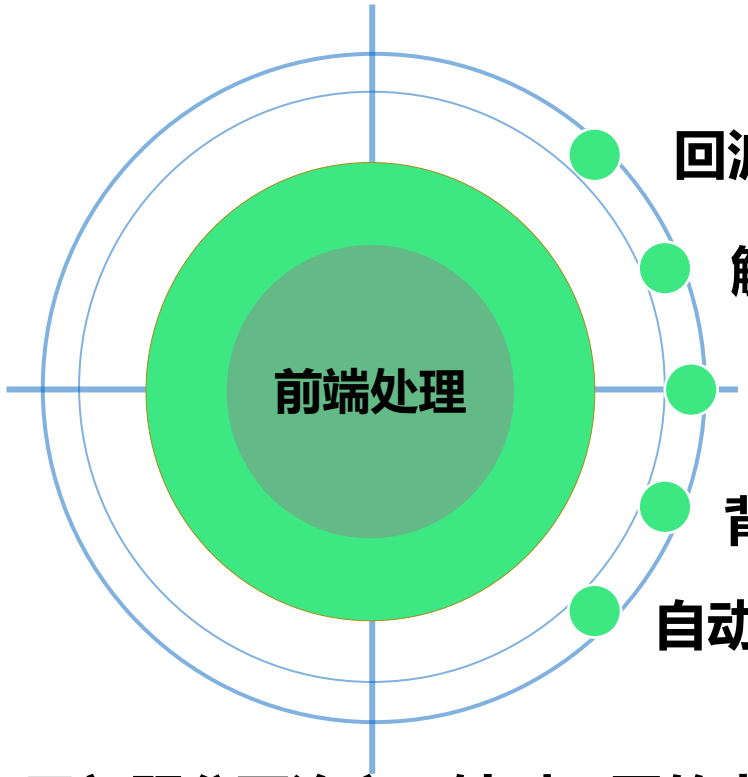
技术路线（1）



传统信号处理

（最小化均方误差）

传统端侧信号处理



回波抵消

解混响

宽带空域滤波

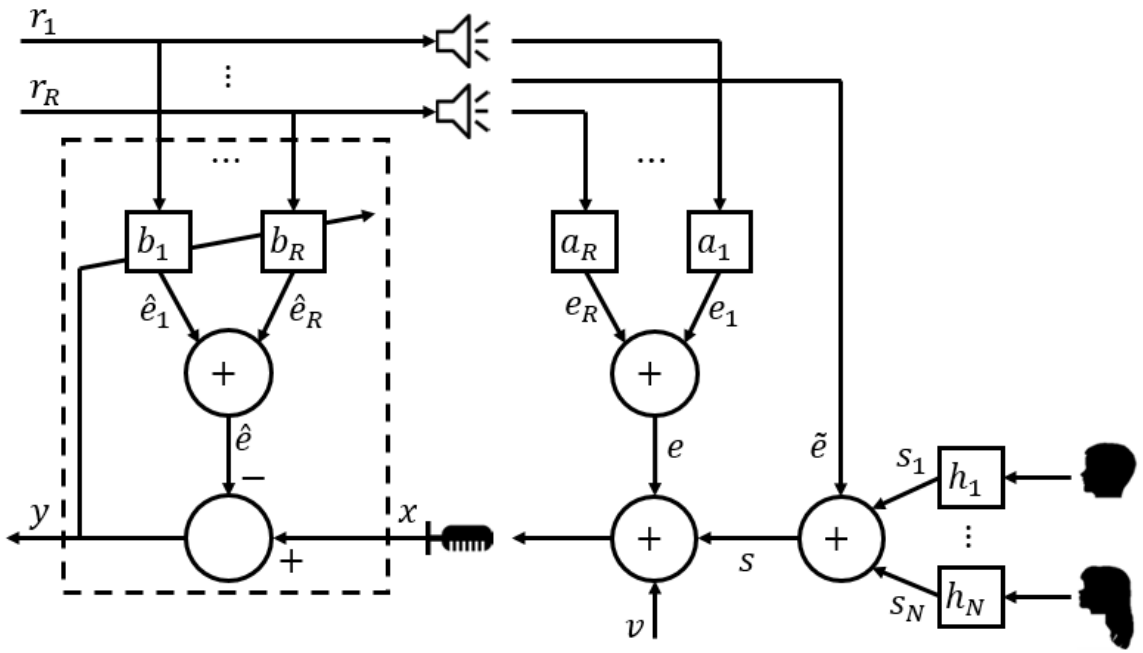
背景噪声抑制

自动增益控制

规则 语音时/频/空域特性

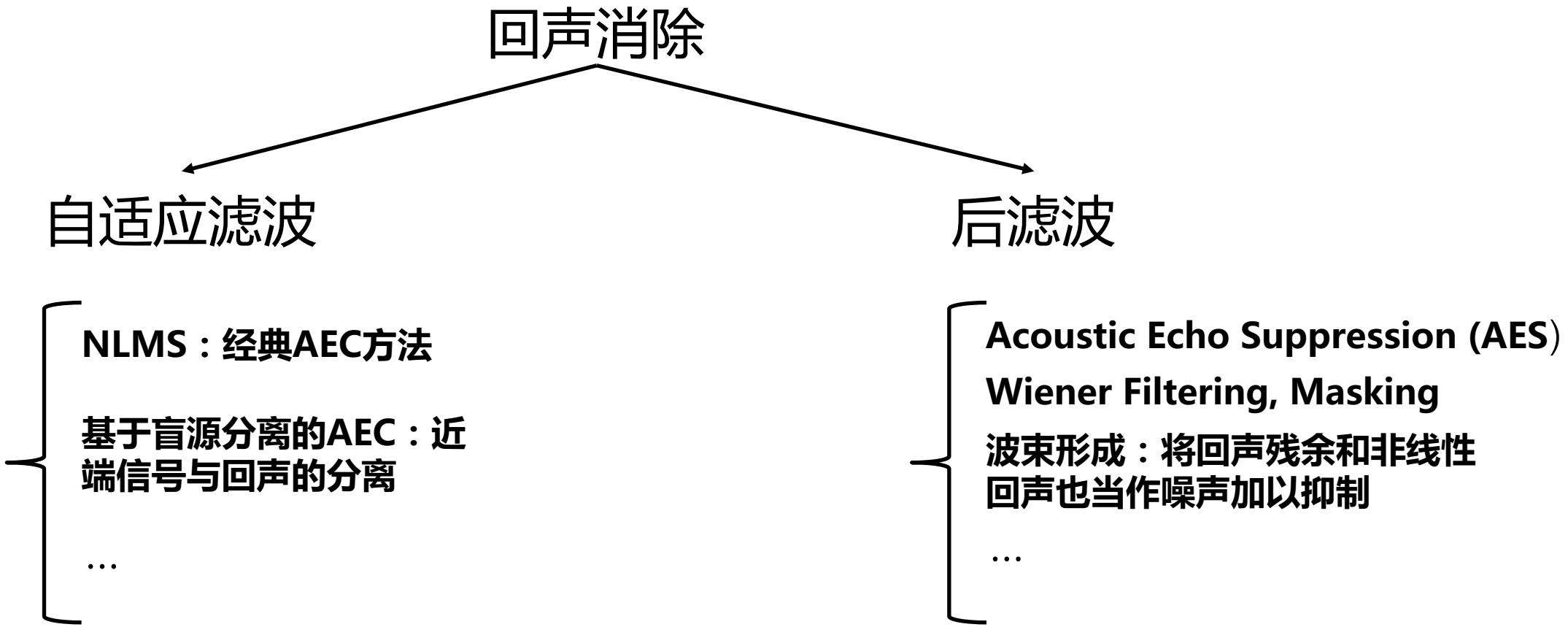
- 子问题分而治之：针对不同的声学影响采用不同的信号处理算法加以解决
- 优化目标：抑制非目标相关成分
- 优化准则：最小化均方误差

回声消除（Acoustic Echo Cancellation, AEC）

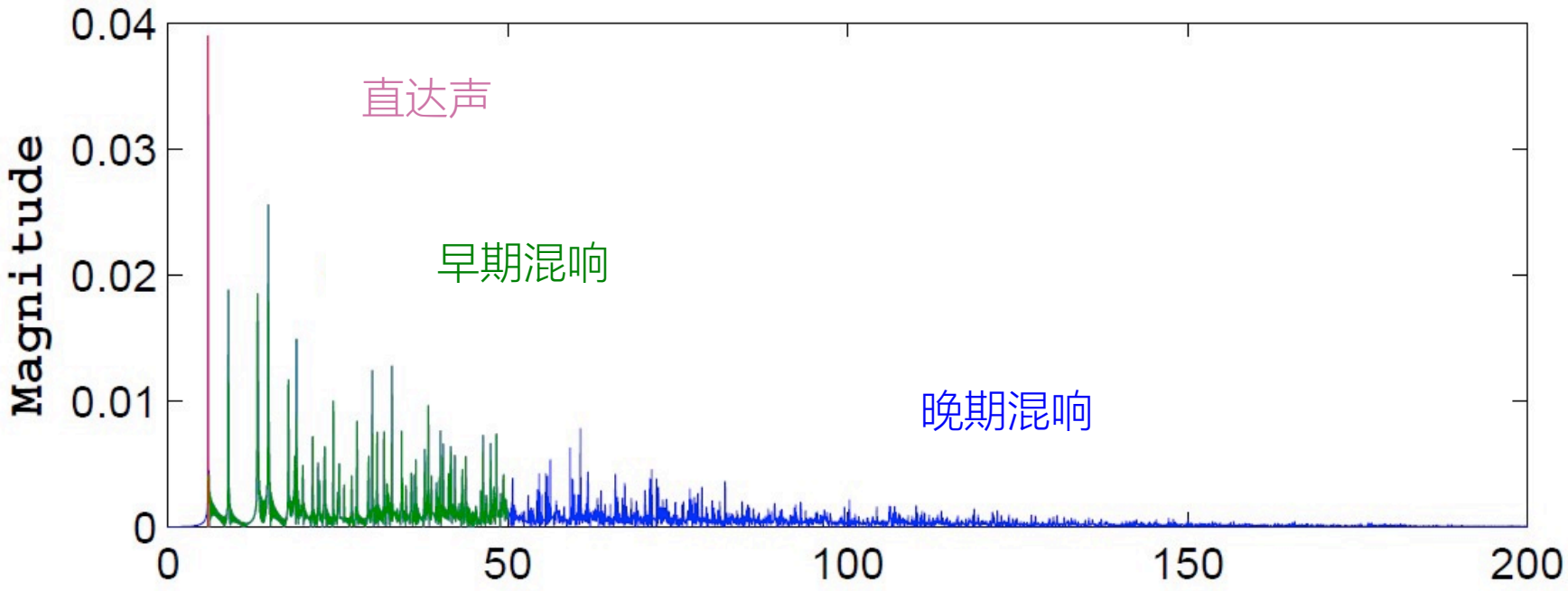


- 由于回放系统的限幅、音效、设备振动等原因，会产生非线性回声$\tilde{e}$，难以通过线性AEC处理。
- 后滤波、非线性AEC、beamforming等技术可用于对非线性回声的抑制。

典型回声消除技术

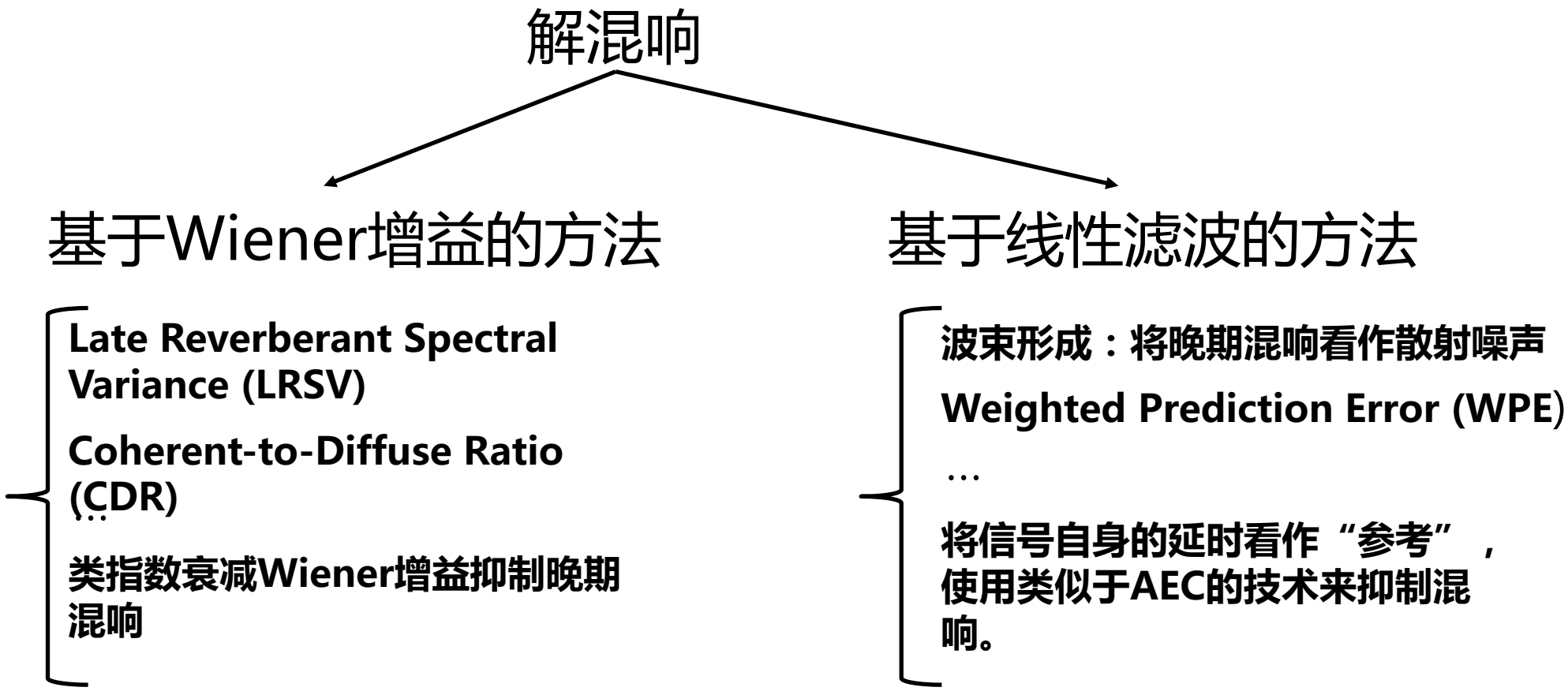

回声消除
自适应滤波
后滤波
NLMS：经典AEC方法
基于盲源分离的AEC：近端信号与回声的分离
…
Acoustic Echo Suppression (AES)
Wiener Filtering, Masking
波束形成：将回声残余和非线性回声也当作噪声加以抑制
…

解混响 (Dereverberation)

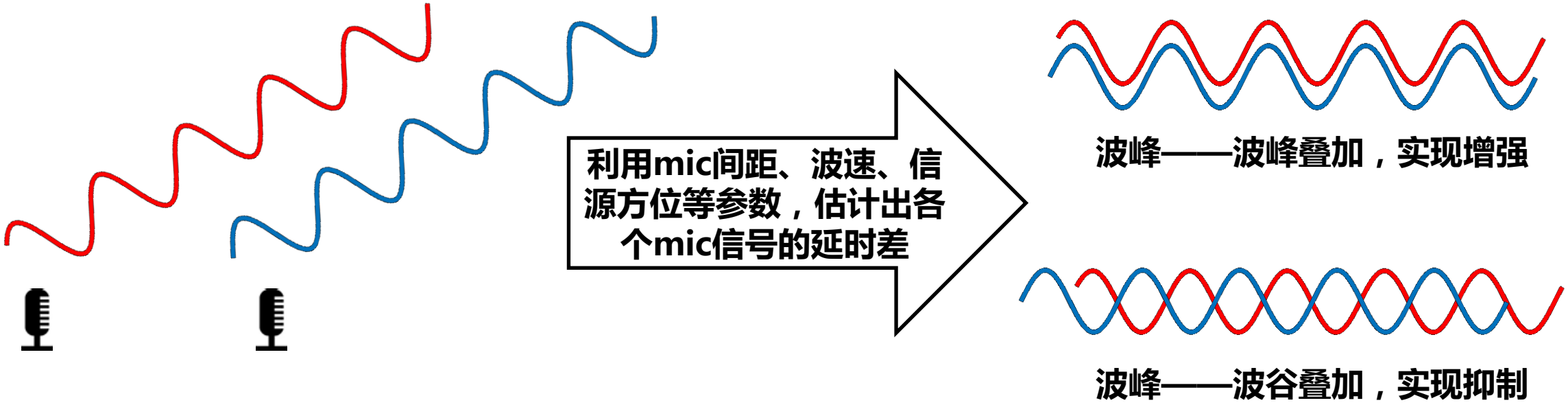


- **典型房间冲击响应可分为直达声、早期混响、晚期混响三部分。**
- **解混响技术用于抑制晚期混响，提高语音信号质量和可懂度。**

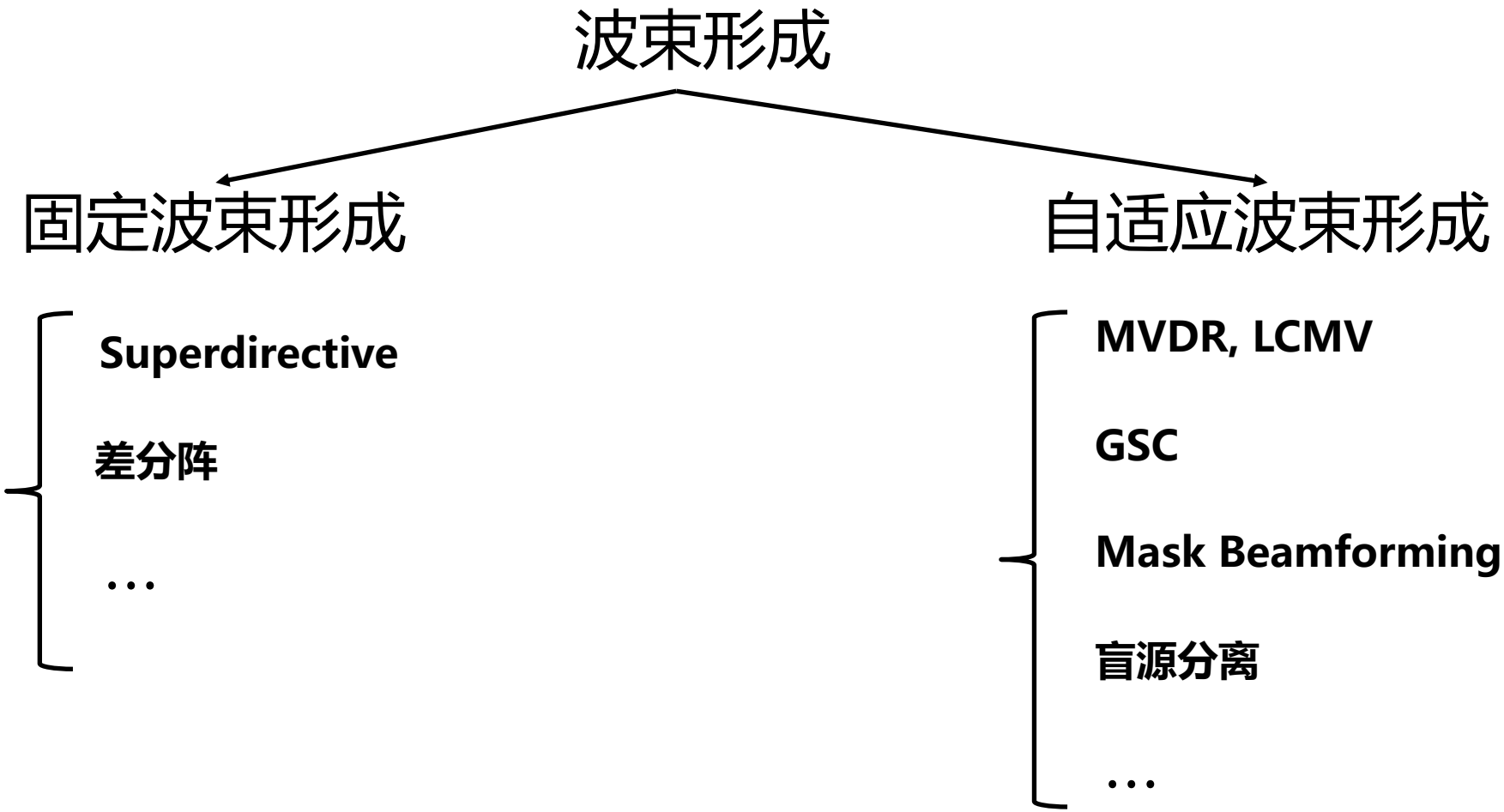
典型解混响技术


解混响
基于Wiener增益的方法
Late Reverberant Spectral Variance (LRSV)
Coherent-to-Diffuse Ratio (CDR)
…
类指数衰减Wiener增益抑制晚期混响
基于线性滤波的方法
波束形成：将晚期混响看作散射噪声
Weighted Prediction Error (WPE)
…
将信号自身的延时看作“参考”，使用类似于AEC的技术来抑制混响。

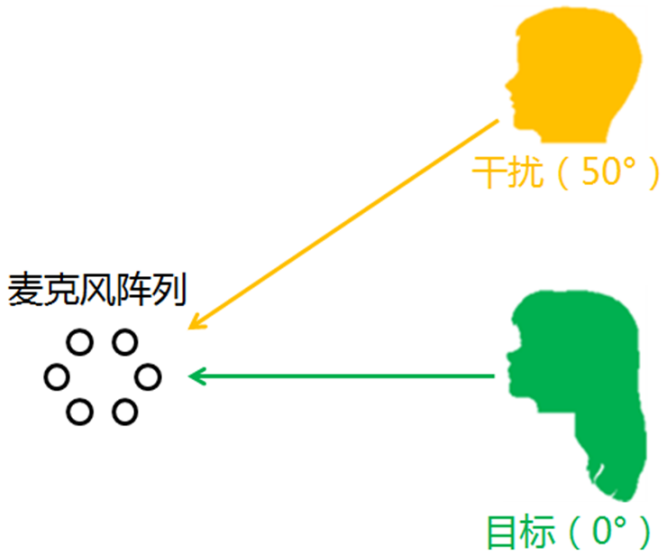
波束形成（Beamforming）的工作原理
——直观解释

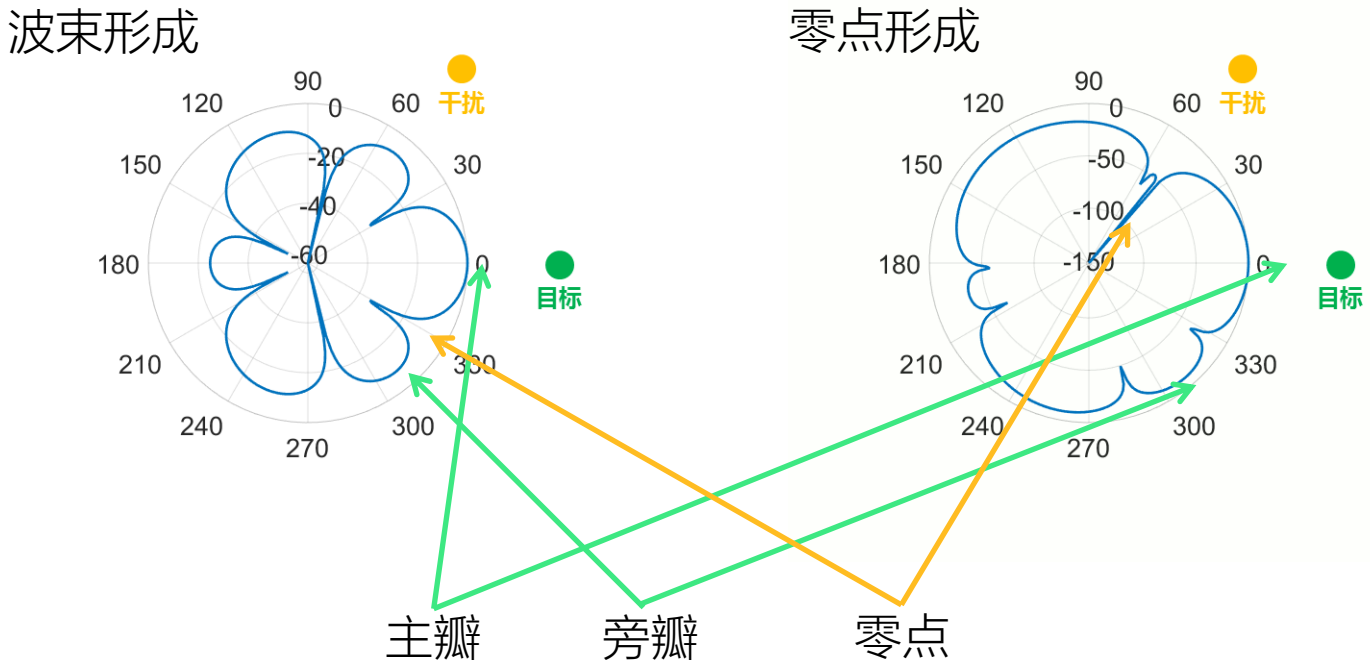


典型的波束形成技术


波束形成
固定波束形成
Superdirective
差分阵
…
自适应波束形成
MVDR, LCMV
GSC
Mask Beamforming
盲源分离
…

波束（BEAM）与零点（NULL）形成

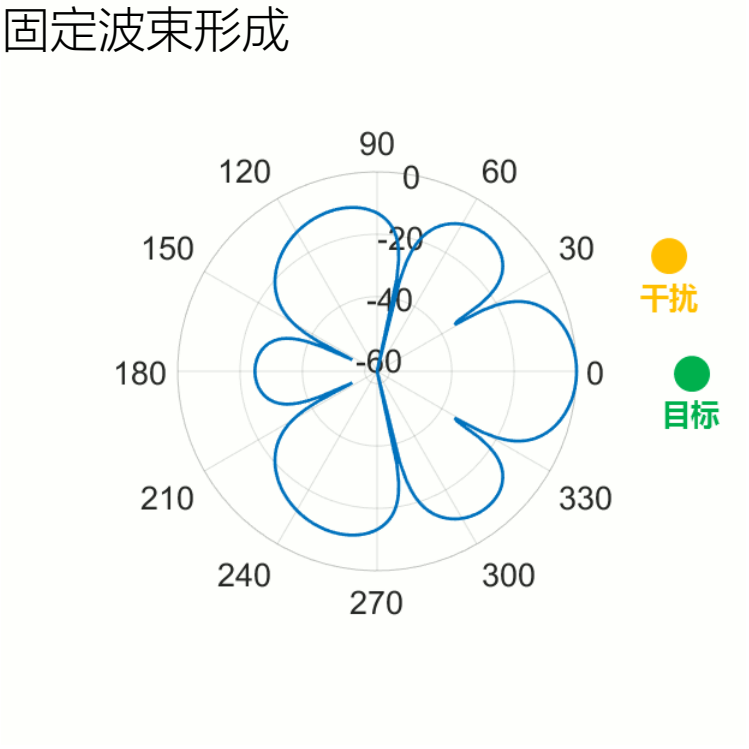




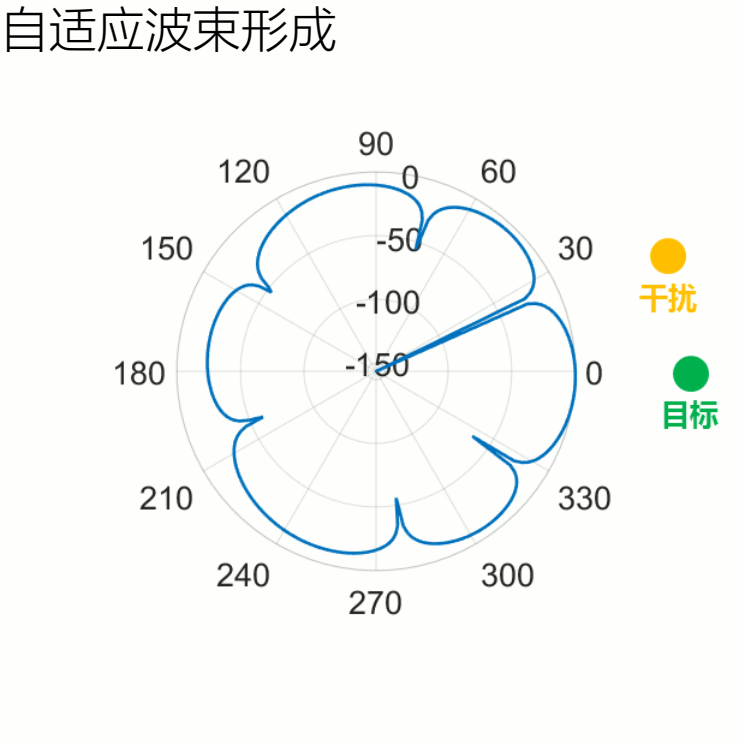
- **波束形成：重点在于形成主瓣，抑制主瓣外的信号，但对来自于旁瓣处的信号抑制能力有限**
- **零点形成：重点在于构造零点，显著抑制位于零点处的信号，但主瓣较宽，对散射噪声的抑制能力有限**

固定与自适应波束形成

固定波束形成

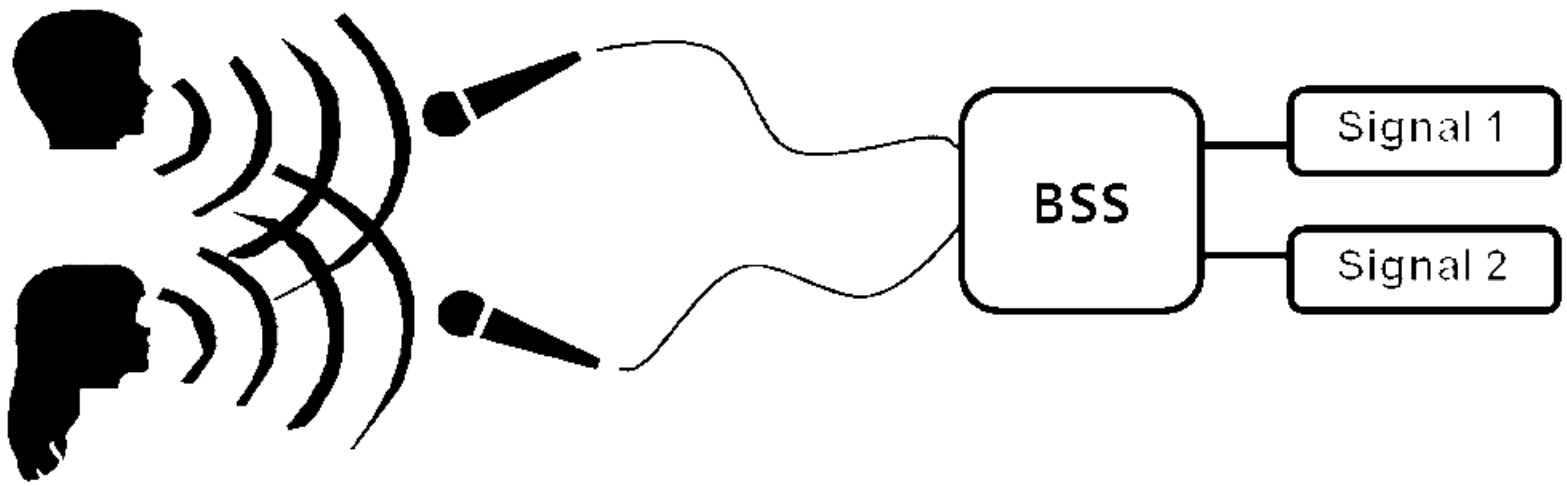


自适应波束形成



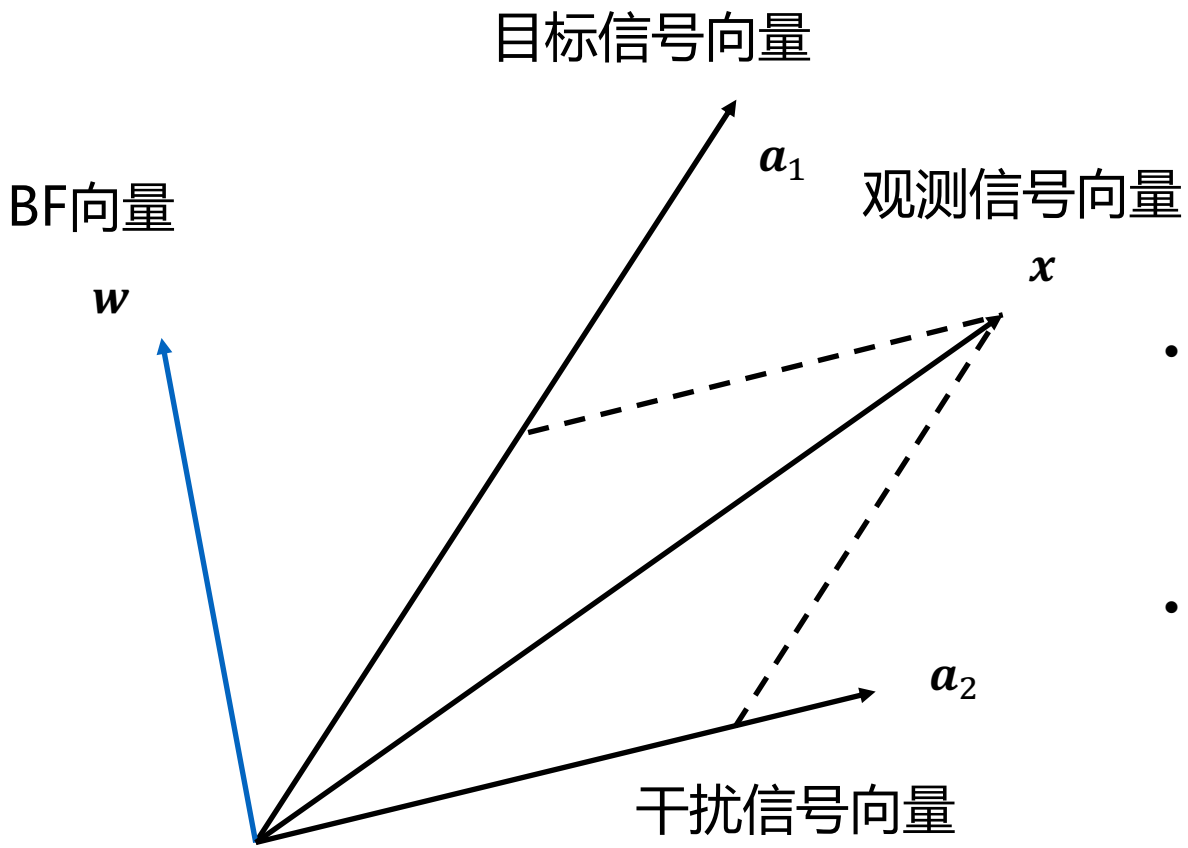
- **固定波束形成：根据阵列的拓扑结构预先设计好波束形状，波束不随实际环境的变化而改变。**
- **自适应波束形成：根据环境自适应调整波束形状，在目标处形成主瓣，在干扰处形成零点。**

盲源分离（Blind Source Separation, BSS）



- 不同物理过程发出的信号可看作是相互独立的
- 通过最大化输出信号间的独立性来达到分离的目的
- 盲源分离也可以看作一种自适应零点的波束形成

波束形成的工作原理——几何解释



- **Beamformer与观测信号作用，相当于向量的投影操作；**

  Beamformer与噪声成分正交，所以可以抑制噪声成分

- **理想条件下，不同的自适应波束形成算法得到的**

  Beamformer方向是平行的，只是向量的大小不同。即

  自适应波束形成算法之间具有“等价性”

$w \perp a_2$，对干扰有抑制作用

问题与趋势

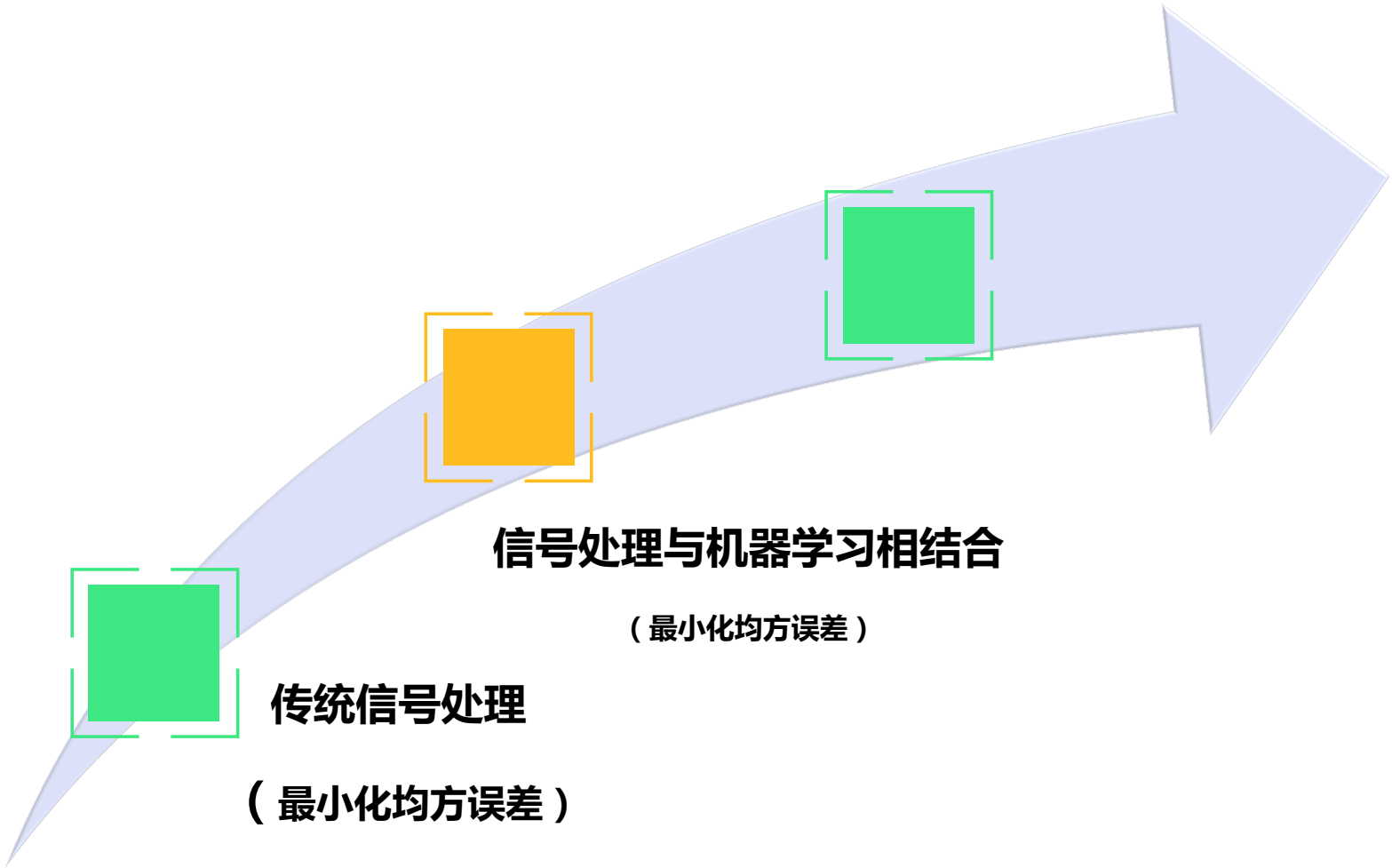
01 极低信噪比

02 多声源干扰

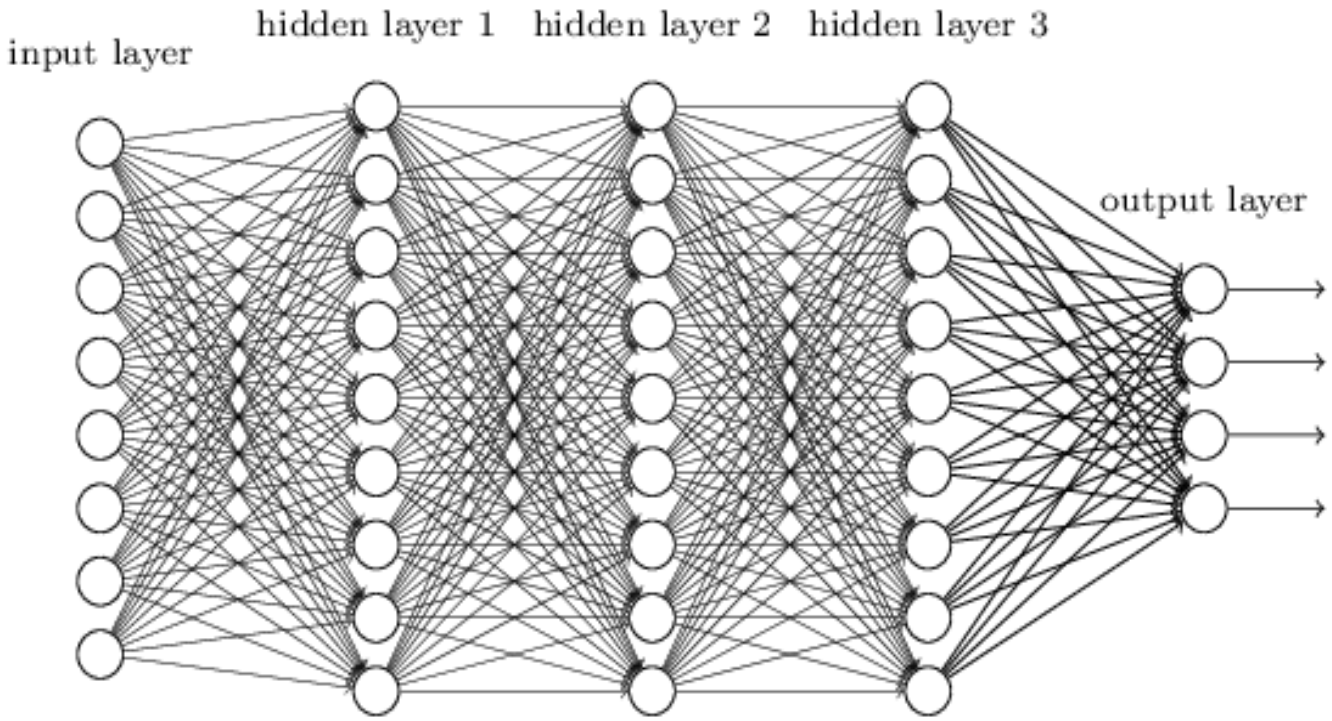
03 移动声源干扰

04 端到端？

技术路线（2）



关于深度学习



深度学习：

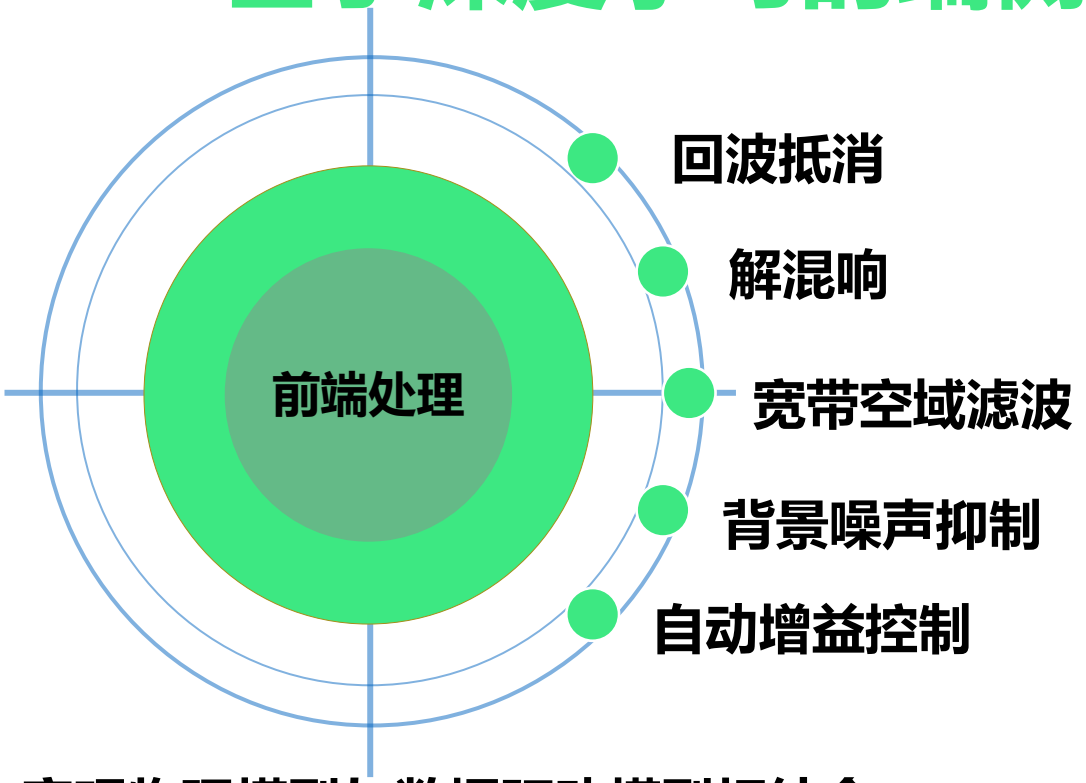
- **Hot but not so new**
- **多层非线性自适应滤波器/状态空间模型**

  模拟大脑的数据处理过程
  - **更多的数据**
  - **更强的计算能力**

突破性进展：

- **语音识别与对话系统**
- **机器翻译**
- **自然语言理解**

基于深度学习的端侧信号处理



规则 + 学习

客观物理模型

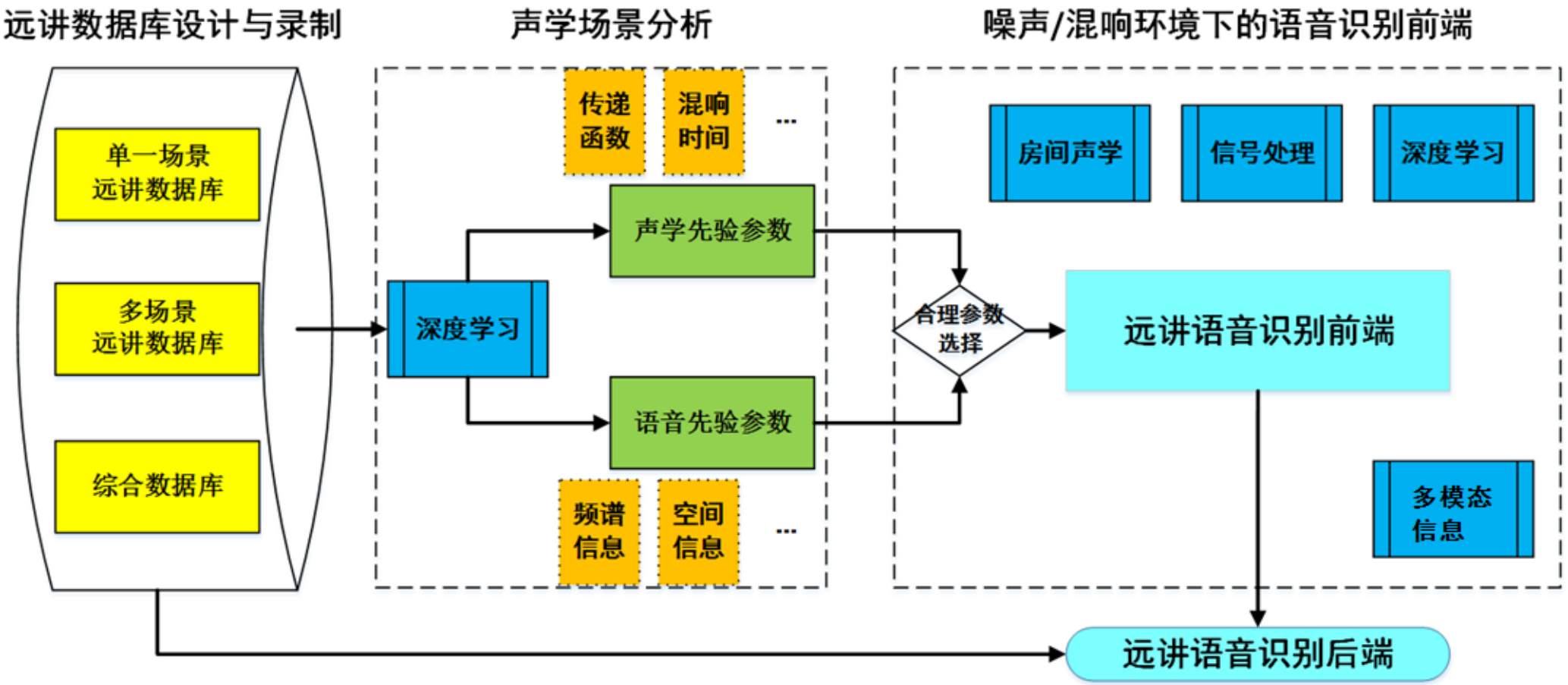
音频时频空域特性

+

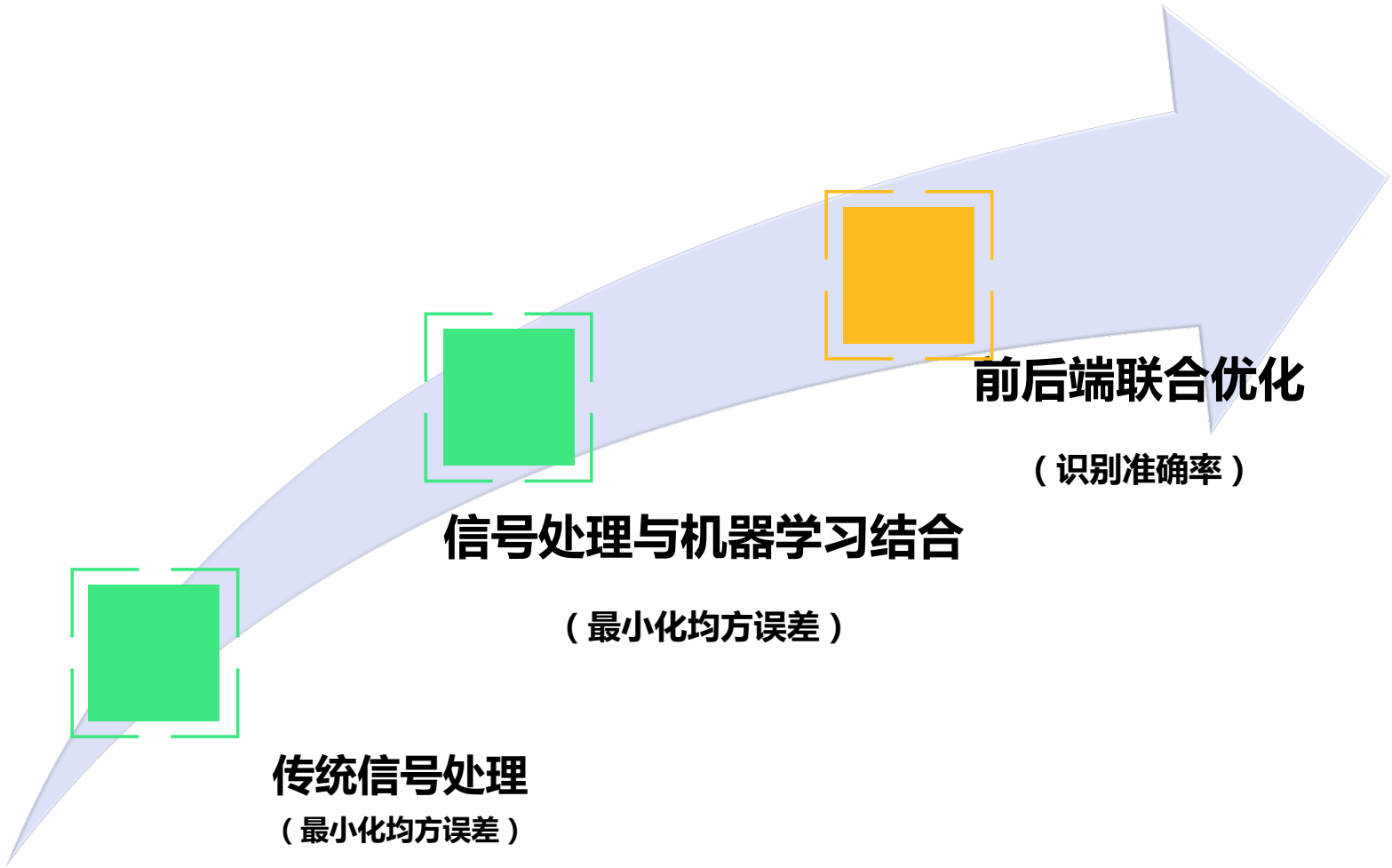
海量音频先验信息

- 客观物理模型与数据驱动模型相结合
- 既遵从了声源和声传播的物理规律，又利用了先验数据统计建模带来的稳健性和性能提升
- 优化准则未变，依然是最小化均方误差

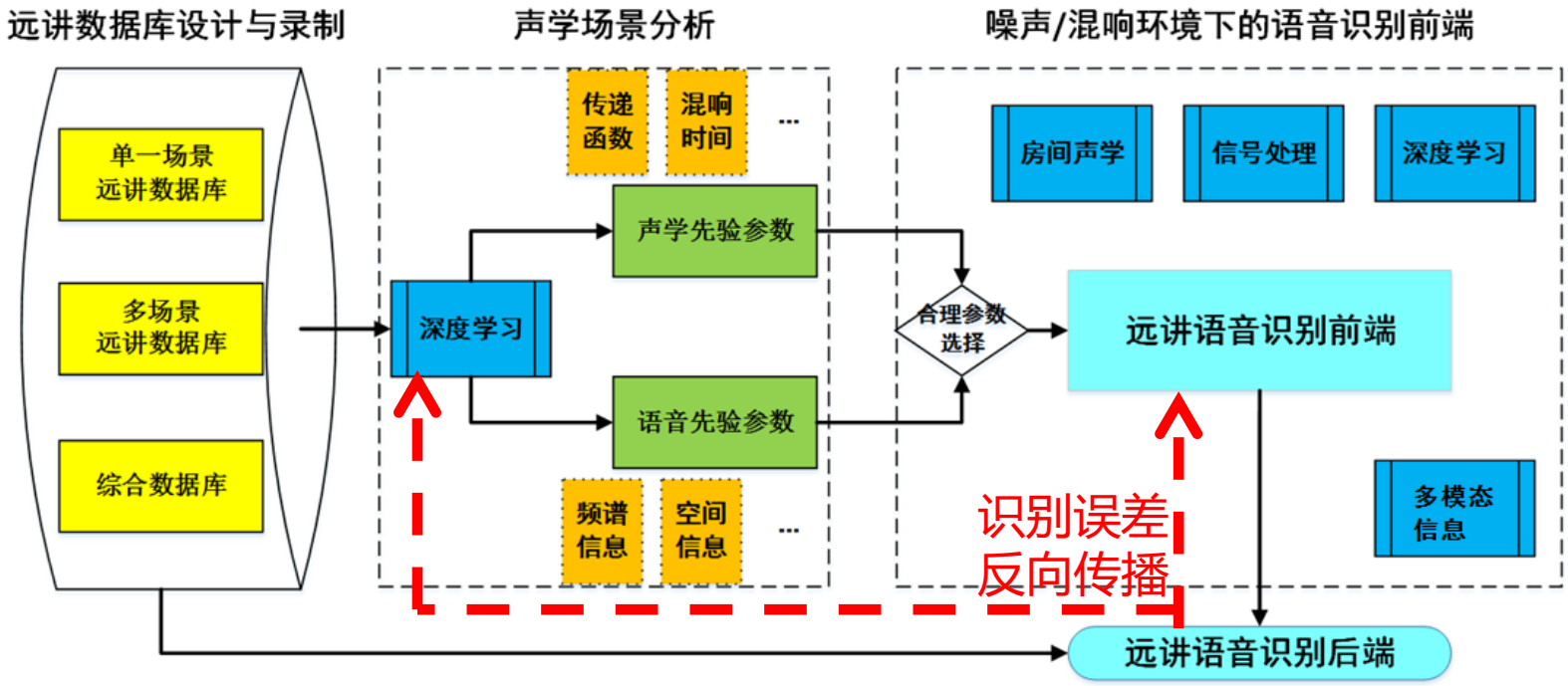
深度学习+前端处理系统



技术路线（3）



深度学习框架下的前后端联合优化



- **前端和后端都以语音识别准确率为优化目标：识别误差从后端声学模型反向传播回前端，用于指导前端的优化**
- **途径1：端到端，前后端融合成一个统一的模型，输入为原始语音，输出为识别结果**
- **途径2：将后端声学模型的梯度反向传播到前端，用于指导前端的神经网络训练**
- **途径3：推理阶段，后端声学模型给出实时反馈信息用于指导前端参数更新**

感谢聆听！