中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

# 语音交互技术
# ——语音增强

中国科学院自动化研究所
模式识别国家重点实验室

陶建华
jhtao@nlpr.ia.ac.cn

# 本节课提纲

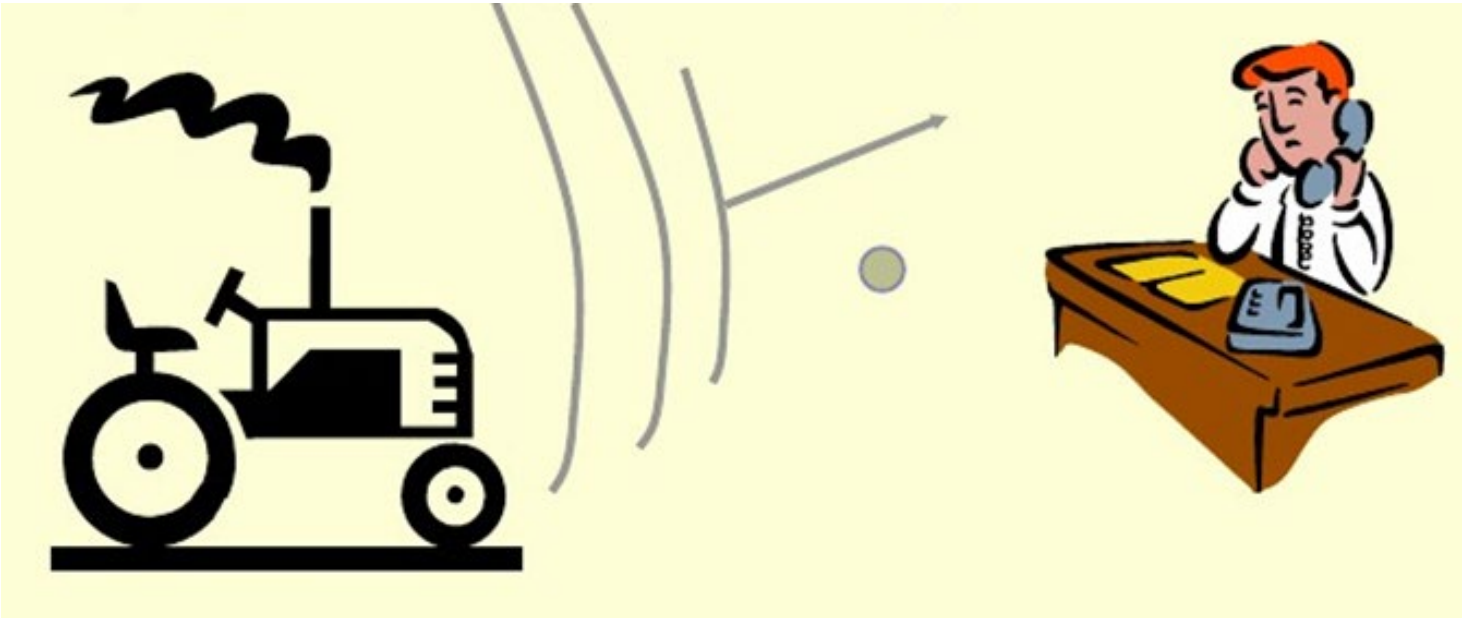
- 语音增强的定义
- 语音增强的任务
- 语音增强技术概述

# 本节课提纲

- 语音增强的定义
- 语音增强的任务
- 语音增强技术概述

# 什么是语音增强？

语音增强是指当语音信号被不同干扰、甚至淹没后，从复杂背景中提取有用的语音信号，抑制干扰的技术。



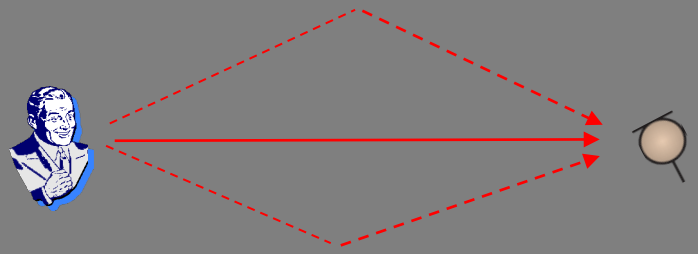
# 语音增强有什么意义？

- 在国家和社会安全方面，侦听信号常常含有较大的噪声，语音增强有助于提高侦听系统的效果；
- 飞机驾驶人员与地面指挥的语音通信，常受到噪声干扰，需要语音增强保证语音的可靠传达
- 有效的语音增强系统能够提高语音通讯的抗干扰能力，扩展移动通信的适应能力和应用范围
- 室内回声在比较严重的情况下会和原始语音一起播放，影响收听效果，需要语音增强算法抑制回声
- 语音增强在语音识别、语音编码等领域有着重要的应用，是语音交互系统中最前端的预处理模块。

# 典型声学环境

## 噪声类型

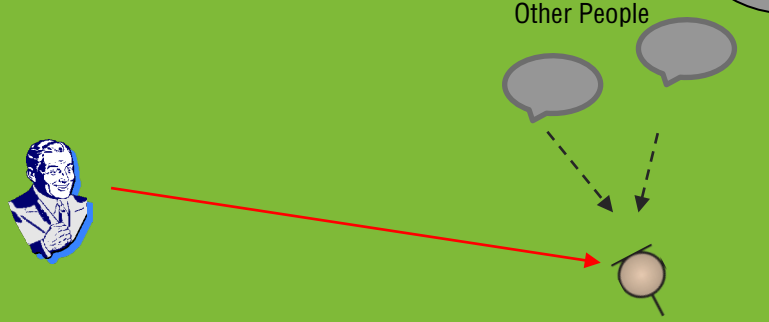
混响： Reverberation



背景噪声： Background Noise


噪声

人声干扰： Interference




回声： Echo

# 语音增强的典型应用

- 语音通信: 移动设备
- 语音识别: 智能家居
- 语音修复: 助听器








Home assistants

Game consoles



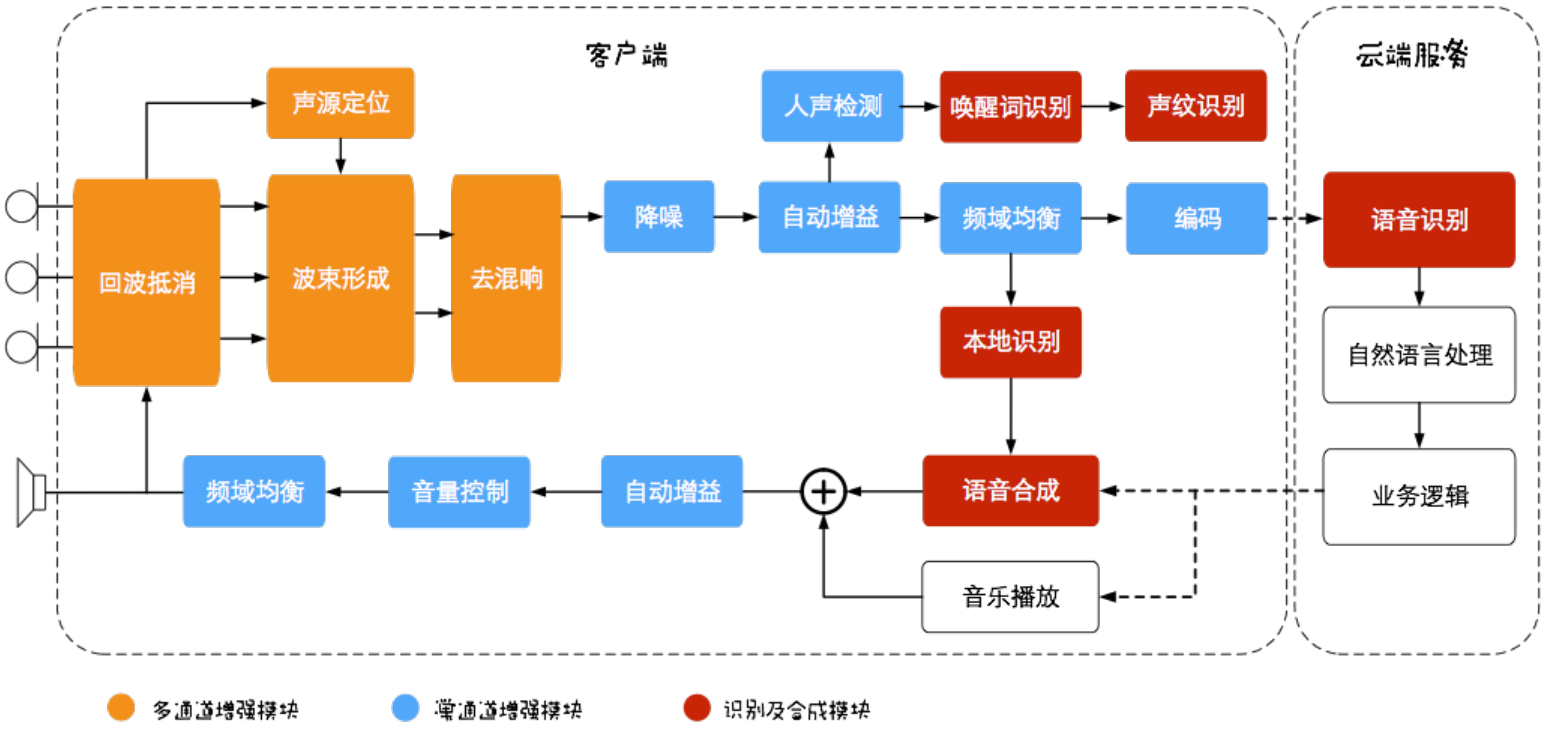
Robots

Voiced controlled appliances

# 本节课提纲

- 语音增强的定义
- 语音增强的任务
- 语音增强技术概述

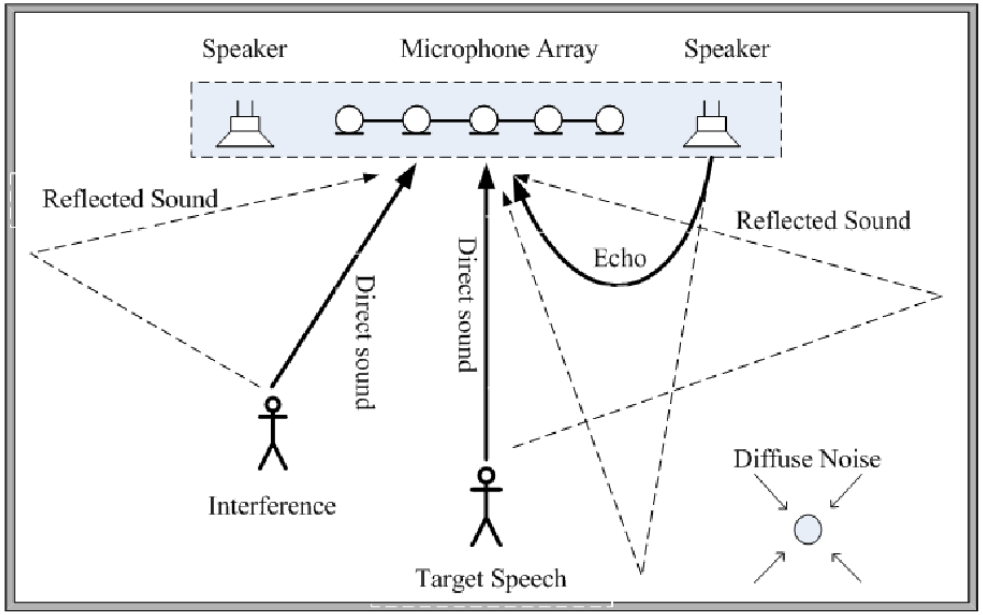
# 典型语音交互系统架构




图中展示了四种主要的干扰源：混响、背景噪声、人声干扰和回声，在真实的语音交互系统中，上述四种噪声同时存在，这给真实环境下的语音识别带来了较大的挑战。

# 语音增强的任务分解

- 更自然的语音交互和通信
  - 解决技术痛点，实现释放双手的语音交互
- 语音增强涉及的主要问题包括
  - 传感、声源定位、声源跟踪、语音降噪、去混响、声源分离、回声对消、场景分析等



# 本节课提纲

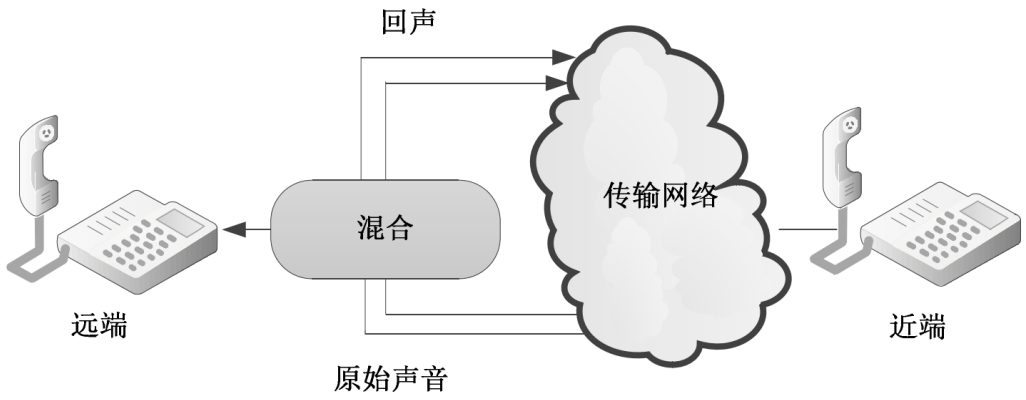
- 语音增强的定义
- 语音增强的任务
- 语音增强技术概述

# 语音增强技术概述

- 语音增强技术主要分为：
  - 回声消除
  - 混响消除
  - 语音降噪
- 语音降噪又分为单通道语音降噪和多通道语音降噪。
  - 单通道是指只有一个麦克风可以利用。单通道语音降噪主要包括传统的信号处理的方法，比如：谱减法、最小均方误差和维纳滤波法等；非负矩阵分解；基于深度学习的方法等。
  - 多通道则是指有多个麦克风收集到的语音信号。多通道语音降噪的方法主要包括：延时求和法、最小方差无失真响应、广义旁瓣滤波和数据驱动的多通道语音降噪方法。

# 回声消除

- 最早应用于语音通信的方法
- 难点：
  - 远端信号与近端信号同步
  - 双讲模式下回波消除
- 挑战：
  - 非线性失真
  - 声音传输衰减
  - 噪声和回声同时存在
  - 信号的时变和非稀疏



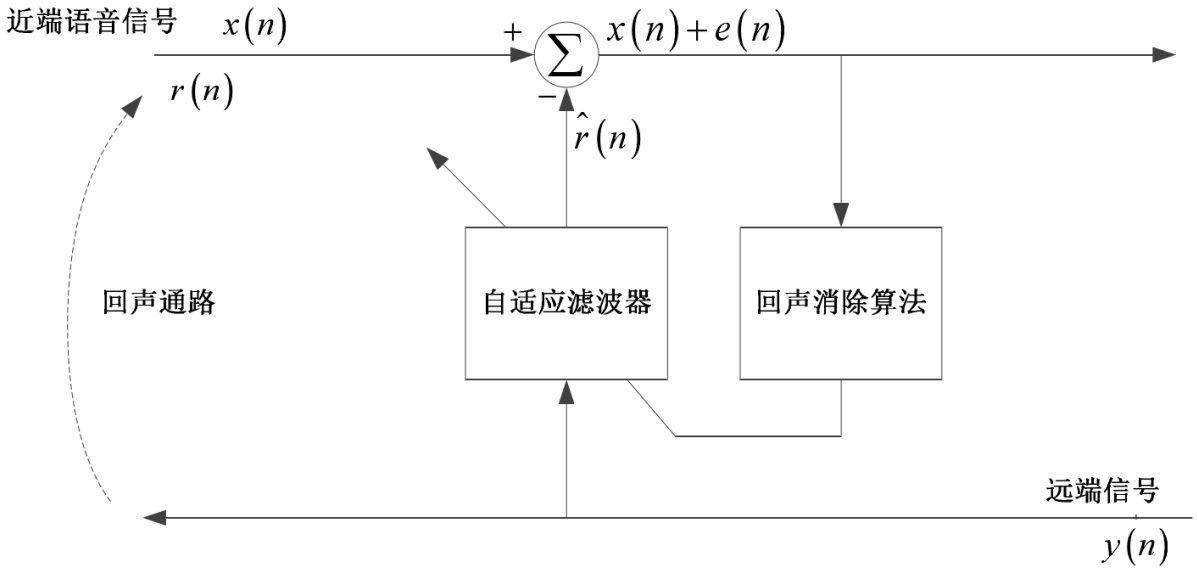
# 常见的回声消除方法

- 对周围环境进行特殊的处理
  - 尽可能的减少扬声器发出的声音直接进入到麦克风
  - 减少室内环境对声音的反射作用
- 回声隔离器
  - 相对较为古老的一种方法，没有任何算法
  - 简单的交替关闭麦克风和扬声器

# 常见的回声消除方法

## 回声抵消器

### 自适应滤波算法




自适应算法通过回声信号和滤波器输出信号相减以后的误差信号来调整自适应滤波器的权值，从而使得滤波器的输出信号更接近于回声信号

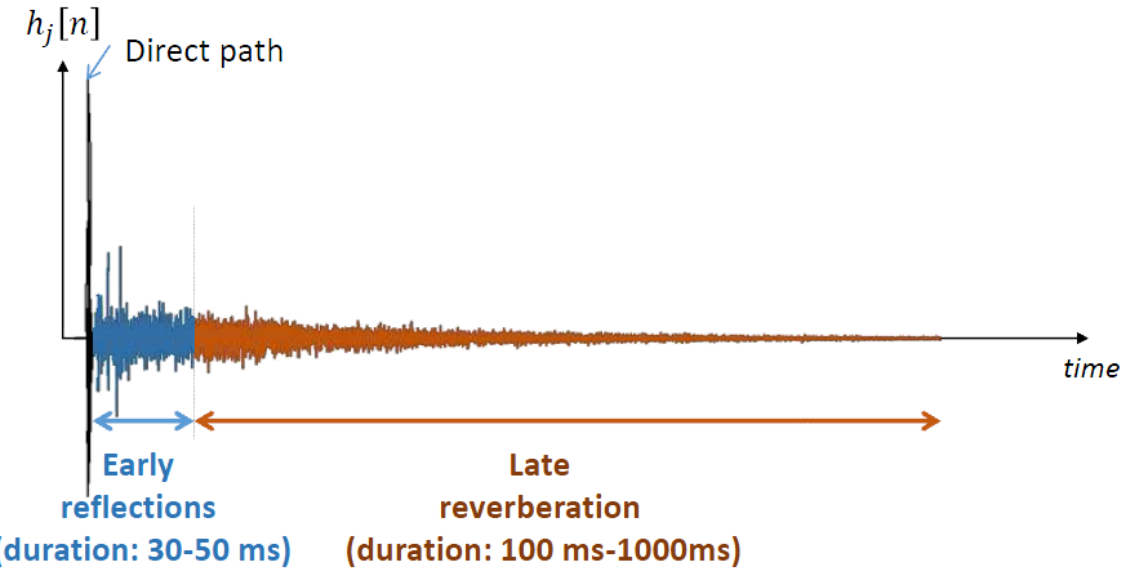
# 混响消除

- 所谓混响就是声音的直达声与反射声很紧凑的重合在一起时人耳所听到的声音，这个效果在语音的后期处理时特别有用。

- 房间脉冲响应（Room Impulse Response）

声音在传输过程中会经过墙壁或其它障碍物的反射后到达麦克风

- 通过T60描述混响时间
- 房间混响时间通常在 200-1000 ms范围
- 下图是一个典型的房间脉冲响应

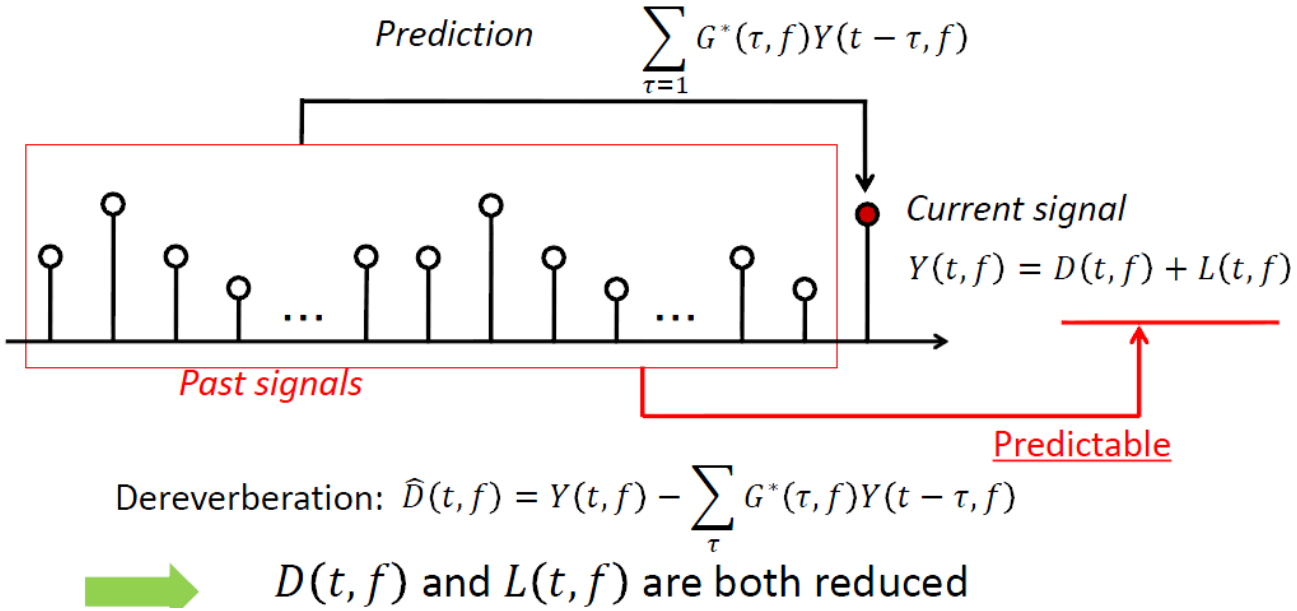


# 混响消除

- 混响消除的主流方法
  - 基于波束形成方法
  - 基于逆滤波方法
  - 基于语音增强方法
  - 基于深度学习方法
- 加权预测误差方法（WPE）

# 混响消除-加权预测误差(WPE)方法

## 线性预测




Dereverberation: $\widehat{D}(t,f) = Y(t,f) - \sum_{\tau} G^*(\tau,f)Y(t-\tau,f)$

$D(t,f)$ and $L(t,f)$ are both reduced

这种方法的思想和语音编码中的线性预测系数有些相似，如图所示，混响语音信号Y可以分解为安静语音成分D混响成分L，L可以通过先前若干点的Y加权确定，G表示权重系数；WPE算法的核心问题是确定G，然后估计出混响消除后的语音。

# 混响消除-加权预测误差(WPE)方法

## 估计线性预测系数

该算法通过如下目标函数估计滤波器系数，具体推倒过程如下所示

通过如下目标函数估计滤波器系数

$$\{\hat{G}(\tau,f)\} = \underset{\{G(\tau,f)\}}{\operatorname{argmin}} \sum_t \left\| Y(t,f) - \sum_{\tau=1} G^*(\tau,f) Y(t-\tau,f) \right\|^2$$

通过如下目标函数估计滤波器系数

$$\mathbf{g}_f = \mathbf{R}_f^{-1}\mathbf{r}_f$$

where,

$$\mathbf{g}_f = [G(1,f) \dots G(L,f)]^T$$

$$\mathbf{R}_f = \sum_t \mathbf{y}_{t,f}\mathbf{y}_{t,f}^H$$

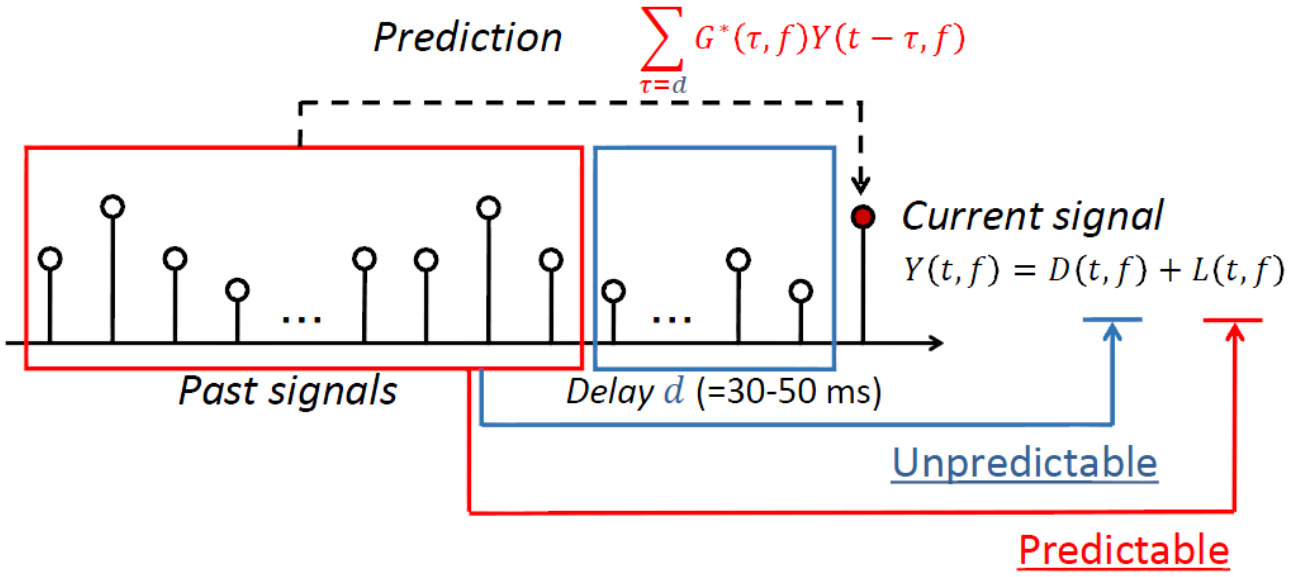
$$\mathbf{r}_{f,d} = \sum_t \mathbf{y}_{t,f} Y^*(t,f)$$

$$\mathbf{y}_{t,f} = [Y(t,f) \dots Y(t-L,f)]^T$$

更为详细的算法流程可以参考一下网址推荐的论文。
http://www.kecl.ntt.co.jp/icl/signal/takuya/research/dereverberation.html

# 混响消除-加权预测误差(WPE)方法

- 延时线性预测：保留早期混响成分

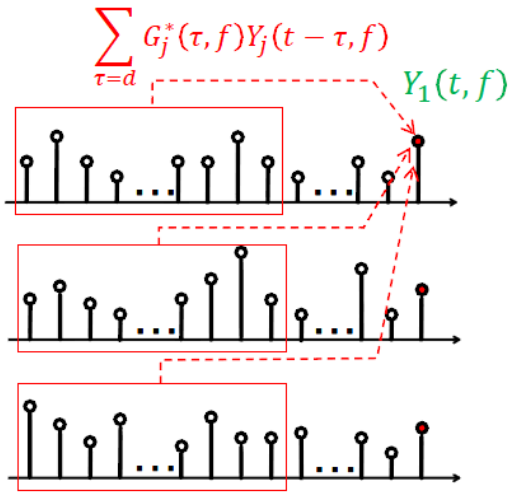



由于早期混响成分有助于提高语音的可懂度，因此可以对上一页的方法进行改进，只抑制晚期混响成分。如图所示D同时包括安静语音成分和早期混响成分，通过先前若干点的Y确定L时没有考虑早期混响成分。

# 混响消除-加权预测误差方法

## 多通道加权预测误差方法

多输入多输出混响消除方法



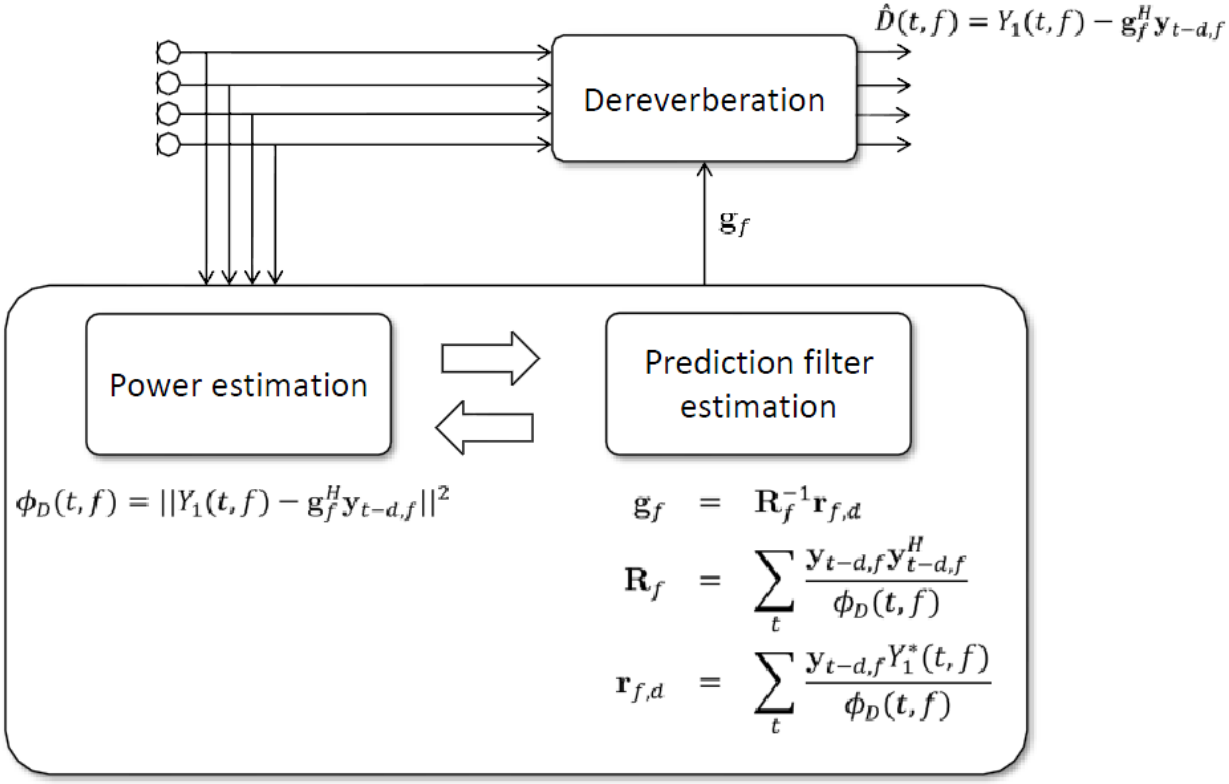

$$\widehat{D}(t,f) = Y_1(t,f) - \sum_{j=1}^{J}\sum_{\tau=d} G_j^*(\tau,f)Y_j(t-\tau,f)$$
$$= Y_1(t,f) - \mathbf{g}_f^H \mathbf{y}_{t-d,f}$$

$$\mathbf{y}_{j,t,f} = [Y_j(t,f) \dots Y_j(t-L,f)]^T$$
$$\mathbf{y}_{t,f} = [\mathbf{y}_{1,t,f}^T, \dots, \mathbf{y}_{J,t,f}^T]^T$$
$$\mathbf{g}_{j,f} = [G_j(1,f) \dots G_j(L,f)]^T$$
$$\mathbf{g}_f = [\mathbf{g}_{1,f}^T, \dots, \mathbf{g}_{J,f}^T]^T$$
$$\hat{\mathbf{g}}_f = \underset{\mathbf{g}_f}{\operatorname{argmin}} \sum_t \frac{\left\| Y_1(t,f) - \mathbf{g}_f^H \mathbf{y}_{t-d,f} \right\|^2}{\phi_D(t,f)}$$

在此基础上将WPE方法扩展到多通道混响消除模式，此时某一通道的晚期混响成分L可以通过各个通道先前若干点的Y加权确定，通过估计最优的权重系数G，消除晚期混响成分的干扰。

# 混响消除-加权预测误差方法

## 加权预测误差方法流程图




图中为基于WPE的多通道混响消除的流程，如果所示需要经过多次迭代确定出滤波器系数g，生成出混响消除后的语音。输出的去混响后的各通道语音可以作为波束形成算法的输入。

# 单通道语音降噪

- 单通道是指只有一个麦克风可以利用。单通道语音降噪主要包括传统信号处理的方法，比如：
  - 谱减法
  - 最小均方误差
  - 维纳滤波法
  - 非负矩阵分解

# 谱减法

## 基本原理

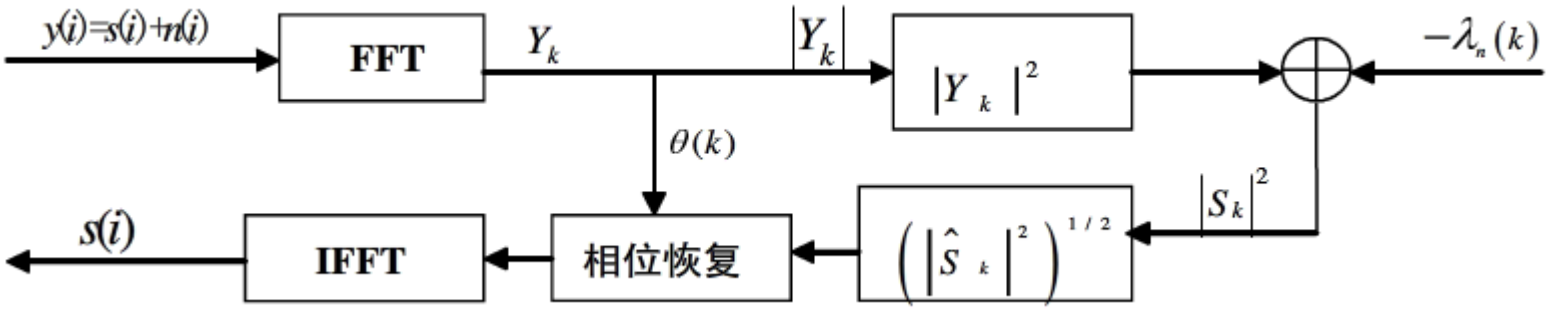
- 将含噪语音信号和VAD判别得到的纯噪声信号进行DFT变化，从含噪语音谱幅度特征中减掉纯噪声的幅度谱特征，得到增强的幅度谱特征，再借用含噪语音的相位进行IDFT变化，得到增强的语音。

## 谱减法假设

- 语音和噪声信号是线性叠加的
- 噪声是平稳的，噪声与语音信号不相关

# 谱减法

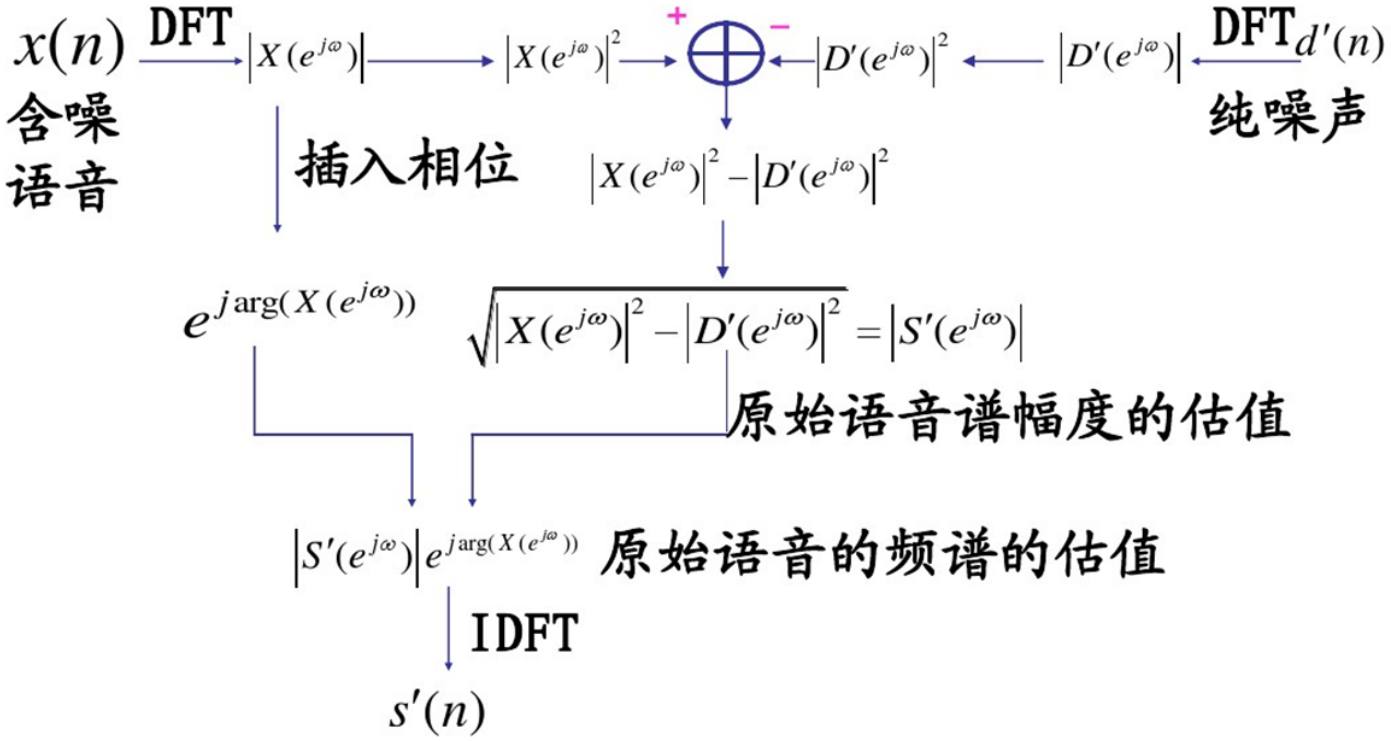
- 谱减法相当于对带噪语音的每一个频谱分量乘以一个系数。信噪比高时，含有语音的可能性大，衰减系数小；反之衰减系数大。



# 谱减法

幅值谱：增强，
相位谱：原始的相位来重构增强语音，因为相位对增强
增强语音等于原始相位谱和幅值谱联合生成增强语音，在做反傅里叶变换，生成语音

$$x(n)=s(n)+d(n)$$



# 谱减法

## 谱减法特点

- 原理简单，算法计算复杂度低
- 需要VAD判决，在信噪比大的情况下，使用短时平均能量等参数可以达到效果
- 当噪声特性变化时，降噪效果会变差
- 由于噪声的随机分布范围广，若语音帧某频率点上的噪声分量大时，就会有一部分残留噪声，它会损伤语音质量，降低语音可懂度

# 维纳滤波法

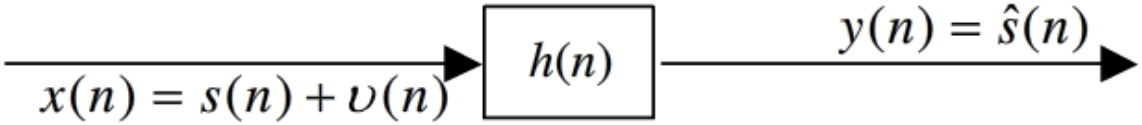
- 维纳滤波算法，由数学家Norbert Wiener提出，该方法是在最小均方误差准则下用Wiener滤波器实现对语音信号的估计，使带噪语音信号经过该滤波器的输出能够满足均方误差最小。维纳滤波算法的本质就是从噪声中提取信号的过滤和预测的方法。

- 在最小均方准则下用维纳滤波器实现对语音信号的估计，即对带噪语音信号y(t)=s(t)+n(t)，确定滤波器的冲击响应h(t)，使得带噪语音信号经过该滤波器的输出能够与s(t)的均方误差最小。

$$x(n) = s(n) + \upsilon(n)$$

$$y(n) = \sum_{m} h(m)x(n-m)$$

$$y(n) = \hat{s}(n)$$




$$\boldsymbol{x}^T = [x(n), x(n-1) \cdots x(n-M+1)]$$

# 维纳滤波法

- 采用最小均方误差法求解：

$$E\left[e^2(n)\right]=E\left[(s(n)-\sum_{m=0}^{N-1}h(m)x(n-m))^2\right]$$

- 对上式求偏导数，进一步变换可得：

$$\begin{bmatrix} R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(N-1) \\ R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(N-2) \\ \vdots & \vdots & \cdots & \vdots \\ R_{xx}(N-1) & R_{xx}(N-2) & \cdots & R_{xx}(0) \end{bmatrix}\begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(N-1) \end{bmatrix}=\begin{bmatrix} R_{xs}(0) \\ R_{xs}(1) \\ \vdots \\ R_{xs}(N-1) \end{bmatrix}$$

求解维纳滤波器系数问题等价于求解维纳-霍夫方程问题，确定h

# 维纳滤波法

## 维纳滤波法特点

- 计算复杂度低，满足实时性要求
- 算法要求输入信号具有平稳特性
- 算法要求带噪语音和安静语音存在线性关系
- 在处理非平稳噪声时，降噪效果会变差
- 在复杂环境下难以跟踪非平稳噪声变化轨迹

tips:平稳噪声：其统计特性不随时间变化的噪声

非平稳噪声：其统计特性随时间变化而变化的噪声

# 非负矩阵分解

- 原理：在矩阵中所有元素均为非负的条件下对其实现非负分解

- 假设处理m个n维空间的样本数据，用V表示。该矩阵中各个元素均为非负。对X进行线性分解

$$V \approx WH \quad V \in \mathbb{R}^{M \times N} \quad W \in \mathbb{R}^{M \times r} \quad H \in \mathbb{R}^{r \times N}$$

  - 其中W为基矩阵，H为系数矩阵
  - 任何一个矩阵可以分解为各个基矩阵W线性加权的形式

# 非负矩阵分解

- 在语音增强中对功率谱进行矩阵分解，分为两个步骤
  - 利用训练数据，离线训练得到最优的基矩阵，基矩阵由不同基矢量组成，各个基矢量的维度与功率谱维度相同（基矢量包括两类：语音基矢量和噪声基矢量）
  - 利用基矩阵，对测试语音逐帧（段）进行增强，通过在线更新矩阵系数H（各个基矢量权重），得到增强后的功率谱，并联合相位谱生成增强后的语音

# 非负矩阵分解

## 功率谱基矩阵确定过程

- 将训练数据功率谱特征组成矩阵V，迭代优化确定基矩阵和系数矩阵
- 其中基矩阵分为两部分 $W_S$ 和 $W_N$ ，分别对应安静语音和噪声
- 采用Multiplicative update准则进行迭代，确定基矩阵

$$H_i \leftarrow H_i \otimes \frac{W_i^T \frac{V_i}{W_i H_i}}{W_i^T 1} \qquad W_i \leftarrow W_i \otimes \frac{\frac{V_i}{W_i H_i} H_i^T}{1 H_i^T} \qquad W = [W_S \, W_N]$$

## 在线语音增强阶段

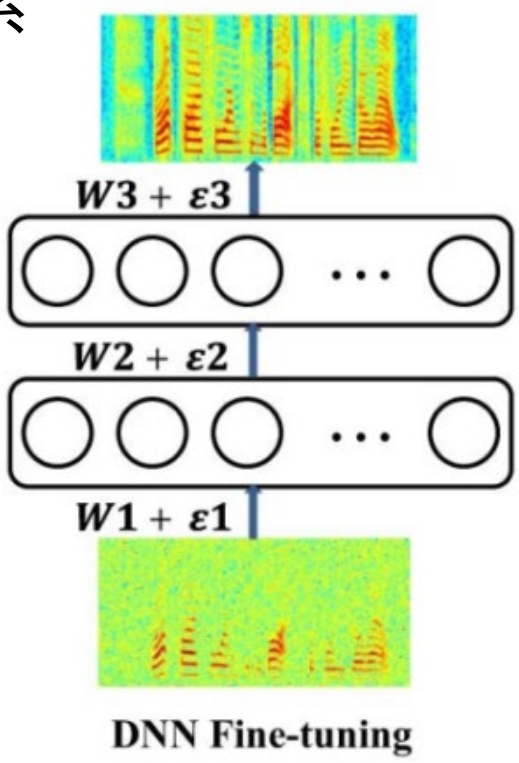
- 保持 W 固定，只迭代更新 H
- 带噪语音可以表示为安静语音基矢量和噪声基矢量的线性组合，只对语音基矢量加权，噪声基矢量权重置为0，生成增强语音

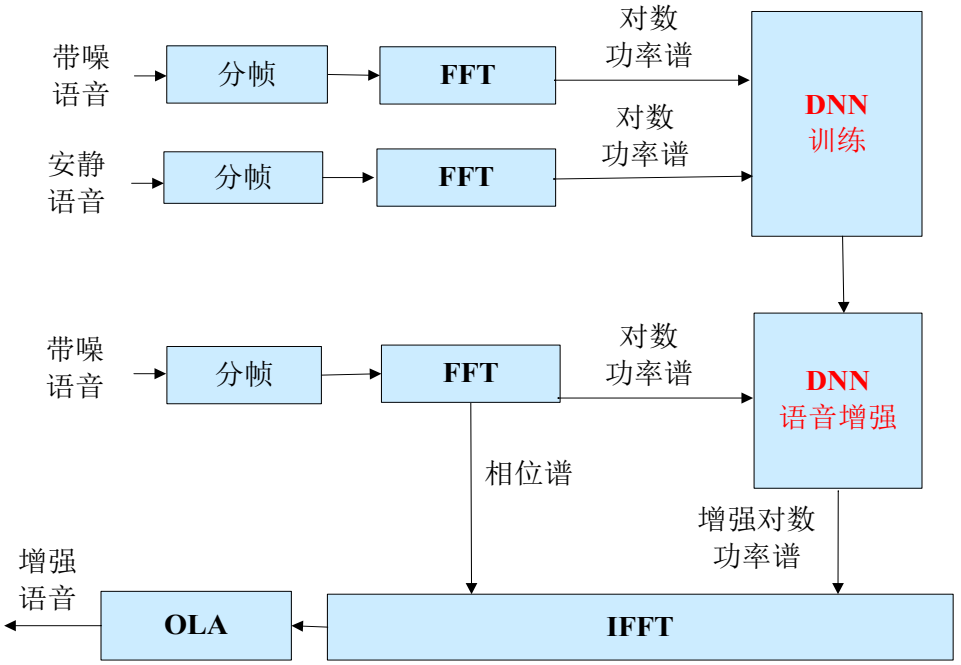
# 非负矩阵分解

- 非负矩阵分解特点：
  - 增强的谱参数通过语音参数基矢量加权得到，可以抑制过平滑问题
  - 建立的基矩阵可以通过扩帧来考虑相邻帧的特征，从而捕获噪声变化轨迹
  - 相对于其它数据驱动方法，不需要大数据进行训练
  - 算法计算复杂度高，实时性难以满足要求

# 基于深度学习的语音增强

- 算法原理框图：建立带噪语音与安静语音对数功率谱的映射关系

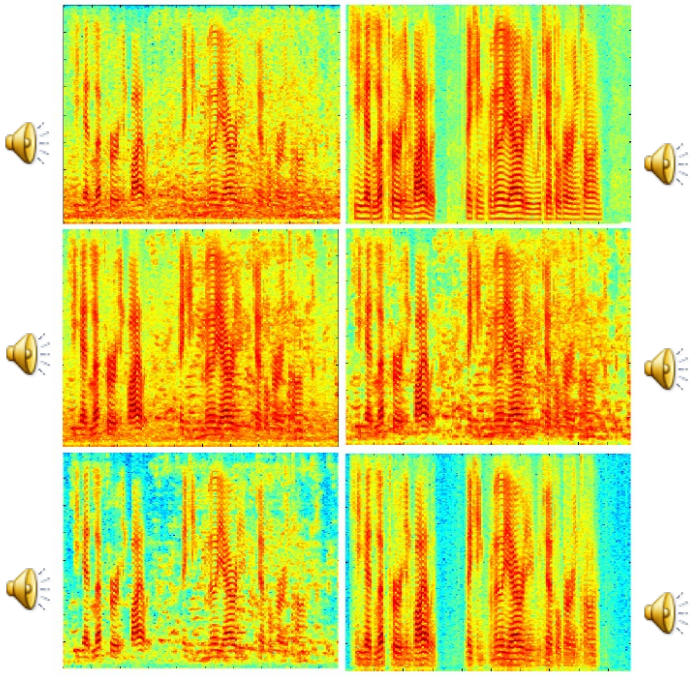



在增强阶段，输入是带噪语音提取对数傅里叶谱特征，经过DNN解码后，获得增强后的特征，然后利用带噪语音的相位信息，经过傅里叶逆变换得到增强后的语音波形。

# 基于深度学习的语音增强

- 语谱图对比分析：5dB信噪比babble噪声语谱图
  - 左上：带噪语音
  - 右上：安静语音
  - 左中：子带谱减法
  - 右中：维纳滤波法
  - 左下：logmmse法
  - 右下：深度学习方法



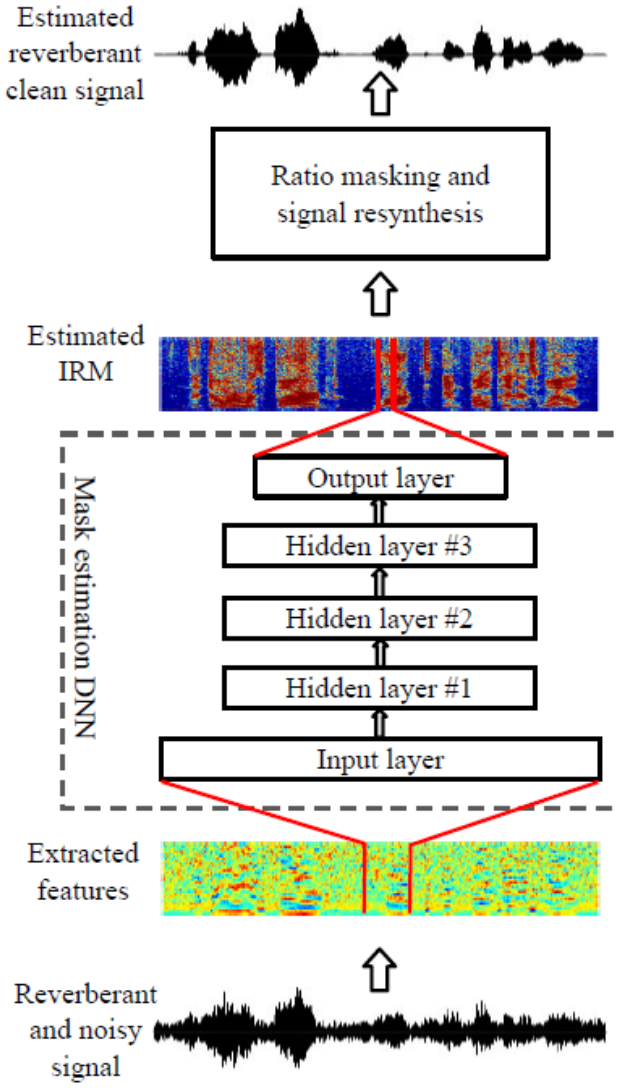
- 基于DNN的语音增强方法可以有效的抑制非平稳噪声

# 基于深度学习的语音增强

- 预测掩蔽值信息

模型的输出：二值型掩蔽值；对于二值型屏蔽值，如果某个时频单元能量是语音主导，则保留该时频单元能量，如果某个时频单元能量是噪声主导，则将该时频单元能量置零；

优势：共振峰位置处的能量得到了很好的保留，增强后的语音具有较高的可懂度；

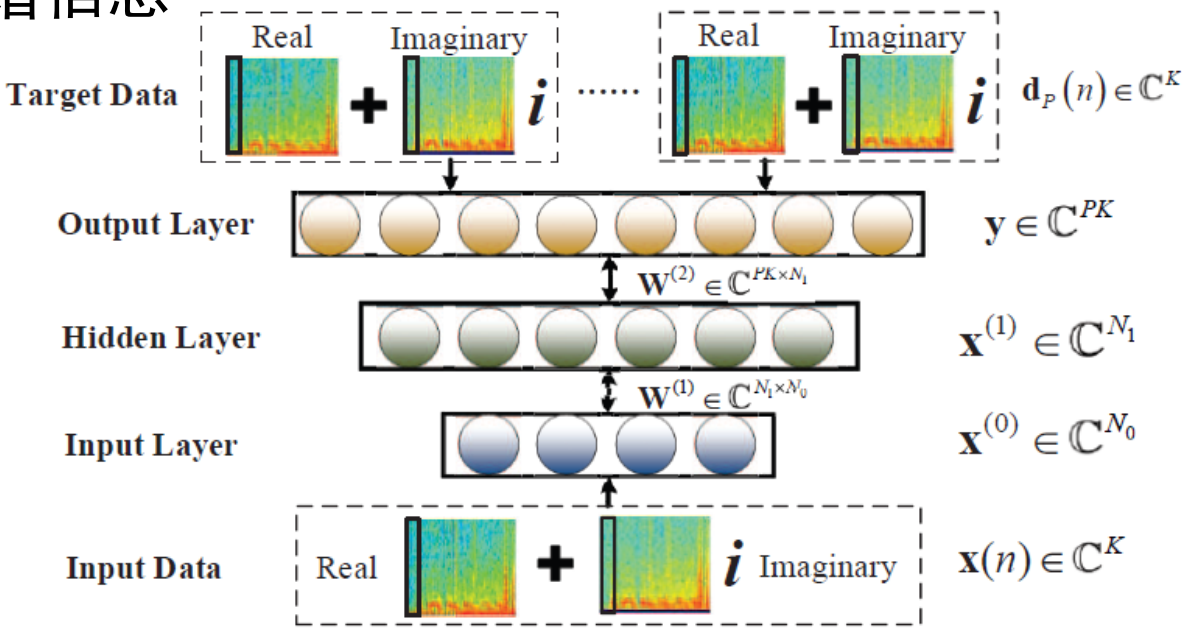



增强前

增强后

# 基于深度学习的语音增强

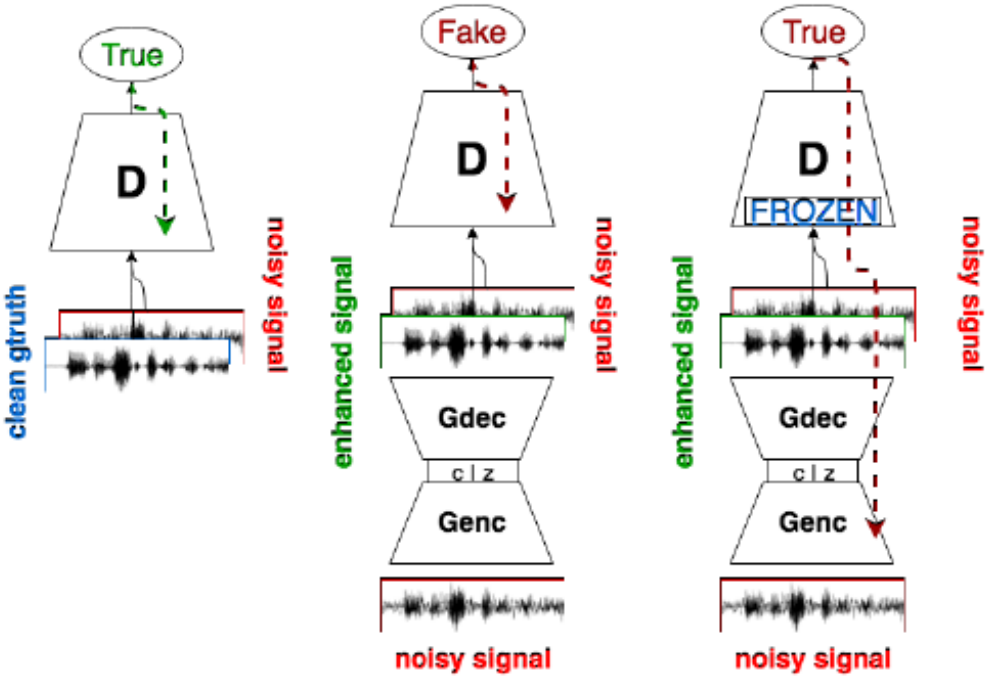
- 预测复数谱信息




目前主流的语音增强方法更多的关注于对幅值谱相关特征的增强而保留原始语音的相位谱；复数神经网络模型可以对复数值进行非线性变换，而语音帧的复数谱能够同时包含幅值谱信息和相位谱信息，可以通过复数神经网络建立带噪语音复数谱和干净语音复数谱的映射关系，实现同时对幅值信息和相位信息的增强。

# 基于深度学习的语音增强

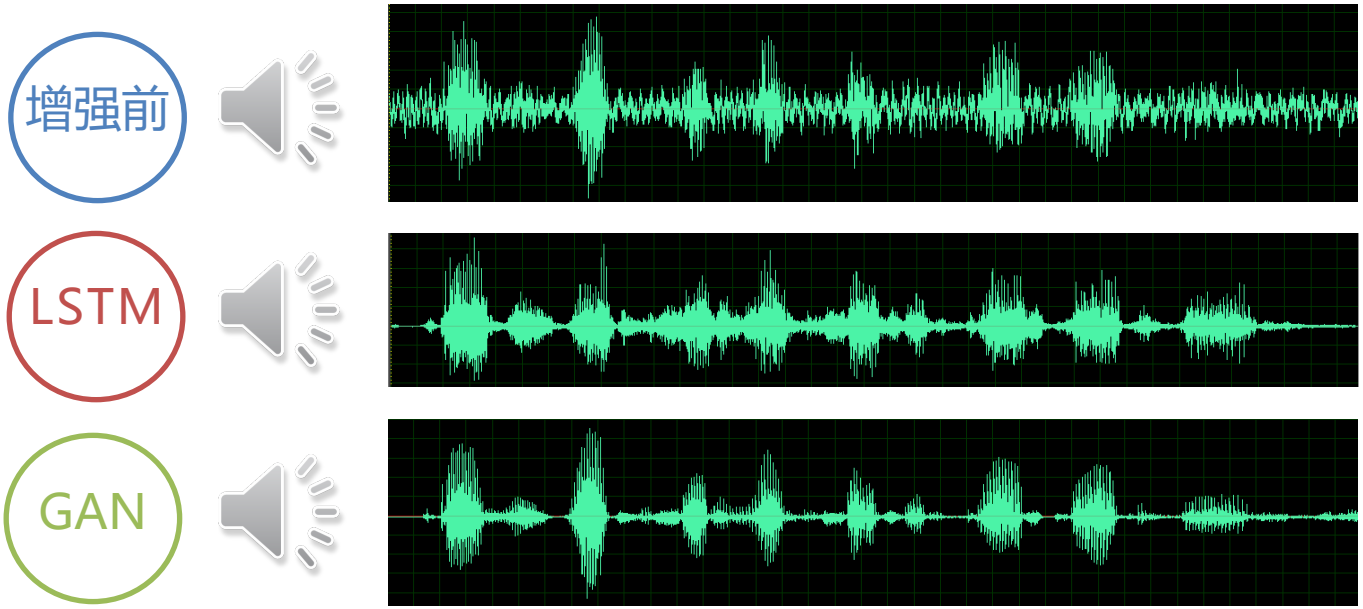
## 基于对抗网络的语音增强




直接处理原始音频的端到端方法。生成网络用于增强，它的输入是含噪语音信号和潜在表征信号，输出是增强后的信号。在训练过程中，鉴别器负责向生成器发送输入数据中真伪信息，使得生成器可以将其输出波形朝着真实的分布微调，从而消除干扰信号。

# 基于深度学习的语音增强

- 基于对抗网络的语音增强



# 鸡尾酒会问题

- 在同时有多个说话人说话时，人类可以将注意力集中到他感兴趣的声音上因此自然过滤掉其他声音，但是这个问题对于机器来说就很困难，这种现象被称为鸡尾酒会问题。
- 基于传统方法对解决鸡尾酒会问题的效果十分有限
  - 计算声学场景分析
  - 维纳滤波
  - 非负矩阵分解
  - . . .

# 鸡尾酒会问题

- 基于深度学习的方法：深度聚类（DPCL）
  - 假设对于混合语音的每一个时频(T-F)块只属于一个说话人。假如我们令相同颜色的块属于同一个说话人，那么对于每一个说话人可以通过对谱聚类来得到。
  - 定义输入特征|Y|，即为混合的幅值谱特征。网络的输出为一个嵌入式向量V。其目标函数为

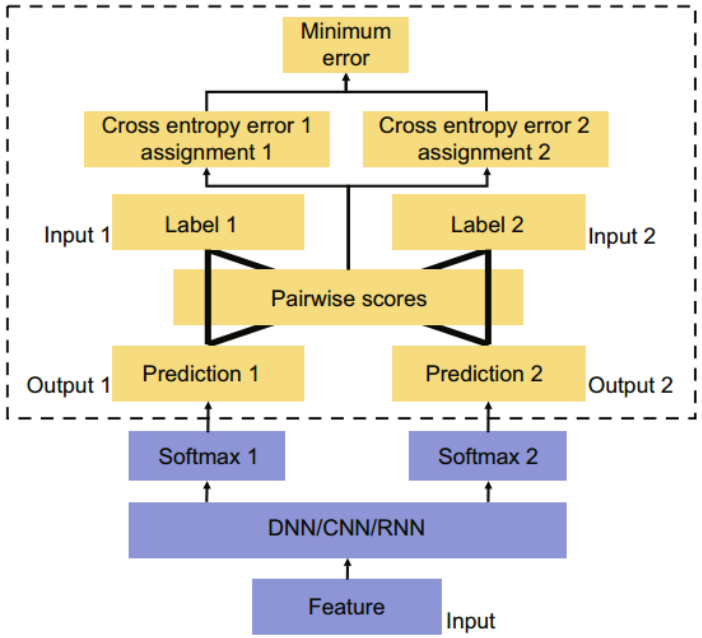
$$J = ||VV^T - BB^T||_F^2$$

# 鸡尾酒会问题

## 基于排列不变性训练的方法（PIT）

- 对于单通道说话人独立的语音分离，其难点在与输出的不确定性或者排列组合问题
- PIT处理语音分离作为一个多类回归问题，它提供了一个排列组合的集合而不是一个固定顺序的列表。
- 在训练阶段，PIT解决排列组合问题是通过选择一个最小均方误差(MSE)的排列来作为目标函数，定义如下：

$$J = \min\left(\frac{1}{F \cdot T \cdot S}\left\| \left|\widetilde{X_{S'}}\right| - \left|X_{S'}\right| \right\|\right) \; S' \epsilon \; permu(S)$$

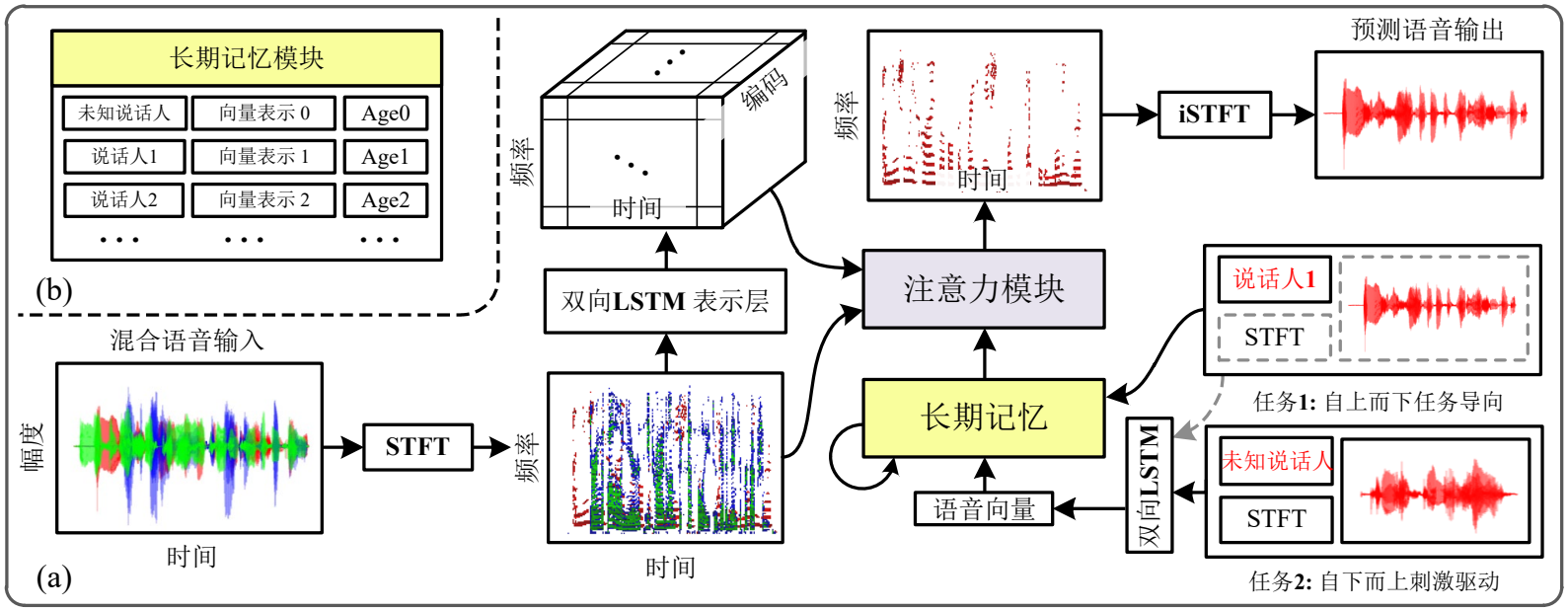


# 鸡尾酒会问题

- 基于注意和记忆的听觉注意模型(ASAM)
  - 自上而下任务导向型听觉注意是由主观目的导向的
    - 在酒会中与朋友进行聊天，我们会有意识选择性地倾听自己所熟知朋友的声音，而忽略其他人的声音
  - 自下而上刺激驱动型听觉注意是由显著音驱动的
    - 在酒会中忽然有人喊我们的名字，或者身边有玻璃杯打碎的声音时，我们会被当前场景下显著的声音所吸引而去关注该声音事件

# 鸡尾酒会问题

## 基于注意和记忆的听觉注意模型（ASAM）




将自上而下任务导向型听觉注意和自下而上刺激驱动型听觉注意整合到一个统一的框架中，提出基于注意和记忆的听觉注意模型

# 鸡尾酒会问题

- 基于注意和记忆的听觉注意模型（ASAM）
- 在自上而下任务导向型任务中，根据给定已知说话人的身份标签，从长期记忆中提取声纹特征进行注意力计算。
- 而在自下而上刺激驱动型任务中，根据给定的显著音片段，通过双向LSTM提取声纹特征进行注意力计算。注意力计算的公式如下：$\alpha_{t,f} = \mathrm{sigmod}(g^T \cdot \tanh(W \cdot v + U \cdot h_{t,f}))$

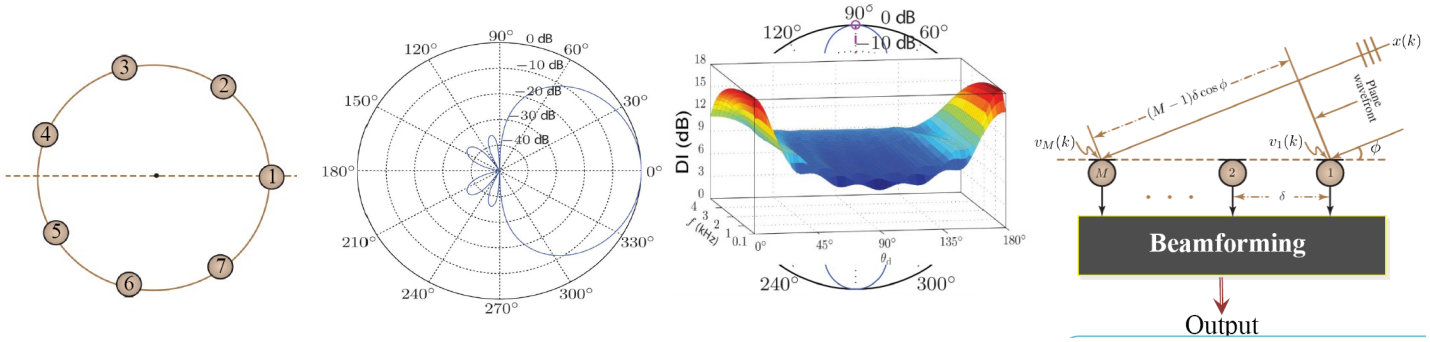
  其中$v \in \mathbb{R}^{d\times 1}$是声纹特征，$h_{t,f} \in \mathbb{R}^{d\times 1}$是时频单元$X_{t,f}$的d维向量表示，$g \in \mathbb{R}^{d\times 1}$ $W \in \mathbb{R}^{d\times d}$和$U \in \mathbb{R}^{d\times d}$是可学习参数。
- 给定目标语音语谱时，模型的目标函数是：

$$J = \sum_{t,f} \left\| S_{t,f} - X_{t,f} \times \alpha_{t,f} \right\|_2^2$$

# 多通道语音降噪

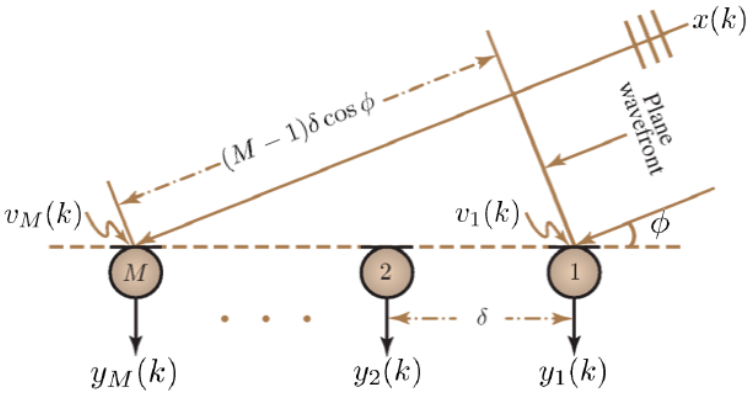
- 多通道语音降噪指有多个麦克风收集到的语音信号。多通道语音降噪的方法主要包括：延时求和、最小方差无失真响应、广义旁瓣滤波和数据驱动的多通道语音降噪方法。

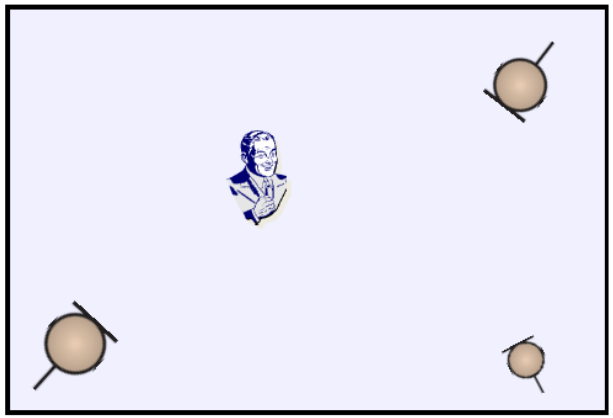
- 多通道语音增强算法通常需要跟麦克风阵列的几何结构适配，在智能语音领域比较常用的麦克风阵列结构包括线阵、环阵、双环阵等。



# 多通道语音降噪

- 阵列系统与分布式网络：



几何尺寸已知

麦克风响应一致

同步采样

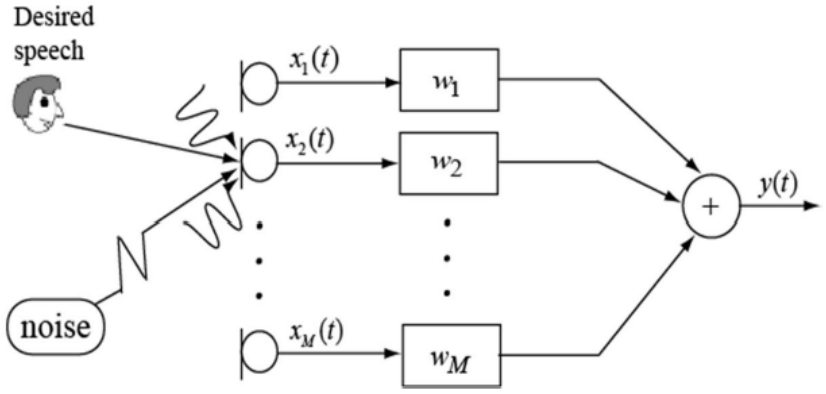
几何尺寸未知或时变

麦克风频响可能不同

异步采样

# 多通道语音降噪

- 多个麦克风同时采集语音:




$$Y(k,\ell)=\sum_{i=1}^{M}W_i(k,\ell)\times X_i(k,\ell)=\underbrace{\sum_{i=1}^{M}W_i(k,\ell)\times A_i(k,\ell)\times S(k,\ell)}_{f(k,\ell)}+\underbrace{\sum_{i=1}^{M}W_i(k,\ell)\times N_i(k,\ell)}_{Y_v(k,\ell)}$$

$f(k,\ell)$= 语音成分变换函数

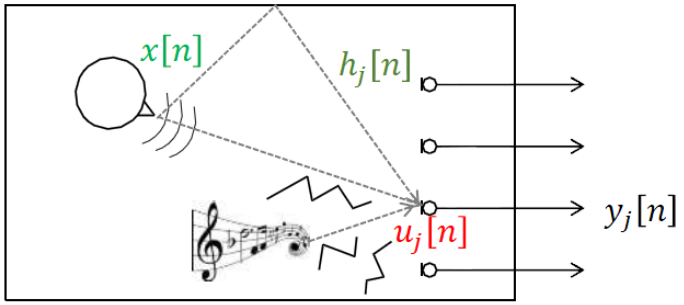
$Y_v(k,\ell)$= 残留的噪声成分

图中为麦克风阵列采集语音的示意图，各个通道的信号通过滤波器加权融合，Y为多通道融合增强后的语音，可以将其分解为两部分：目标语音成分和残留噪声成分；残留噪声成分可以通过后置滤波算法进一步处理，也可以通过改进麦克风阵列波束形成算法使这一成分得到有效抑制。

# 多通道语音降噪

## 多通道语音处理模型（时域）

- 麦克风阵列采集语音




- 第j个麦克风的表达式为：

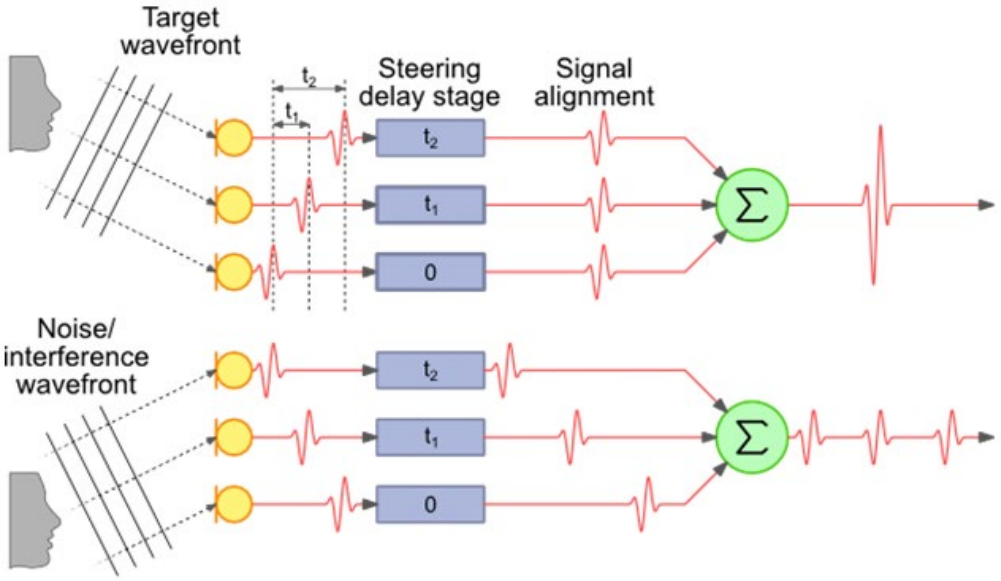
$$y_j[n] = \sum_l h_j[l]x[n-l] + u_j[n] = h_j[n] * x[n] + u_j[n]$$

- $x[n]$ Target clean speech
- $h_j[n]$ Room impulse response
- $u_j[n]$ Additive noise (background noise, ...)
- $n$ Time index

x表示安静语音，h表示房间响应函数，u表示其它噪声干扰。接下来介绍的算法将更多的侧重于对噪声源u的抑制。

# 多通道语音降噪

- 波束形成法：建立空间滤波器模型，它的作用包括
  - 将多个麦克风采集的信号进行同步，生成单通道信号
  - 只增强目标方向的信号，对其它方向的信号进行抑制



# 多通道语音降噪-波束形成法

- 空间滤波器的分类：
  - 线性滤波器：
    - 时不变线性滤波

$$Y(t,f) \Rightarrow \hat{X}(t,f) = W^*(f)Y(t,f)$$

  - 非线性滤波器
    - 时变线性滤波器

$$Y(t,f) \Rightarrow \hat{X}(t,f) = W^*(t,f)Y(t,f)$$

    - 非线性变换

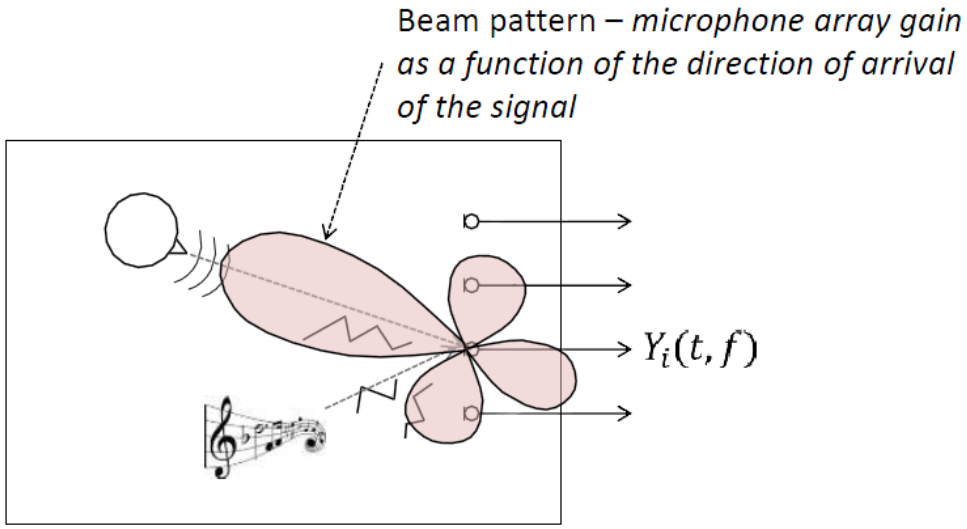
$$Y(t,f) \Rightarrow \hat{X}(t,f) = F(Y(t,f))$$

波束形成算法需要解决的核心问题是估计空间滤波器W，
输入：麦克风阵列采集的多通道语音信号
输出：增强后的单路语音信号。
对空间滤波器进一步细分，可以分为时不变线性滤波、时变线性滤波以及非线性变换模型。最简单的延时求和法属于时不变线性滤波，广义旁瓣滤波法属于时变线性滤波，基于深层神经网络的波束形成属于非线性变换模型。

# 多通道语音降噪–波束形成法

## 波束方向图

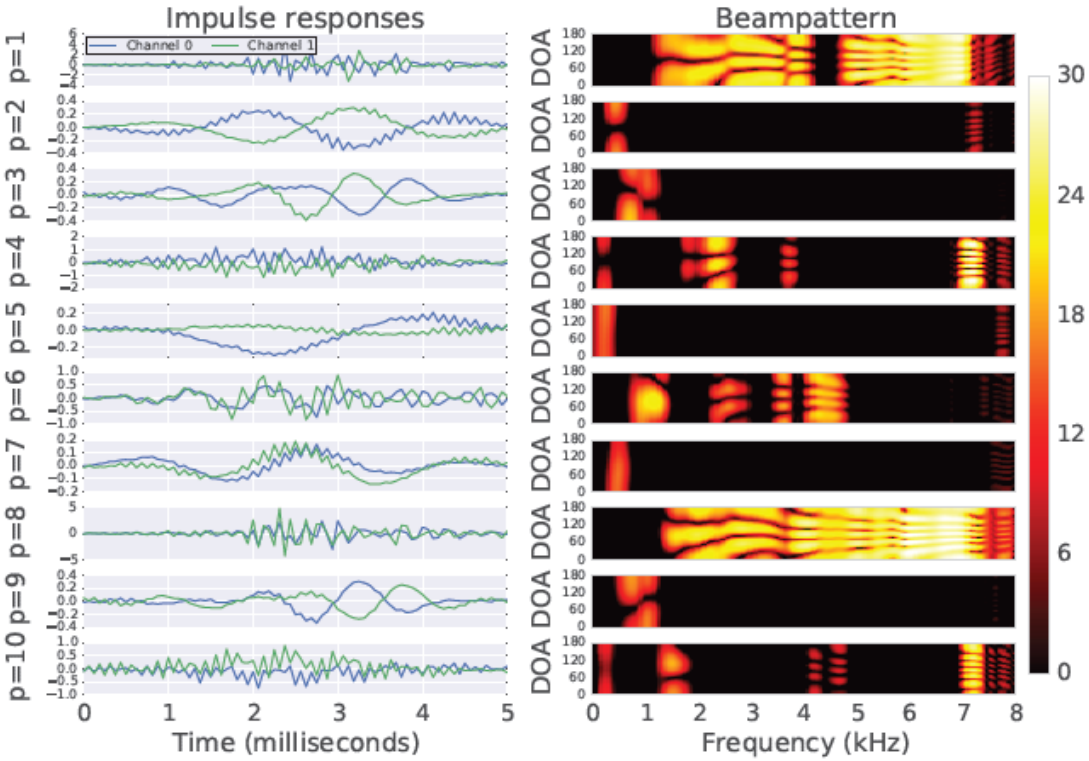
- 拾取目标方向声音
- 抑制非目标方向声音以及无向噪声




上图是一个麦克风阵列算法在f频带上所对应的波束方向图，不同频带对应不同的波束方向图。在声源方向已知的条件下，可以采用固定波束形成的思路，通过设计波束方向图，确定空间滤波器的系数，增强目标方向的声音。

# 多通道语音降噪-波束形成法

- 空间滤波器的时域和频域特征
  - 左侧为时域信号，右侧为频域响应
  - 对于不同频带以及同一频带下不同角度响应均有差异

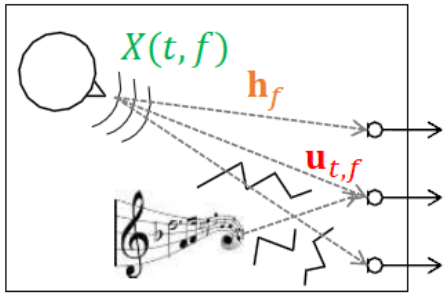


# 多通道语音降噪–麦克风阵列模型

## 模型解析：噪声模型与导向矢量

$$Y_j(t,f) \approx \sum_m H_j(m,f)X(t-m,f) + U_j(t,f)$$
$$= \underbrace{H_j(f)X(t,f)}_{O_j(t,f)} + U_j(t,f)$$

source image at microphone $j$




$$\mathbf{y}_{t,f} = \begin{bmatrix} Y_1(t,f) \\ \vdots \\ Y_J(t,f) \end{bmatrix} = \underbrace{\mathbf{h}_f \; X(t,f)}_{\triangleq \mathbf{o}_{t,f}} + \mathbf{u}_{t,f}$$

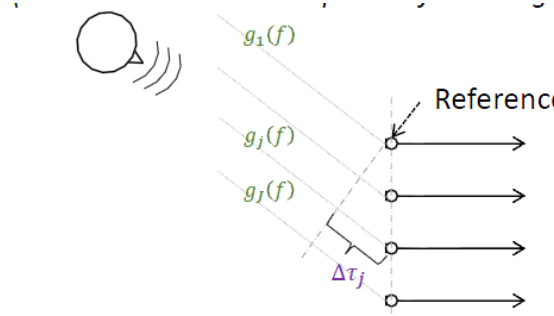
$\mathbf{o}_{t,f}$ Source image at microphones

$\mathbf{h}_f = [H_1(f), \ldots, H_J(f)]^{\mathrm{T}}$ Steering vector

式中Y表示各个麦克风接收到的信号，绿色部分表示声源信号，橙色部分表示声源信号传输到麦克风的变换，红色部分表示各种噪声源的干扰。因此波束形成算法需要在已知Y的条件下，尽可能准确的估计h和u；即估计导向矢量和噪声模型。

# 导向矢量

- 描述从声源到麦克风传输过程中延时、衰减等
- 下图为自由场条件下的平面波模型




$$Y_j(t,f) = g_j(f)e^{-2\pi f\Delta\tau_j}X(t,f) + U_j(t,f)$$

$\Delta\tau_j$ time difference of arrival (TDOA), frequency independent.

$g_j(f)$ signal gain, frequency dependent.

$U_j(t,f)$ reverberation and noise

Steering vector given as :

$$\mathbf{h}_f = \begin{bmatrix} g_1(f)e^{-2\pi f\Delta\tau_1} \\ \vdots \\ g_J(f)e^{-2\pi f\Delta\tau_J} \end{bmatrix}$$

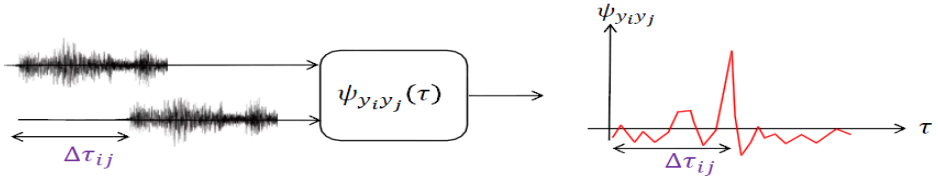
导向矢量是麦克风阵列算法中最为重要的参数，能够反映声源传输的方向性信息，用于描述从声源到麦克风传输过程中延时、衰减等特性；
数学表达式中紫色部分表示声源到达各个麦克风的时间差，绿色部分表示声源向麦克风传输过程中的衰减，导向矢量主要跟这两个因素有关；
对导向矢量进一步处理也可以对声源方位信息进行估计。

# 导向矢量

## TODA估计

- 计算互相关函数峰值

$$\Delta\tau_{ij} = \arg\max_{\tau} \psi_{y_i y_j}(\tau)$$
$$\psi_{y_i y_j}(\tau) = E\{y_i(t) y_j(t+\tau)\}$$



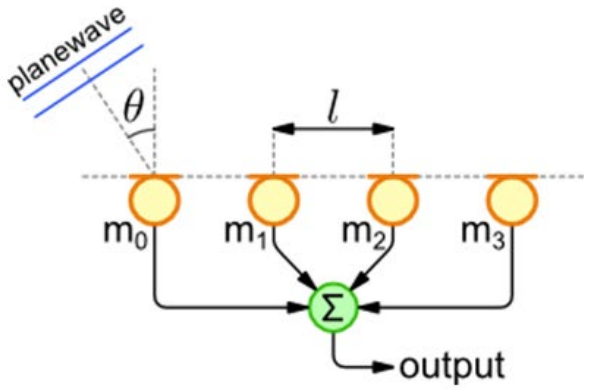

- 进一步提高鲁棒性：GCC-PHAT 系数

$$\psi_{y_i y_j}^{PHAT}(\tau) = IFFT\left(\frac{Y_i(f)Y_j^*(f)}{\left|Y_i(f)Y_j^*(f)\right|}\right)$$

可以通过广义互相关函数确定各个麦克风之间的相对延时，如图所示，寻找广义互相关函数中的峰值点，通过峰值点的位置计算出相对延时。为了进一步提高TDOA估计的鲁棒性，可以采用GCC-PHAT方法，这种方法在已有方法基础上引入了能量归一化机制。

# 延时求和方法

- 基于Delay Sum的波束形成算法
  - 利用麦克风之间的几何关系确定同步后的单通道信号
  - 下图以正弦信号为例进行说明，其中N表示麦克风个数，l是麦克风间距， 是声音入射角




$$output = 20\log_{10}\left(\frac{1}{N}\sum_{i=0}^{N-1} e^{\frac{j\,2\pi f\,i\,l\,\sin(\theta)}{c}}\right)$$

# 延时求和方法

- Delay Sum方法特点：
  - 算法基于外部环境不存在无向噪声和混响的假设设计的
  - Delay Sum算法输出的单通道语音，需要增加后置滤波模块，抑制噪声和混响干扰
  - 算法计算复杂度低，便于实时应用

# 最小方差失真响应波束形成

在保持信号不变的约束下使噪声输出功率最小

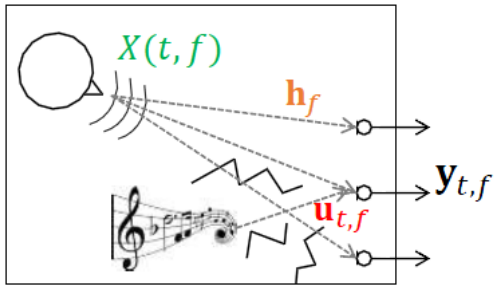
- Beamformer output:

$$\hat{X}(t,f) = \mathbf{w}_f^H \mathbf{y}_{t,f} = \mathbf{w}_f^H \left( \mathbf{h}_f \, X(t,f) \right) + \mathbf{w}_f^H \mathbf{u}_{t,f}$$

Speech $X(t,f)$ is unchanged (distortionless): $\mathbf{w}_f^H \mathbf{h}_f = 1$

Minimize noise at the output of the beamformer

$$\Rightarrow \hat{X}(t,f) = X(t,f) + \mathbf{w}_f^H \mathbf{u}_{t,f}$$




- Filter is obtained by solving the following:

$$\mathbf{w}_f^{MVDR} = \underset{\mathbf{w}_f}{\operatorname{argmin}} \, E\{|\mathbf{w}_f^H \mathbf{u}_{t,f}|^2\}, \quad \text{subject to } \mathbf{w}_f^H \mathbf{h}_f = 1,$$

y表示多通道语音，w表示空间滤波器，x表示增强后的单通道语音，这种波束形成算法的假设是期望方向上的语音无失真，也就是w*h这项为1；同时保证对噪声的响应最小，也就是最小化w*u这项。在这两个约束条件下估计最优的空间滤波器w。

# 最小方差失真响应波束形成

- MVDR滤波器：

$$\mathbf{w}_f^{MVDR} = \frac{\left(\mathbf{R}_f^{noise}\right)^{-1}\mathbf{h}_f}{\mathbf{h}_f^H\left(\mathbf{R}_f^{noise}\right)^{-1}\mathbf{h}_f}$$

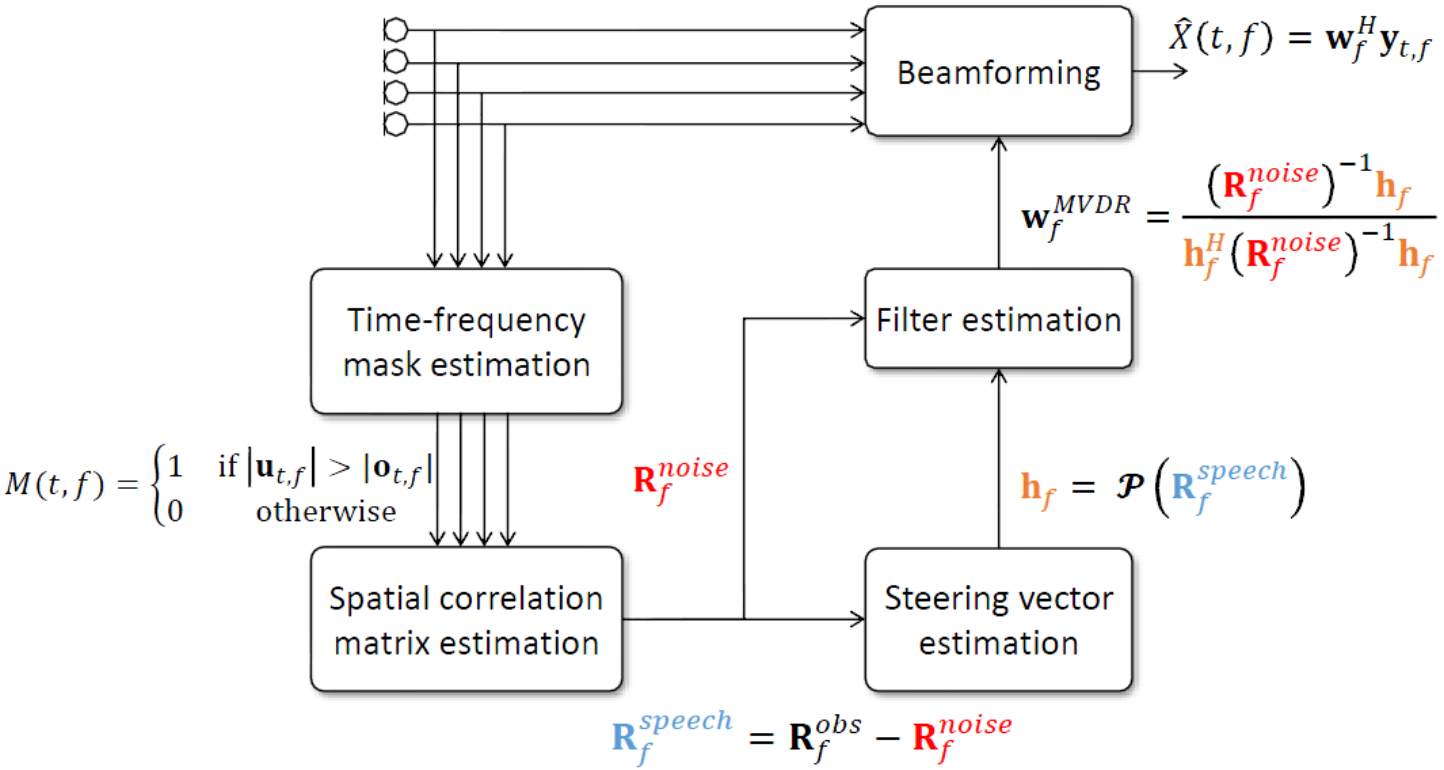
- 噪声相关矩阵：

$$\mathbf{R}_f^{noise} = \sum_t \mathbf{u}_{t,f}\mathbf{u}_{t,f}^H = \begin{bmatrix} \frac{1}{T}\sum_t^T U_1(t,f)U_1^*(t,f) & \cdots & \frac{1}{T}\sum_t^T U_1(t,f)U_J^*(t,f) \\ \vdots & \ddots & \vdots \\ \frac{1}{T}\sum_t^T U_J(t,f)U_1^*(t,f) & \cdots & \frac{1}{T}\sum_t^T U_J(t,f)U_J^*(t,f) \end{bmatrix}$$

经过一系列的变换和推倒，我们能够确定空间滤波器w与噪声协方差矩阵和导向矢量的关系。为了计算噪声协方差矩阵，需要估计出各个通道中信号在各个频带上噪声成分的互相关系数，因此对噪声成分的有效估计将直接影响到波束形成算法的性能。对于导向矢量，可以通过估计声源到达各个麦克风的相对延时来确定。

# 多通道语音降噪-最小方差失真响应波束形成

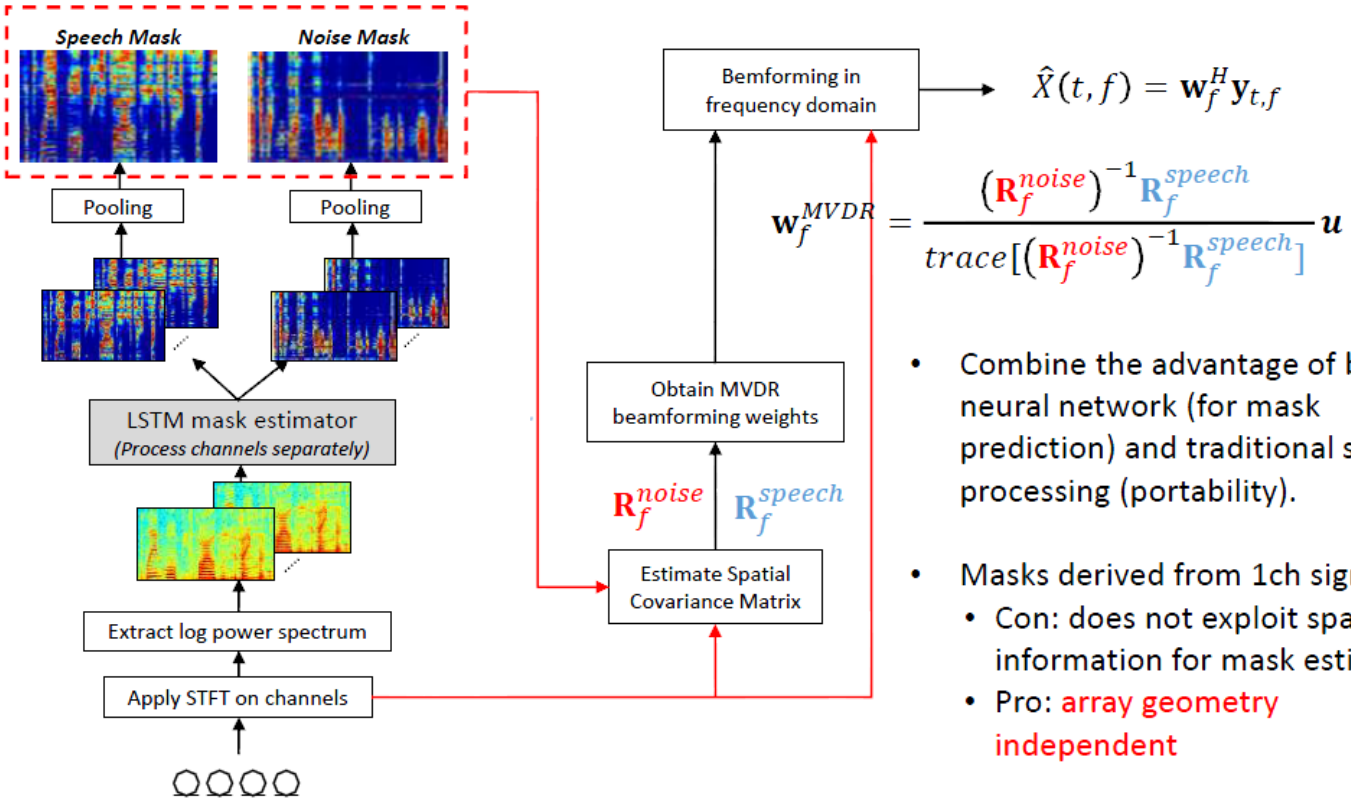
## 算法流程：




图中更详细的描述了重点介绍的基于最小方差失真响应波束形成的流程，对各个通道语音首先进行掩蔽值估计，然后计算噪声协方差矩阵和语音协方差矩阵，进一步确定导向矢量，通过导向矢量和噪声协方差矩阵估计空间滤波器，生成波束形成后的单通道语音。

# 多通道语音降噪

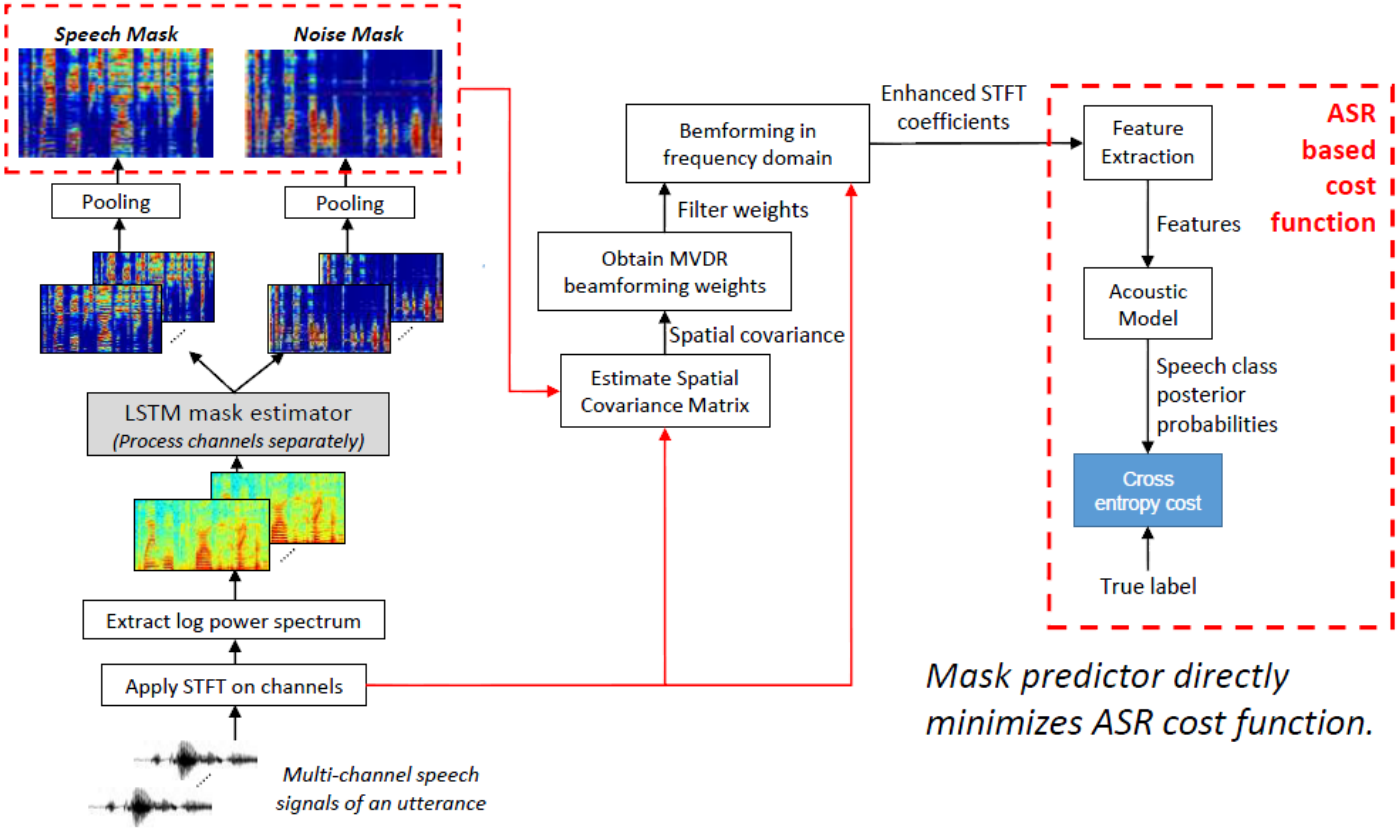
## 基于深度学习的波束形成方法




- Combine the advantage of both neural network (for mask prediction) and traditional signal processing (portability).
- Masks derived from 1ch signals
  - Con: does not exploit spatial information for mask estimation
  - Pro: array geometry independent

为了提高模型的泛化能力，更常用的方法是采用深层神经网络模型对各个通道各个频带的掩蔽值进行估计、融合，进而计算出噪声协方差矩阵，然后再跟传统的波束形成方法对接，如图所示的方法是将深层神经网络方法跟最小方差失真响应波束形成方法对接。

# 多通道语音降噪




采用这种基于深度学习的方法，可以有效的抑制噪声的干扰，提高增强语音的质量。增强后的语音可以输入到语音识别系统，提高语音识别的鲁棒性。

# 本节课程总结

- 语音增强的定义
  - 阐述了语音增强的定义以及语音增强的应用范围
- 语音增强的任务
  - 从语音增强的声学环境出发，分析语音增强需要实现的功能
- 语音增强技术概述
  - 从语音增强的任务出发，分析语音增强所用的技术原理及技术特点

# 推荐书籍

- Speech enhancement theory and practice, Philipos C. Loizou，2007 .
- Microphone Arrays: Signal Processing Techniques and Applications (Digital Signal Processing) by Michael Brandstein, Darren Ward, Springer, 2001.

# 谢谢！