鲁棒语音识别

李军锋

中科院语言声学与内容理解重点实验室
中国科学院声学研究所

本讲主要内容

- 影响语音识别性能的环境变化因素
- 噪声环境下的鲁棒语音识别技术(前端)
- 自适应技术（后端）

现实中的识别环境

中国科学院声学研究所
Institute of Acoustics, CAS

在现实生活中，很多语音识别的情景并不是干净的声学环境。如右图所示，在一个会议转录场景下，噪声、混响、多说话人始终干扰着识别系统的正常工作。

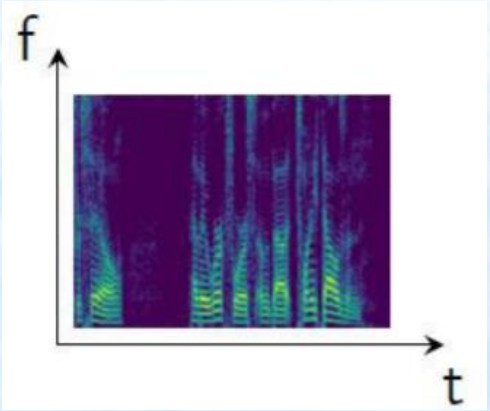
能听清楚在说什么吗

针对这些问题，在语音增强领域发展了噪声抑制、混响消除、多说话人分离这些领域。而针对语音识别，如何鲁棒地在复杂的现实环境中（会议、车载、家居、户外等）运行是最大的挑战。

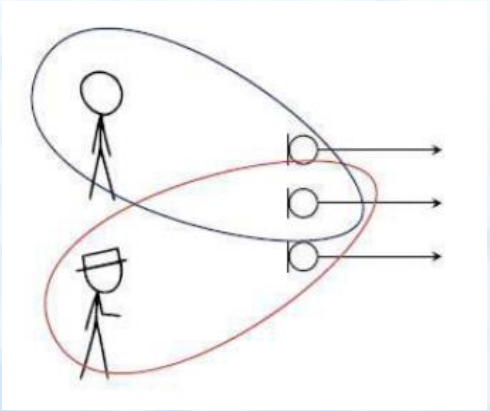
语音增强与鲁棒语音识别

中国科学院声学研究所
Institute of Acoustics, CAS

- 语音增强侧重于听感质量上的提升，语音识别侧重于识别率的提升
- 两者息息相关，但又有所不同。因为人的大脑对不同形式语音谱的理解大大超过计算机，听感的提升在部分情况下未必能在识别系统上有所体现。
- 如果实现语音增强，以多说话人为例（经典的鸡尾酒会问题）：



不同人有不同的声音特性（如基频、语调等），转化为模式识别问题



不同人站在不同的位置说话，转化为空间滤波问题（保留指定方向，抑制其他方向）

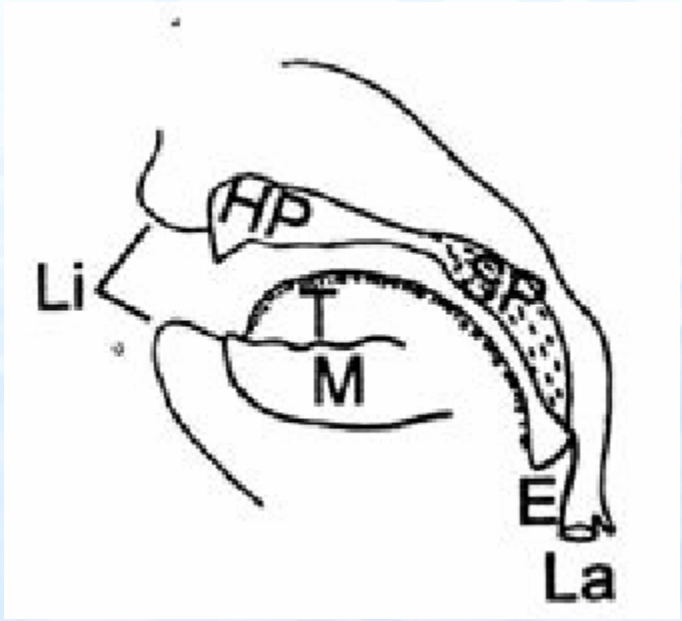
本讲主要内容

- HMM语音识别框架回顾
- 影响语音识别性能的环境变化因素
- 噪声环境下的鲁棒语音识别技术
- 自适应技术

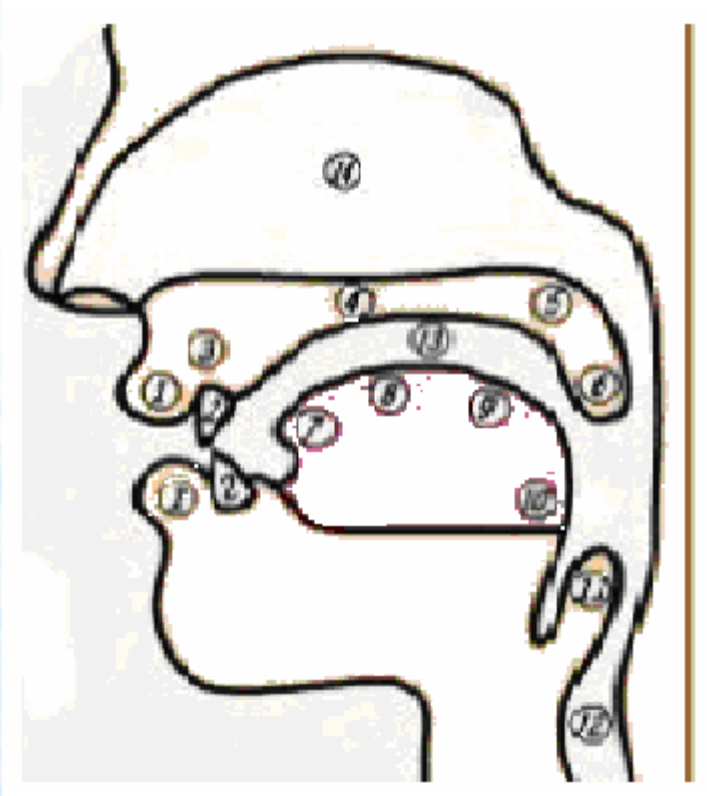
影响语音识别性能的环境变化因素

中国科学院声学研究所
Institute of Acoustics, CAS

- **存在的问题**
  - 说话人差异



（注：Li：唇；HP：硬腭；T：舌；M：下颚；SP：软腭；E：会厌；La：喉）



（注：⑴唇；⑵齿；⑶齿龈；⑷硬腭；⑸软腭；⑹小舌；⑺舌尖；⑻舌前；⑼舌后；⑽舌根；⑾咽头；⑿声带；⒀口腔；⒁鼻腔）

影响语音识别性能的环境变化因素

- **存在的问题**
  - 说话人差异
    - 年龄
    - 性别
    - 身体状况
    - 情绪/心理状况
    - 地域
    - ……

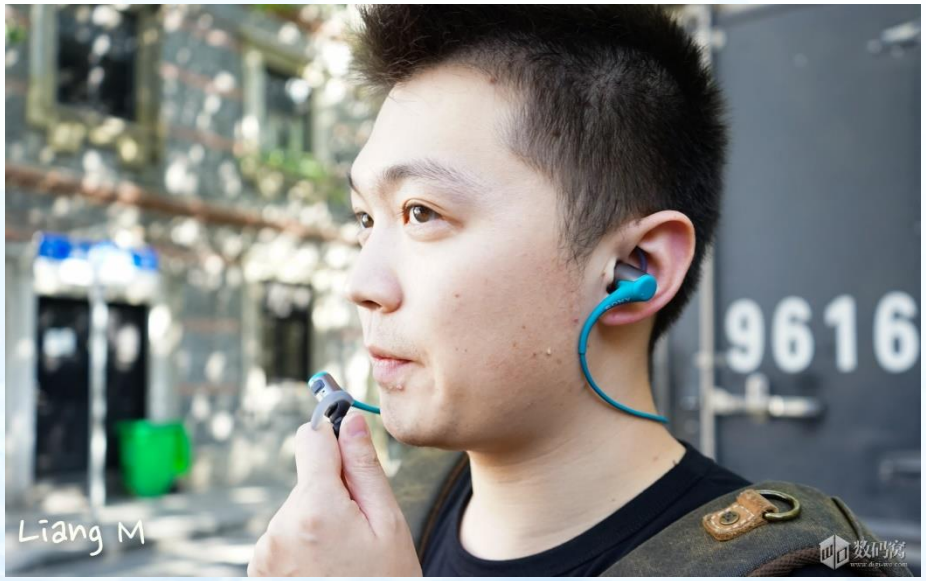
影响语音识别性能的环境变化因素

- 存在的问题
  - 声学环境差异

影响语音识别性能的环境变化因素

中国科学院声学研究所
Institute of Acoustics, CAS

- 存在的问题
  - 传播信道差异



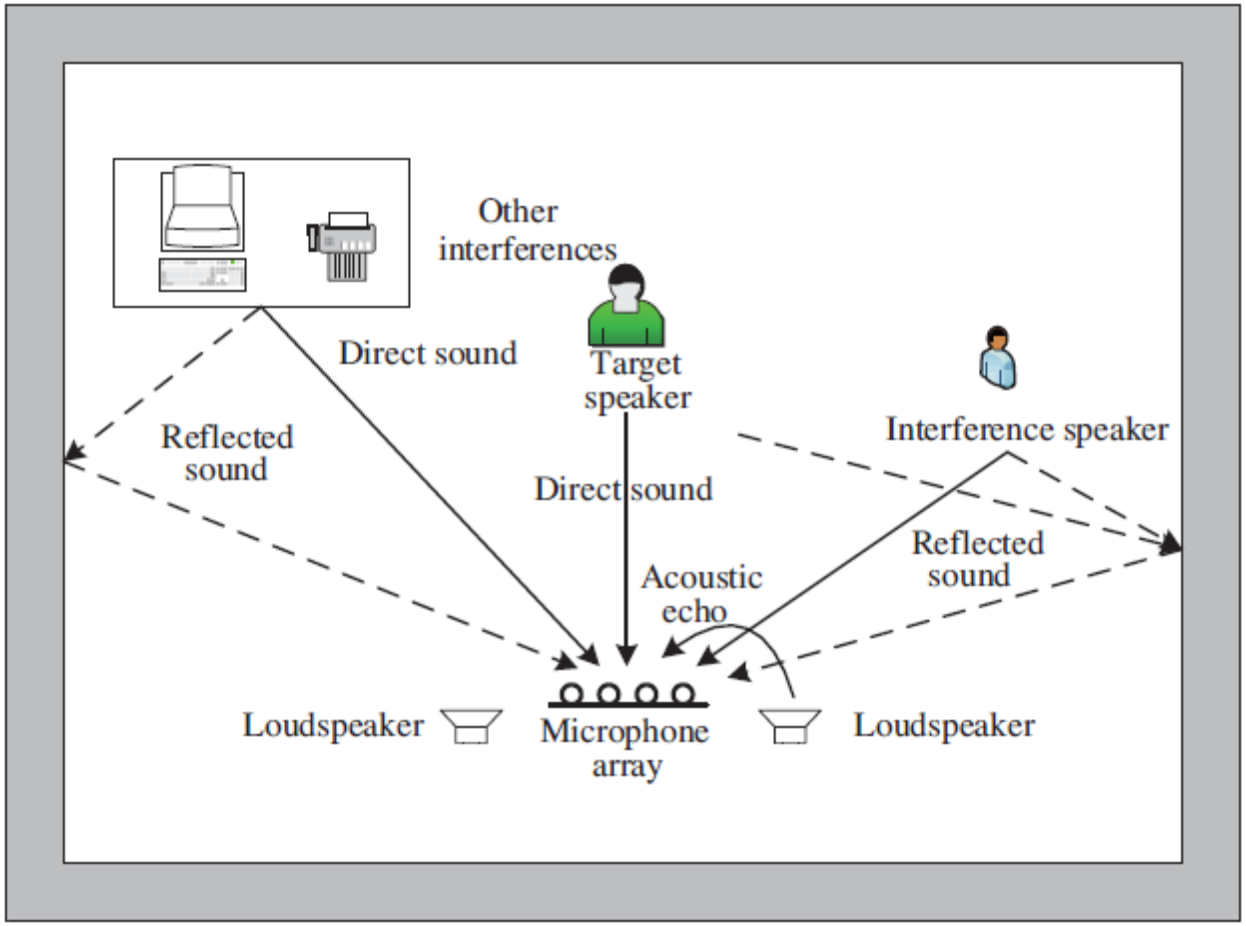
影响语音识别性能的环境变化因素

中国科学院声学研究所
Institute of Acoustics, CAS

- 存在的问题
  - 采音设备的差异

影响语音识别性能的环境变化因素

- 存在的问题
  - 近讲/远讲的差异



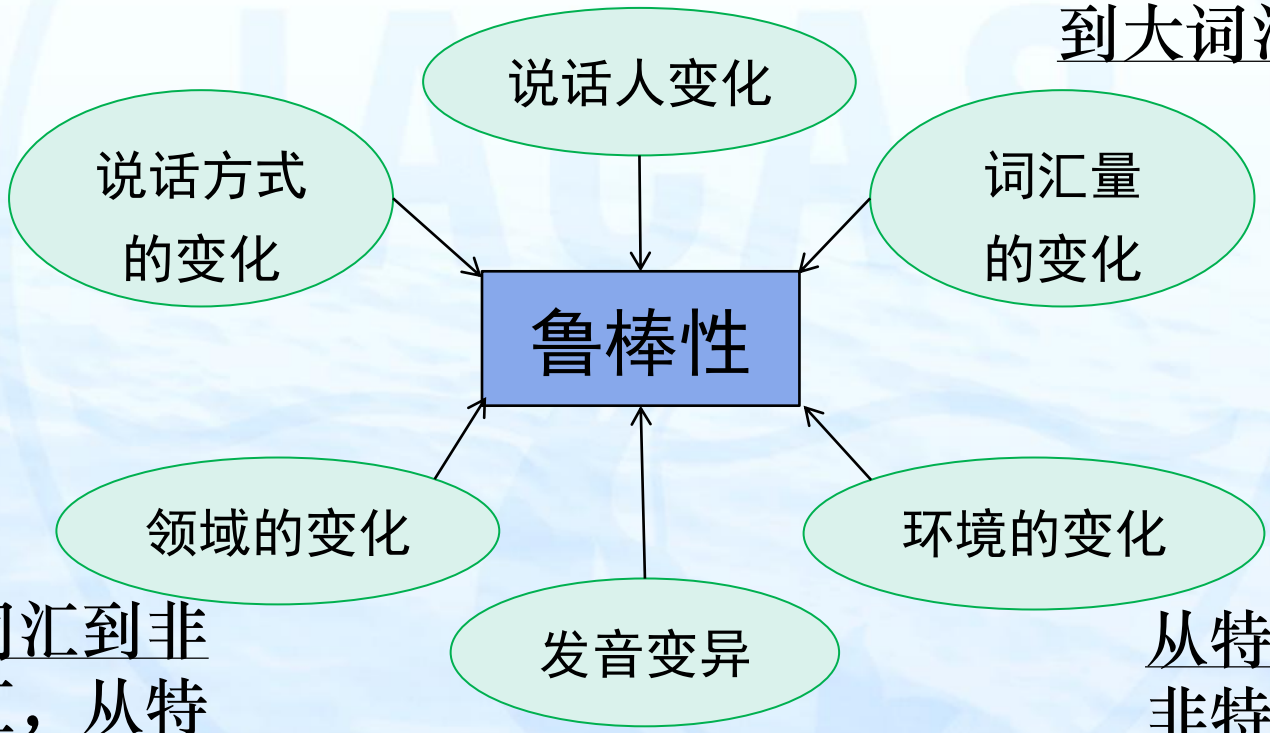
影响语音识别性能的环境变化因素

中国科学院声学研究所
Institute of Acoustics, CAS

从孤立词识别到连续语音识别

从特定说话人到非特定说话人

从小词汇量任务到大词汇量任务



从特定词汇到非特定词汇，从特定领域文法到非特定领域文法

话者由于受到生理、心理、情感等影响而产生的发音变化

从特定环境到非特定环境

影响语音识别性能的环境变化因素

- 加性噪声
- 通道畸变
- 其他因素
  - ✓ 人为因素
  - ✓ 瞬间噪声
  - ✓ 来自其他话者的干扰

本讲主要内容

- HMM语音识别框架回顾
- 影响语音识别性能的环境变化因素
- 噪声环境下的鲁棒语音识别技术
- 自适应技术

噪声环境下的鲁棒语音识别技术

- 基于语音增强方法
  - ✓ 谱减的方法
  - ✓ 维纳滤波

噪声环境下的鲁棒语音识别技术

- 通道畸变的抑制方法
  - ✓ 倒谱均值减CMS方法
  - ✓ 二级CMS的方法
  - ✓ RASTA-PLP技术

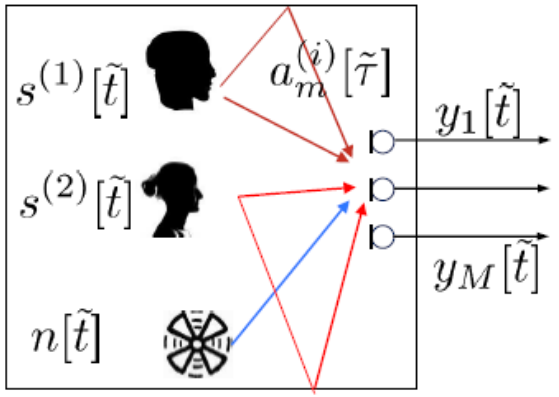
噪声环境下的鲁棒语音识别技术

- 基于模型的补偿方法
  - ✓ 多重风格训练
  - ✓ HMM分解
  - ✓ 并行模型融合PMC
  - ✓ 矢量泰勒级数VTS方法
  - ✓ 基于最小分类错误准则的补偿
  - ✓ 基于最小分类错误的环境特征学习

信号模型与优化目标

中国科学院声学研究所
Institute of Acoustics, CAS

- 时域信号成分
  - $s^{(i)}[\tilde{t}]$ 第i个源信号
  - $a_m^{(i)}[\tilde{\tau}]$ 第i个源到第m个麦克风的房间脉冲响应
  - $n[\tilde{t}]$ 噪声



- 麦克风拾取到的观测信号（对于单通道M=1）

标量形式

$$y_m[\tilde{t}] = \sum_{i=1}^{I} \left(\sum_{\tilde{\tau}=0}^{L-1} a_m^{(i)}[\tilde{\tau}] s^{(i)}[\tilde{t} - \tilde{\tau}] \right) + n_m[\tilde{t}]; \quad m = 1, \ldots, M$$

矢量形式

$$\mathbf{y}[\tilde{t}] = \sum_{i=1}^{I} \left(\sum_{\tilde{\tau}=0}^{L-1} \mathbf{a}^{(i)}[\tilde{\tau}] s^{(i)}[\tilde{t} - \tilde{\tau}] \right) + \mathbf{n}[\tilde{t}]; \quad \mathbf{y}[\tilde{t}] = \begin{pmatrix} y_1[\tilde{t}] \\ \ldots \\ y_M[\tilde{t}] \end{pmatrix}$$

信号模型与优化目标

中国科学院声学研究所
Institute of Acoustics, CAS

从干扰中提取出每个说话人的语音

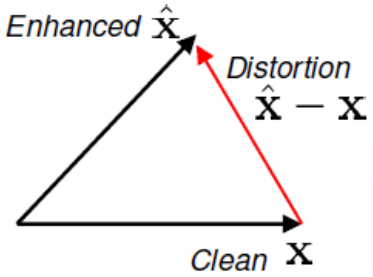
- 源分离：分离出混合语音中的声源
  - 降噪：削弱信号中的噪声部分（分离出语音源）
  - 说话人分离：分离出不同的说话人
- 去混响：削弱信号中的混响部分
- 说话人追踪：检测出不同说话人及说话的时间

$$\mathbf{y}[\tilde{t}] = \sum_{i=1}^{I} \left(\sum_{\tilde{\tau}=0}^{L-1} \mathbf{a}^{(i)}[\tilde{\tau}] s^{(i)}[\tilde{t} - \tilde{\tau}] \right) + \mathbf{n}[\tilde{t}];$$

信号评价指标

- 信号：信号失真比 SDR等

$$SDR = 10\log_{10}\frac{\sum_t|\boldsymbol{x}[t]|^2}{\sum_t|\hat{\boldsymbol{x}}[t]-\boldsymbol{x}[t}$$



优点：充分使用了全频段的信息

缺点：对缩放和相位十分敏感；不能直接反应语音识别性能的提升

- 感知质量：PESQ（语音质量）、STOI（语音可懂度）

优点：针对人的听感知

缺点：只对部分的语音失真敏感

- 语音识别：WER

优点：不需要语音对

缺点：依赖于ASR系统

语音增强的方法

单通道和多通道采用不同的方法，但大都有以下思路

- 统计信号处理
- 深度学习
- 结合统计和深度学习的方法

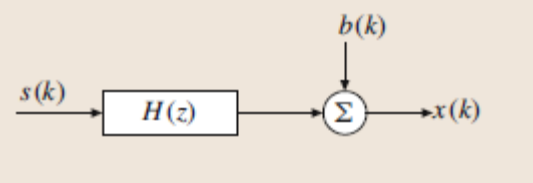
以源分离任务为例

单通道语音增强

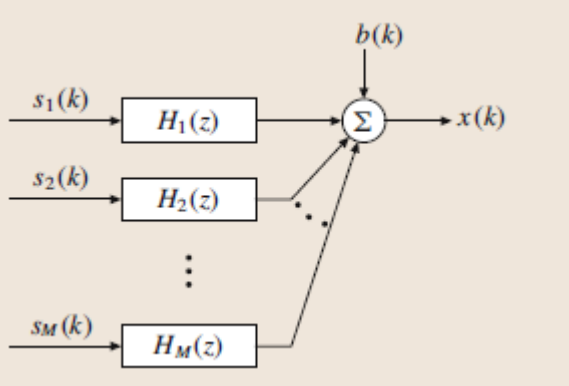
单通道语音增强

中国科学院声学研究所
Institute of Acoustics, CAS

- 统计信号处理——以降噪为例
  1. 计算语音信号和噪声的统计性质（如功率谱等）
  2. 应用滤波器（如维纳滤波）



Single-input Single-output system (SISO)



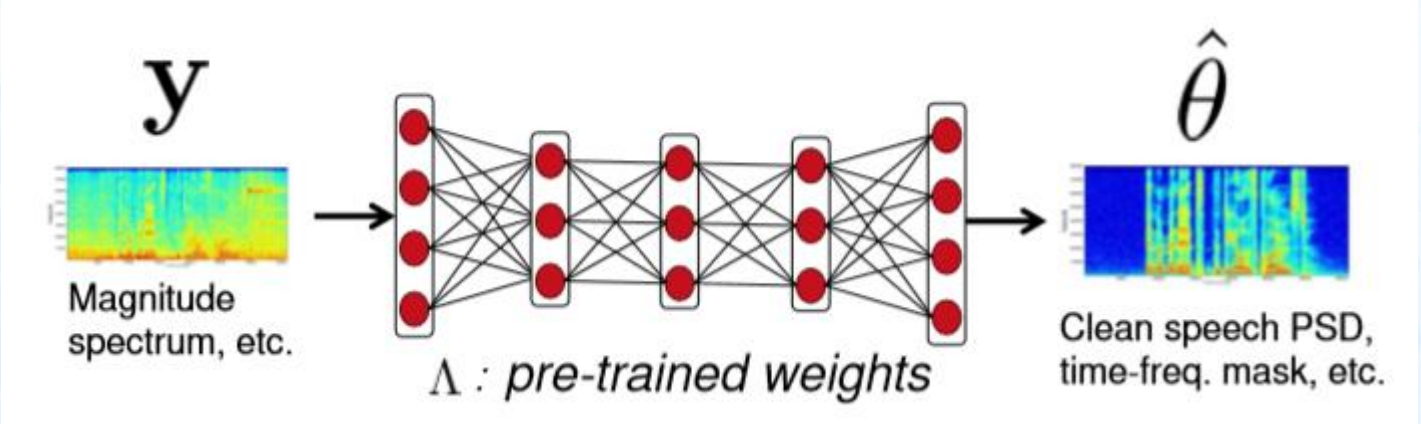
Multiple-input Single-output system (MISO)

Source: Chapter 6 in Springer handbook of speech processing

单通道语音增强——神经网络

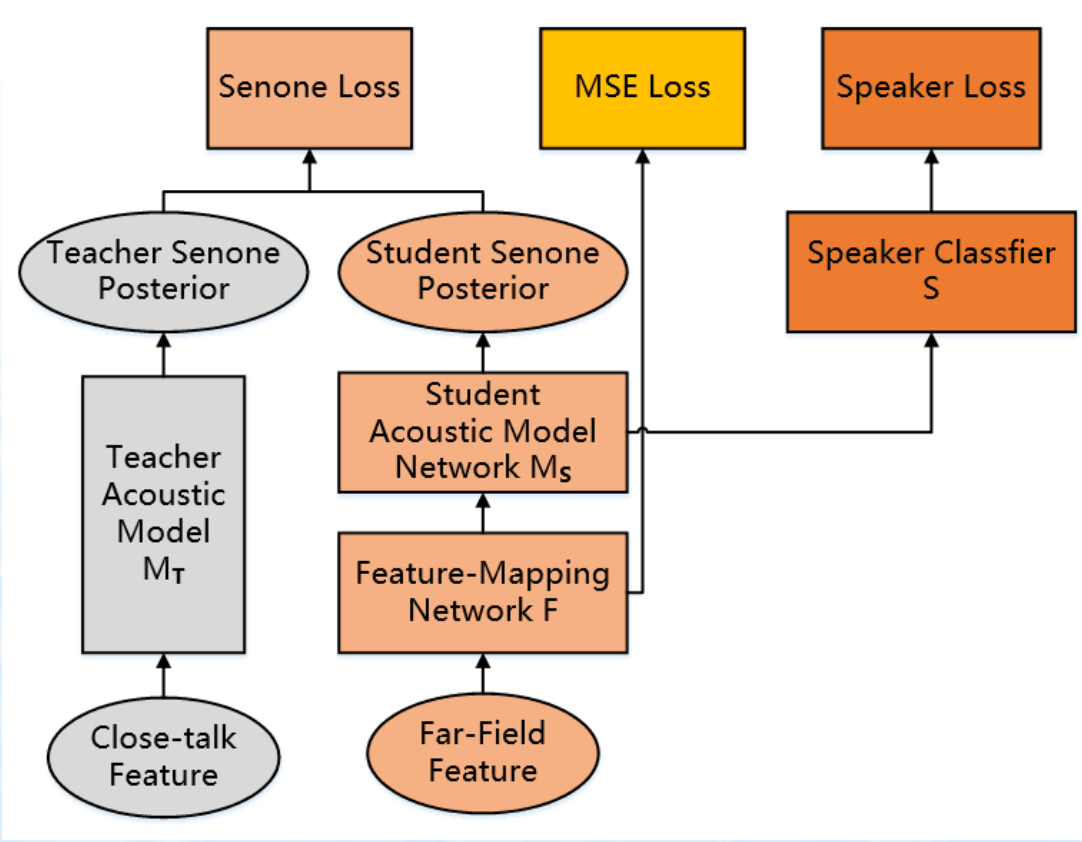
中国科学院声学研究所
Institute of Acoustics, CAS

- 大量的训练数据——数据对（pair）
- 训练神经网络
- 输出为特征或者掩膜



单通道语音增强——前后端联合优化

中国科学院声学研究所
Institute of Acoustics, CAS



Feature Mapping Network
将远场映射为近场，使用MSE loss
Teacher-student Model
使远场模块学习近场的概率分布
Speaker classifier
使用对抗性思想，使学生网络学习“说话人不变”的性质

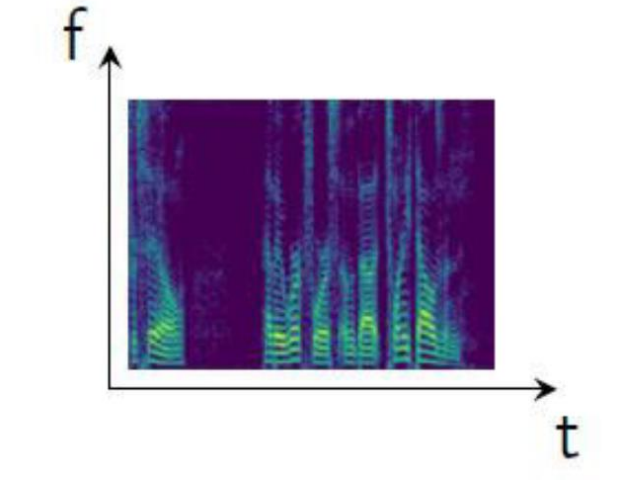
Speaker-Invariant Feature-Mapping for Distant Speech Recognition via Adversarial Teacher-Student Learning
Long Wu, Hangting Chen, Li Wang, Pengyuan Zhang, Yonghong Yan//Interspeech 2019

以源分离任务为例，用麦克风阵列接受信号

多通道语音增强

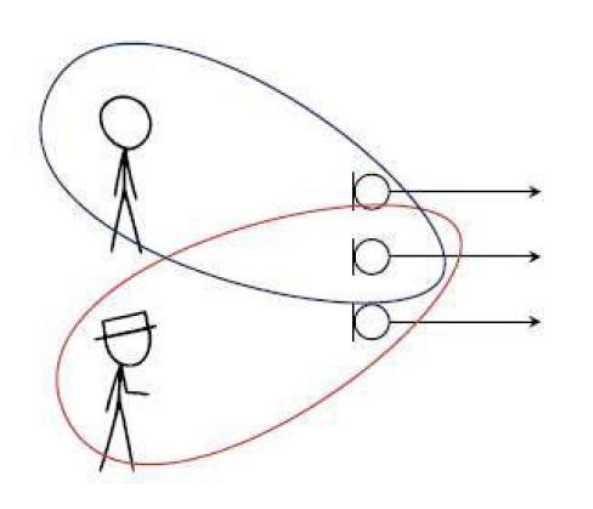
多通道语音增强

中国科学院声学研究所
Institute of Acoustics, CAS



Spectro-temporal

- 同样存在于单通道，不同得说话人/音素有不同的时间-强度信息
- 特定模型建模声源的内在统计性质



Spatial

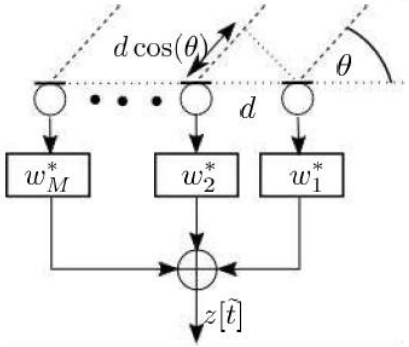
- 不同通道间的相位和强度差，代表了声源的空间信息
- 不建模声源的统计信息，可以对任何声源使用

多通道语音增强

中国科学院声学研究所
Institute of Acoustics, CAS

麦克风阵列信号处理——波束形成

1. 合适的波束形成可以同时做到抑制混响、噪声，并能分离不同方向的声源
2. 波束形成又称为空间滤波，因为他能加强/抑制某一个方向的信号
3. 但是大部分的波束形成都需要一些已知的空间信息（如房间冲击相应、



$$s[\tilde{t}] = \mathrm{e}^{j\omega_0\tilde{t}} = \mathrm{e}^{j\frac{2\pi c}{\lambda_0}\tilde{t}}$$

Signal at mth microphone:

$$x_m[\tilde{t}] = s[\tilde{t} - \tilde{\tau}_m] = \mathrm{e}^{j\omega_0(\tilde{t}-\tilde{\tau}_m)}$$

$$\tilde{\tau}_m = \frac{(m-1)d\cos\theta}{c};\ m = 1, \ldots, M$$

Beamformer output:

$$z[\tilde{t}] = \sum_{m=1}^{M} w_m^* x_m[\tilde{t}]$$

$$= \ldots$$

$$= \mathrm{e}^{j\omega_0\tilde{t}}\mathbf{w}^{\mathrm{H}}\mathbf{v}(\theta, \lambda_0)$$

Beamformer coeff.: $\mathbf{w} = [w_1, \ldots, w_M]^{\top}$

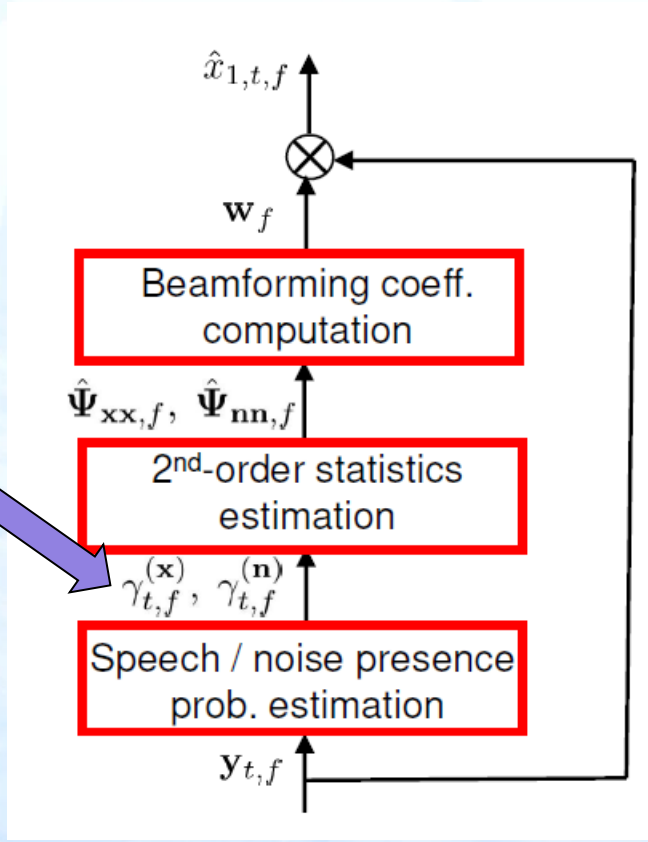
Steering vector: $\mathbf{v}(\theta, \lambda_0) = \left(1 \quad \mathrm{e}^{-j2\pi\left(\frac{d}{\lambda_0}\right)\cos(\theta)} \quad \ldots \quad \mathrm{e}^{-j2\pi\left(\frac{d}{\lambda_0}\right)\cos(\theta)(M-1)}\right)$

波束形成与神经网络结合

中国科学院声学研究所
Institute of Acoustics, CAS

- 基本流程

1. 输入多通道观测信号
2. 由神经网络获得参数
3. 生成式模型（CGMM、cACGMM等）
4. 波束形成
5. 输出干净语音



语音分离与提取技术

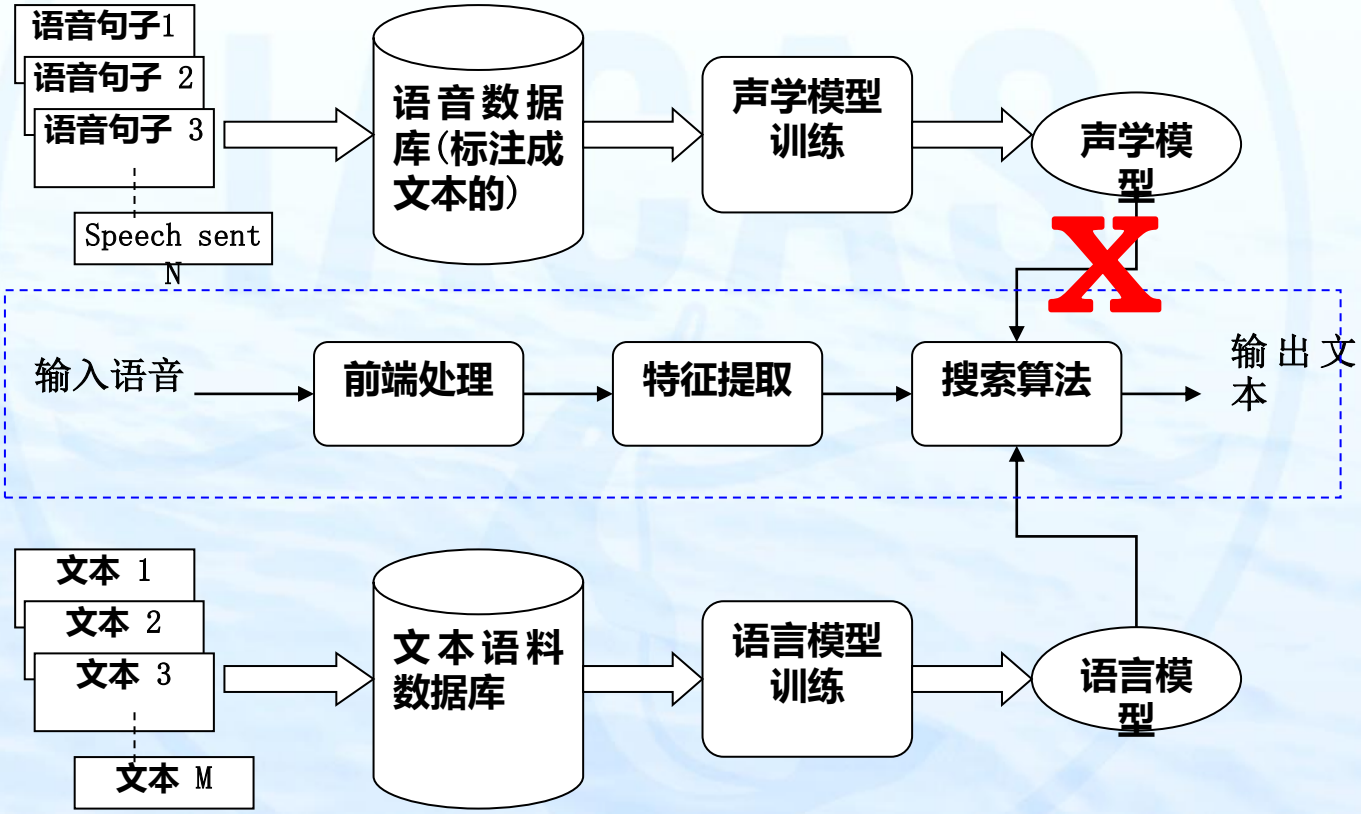
- 传统算法
  - 计算听觉场景分析：CASA（D.L.Wang et al.,2006）
  - 非负矩阵分解：NMF（M.N.Schmidt et al.,2006）
  - 基于生成模型的方法：GMM-HMM（Z.Ghahramani et al., 1997）
- 基于深度学习的算法
  - 语音分离角度
    - 深度聚类：DPCL（J.R.Hershey et al.,2016）
    - 深度吸引子网络：DAnet（Chen Z. et al.,2017）
    - 排列不变训练网络：PIT（Yu D. et al.,2017）
    - 时域分离网络：TasNet（Luo Y. et al.,2018）
  - 说话人相关语音提取角度
    - 深度提取网络：DEnet（Wang J. et al.,2018）
    - Speaker beam（Delcroix M. et al.,2018）
    - Voicefilter（Wang Q. et al.,2018）
    - TEnet（Li WJ. et al., 2019）

本讲主要内容

- 影响语音识别性能的环境变化因素
- 噪声环境下的鲁棒语音识别技术
- 自适应技术

自适应技术

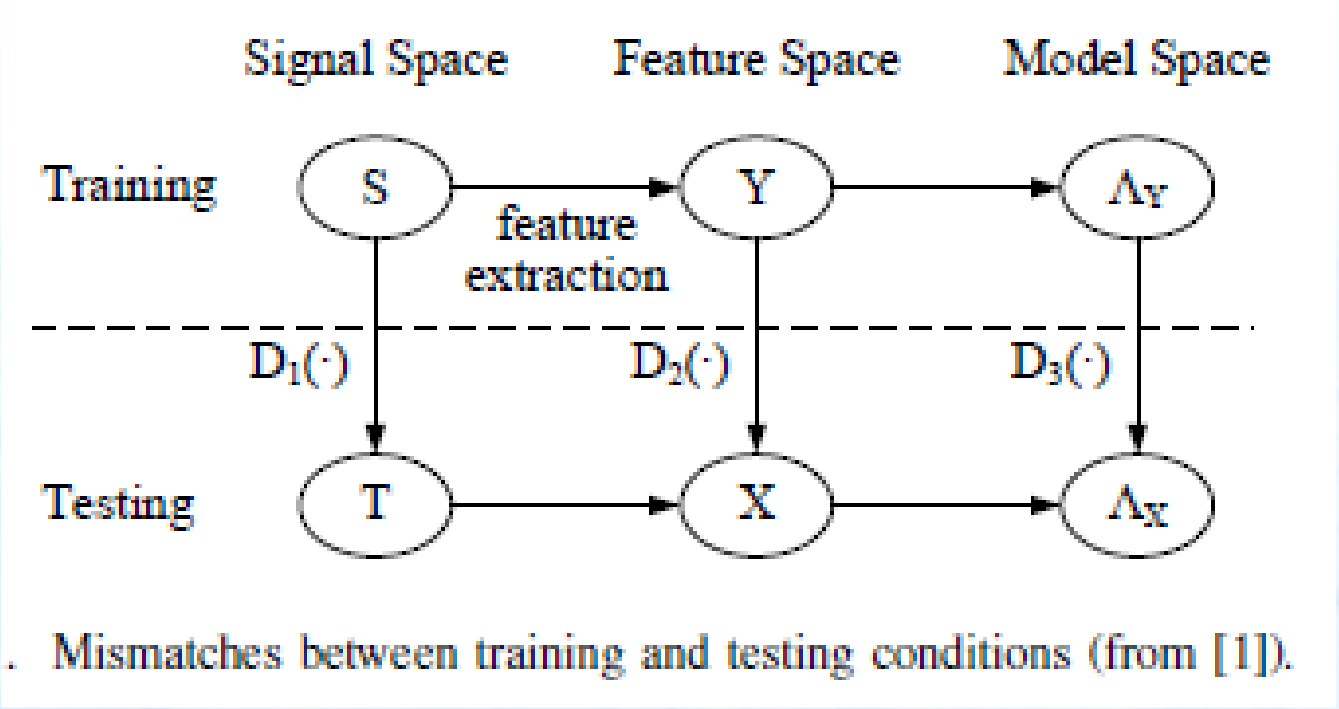
中国科学院声学研究所
Institute of Acoustics, CAS



自适应技术

中国科学院声学研究所
Institute of Acoustics, CAS

❑ 解决思路



. Mismatches between training and testing conditions (from [1]).

自适应技术产生的背景

- 九十年代以来，人们在非特定人大词汇量连续语音识别这一研究领域获得了很大的进展，但与训练得较为充分的特定人系统相比识别性能还是有较大的差距，造成这一差距的主要原因是不同说话人语音之间的差异。

自适应技术的基本含义

说话人无关模型（SI）与说话人相关模型（SD）差异的原因

- 语音学上的差异：由于方言的存在，不同的地方说话人对于同一个句子的发音可能有很大的不同
- 生理上的差异：即使人们采用标准的普通话，不同的说话人的声道形状、声门特性等也存在明显的区别，造成产生的语音频谱特性存在很大的差异
- 发音习惯差异与心理状态差异：每个人有自己发音习惯，说话快慢也很不一样，说话时的心情也往往不同，这些习惯与心态都会对当时说话的语音频谱特征造成影响，从而减低识别系统的性能

自适应技术的基本含义

说话人之间的差异对非特定人语音识别系统造成的影响

- 当某一个使用该系统的说话人语音与训练语音库中的所有说话人的语音都有较大的差别时，对该使用者的语音识别系统的识别性能会有严重的恶化
- 训练一个较好的识别系统需要采集数量很大的说话人的语音用于训练，让训练语音库覆盖更为广泛的语音空间，这样虽然可以减低前面的不良影响，但同时会造成识别系统参数分布较为平缓而不是较为尖锐的分布，产生一个中庸的模型，造成识别系统性能的普遍下降

自适应技术的解释

- 利用系统使用者的少量训练语音，调整系统的参数，使得系统对于该使用者的性能有明显的提高。

- 与说话人相关模型（SD）系统相比，说话人自适应模型（SA）系统引入了SI系统的先验信息，需要用户提供的训练语音数量远低于SD系统，有更好的实用性。

- 因此，非特定人+自适应成为当前各语音识别系统采用的实用框架，自适应算法也成为近年来语音识别领域研究的主要热点之一。

自适应方式的分类

- 按照训练语音获取的不同形式：批处理式、在线式、立即式；
- 按照学习过程有无标注信息：有监督、无监督

- 实用的语音识别系统可以采用批处理+有监督，批处理+无监督，在线式+有监督（对识别结果需要用户验证的系统），在线式+无监督和立即式+无监督的方式。

自适应算法

- 基于最大后验概率的方法（Maximum a posteriori, MAP）

  基本准则是后验概率最大化，利用Bayes学习理论将SI模型的先验信息与目标说话人的信息相结合实现自适应

- 基于变换的方法

  估计SI系统模型与目标说话人之间的空间影射关系，利用SI模型参数和变换矩阵实现模型自适应

- 基于聚类

  用一组HMMs进行自适应的方法

自适应算法

❑ MAP算法

准则：

$$\hat{H}_r = \arg\max_{H_r}(P(H_r \mid O_r))$$

其中： O为训练样本，H_r为第 r 个语音模型的参数，$\hat{H}_r$为模型参数的最大后验概率估计值。

自适应算法

MAP算法

$$\hat{\mu} = \frac{\tau}{\sum_{t=1}^{T}\gamma(t)+\tau} \times \mu + \frac{\sum_{t=1}^{T}\gamma(t)}{\sum_{t=1}^{T}\gamma(t)+\tau} \times \frac{\sum_{t=1}^{T}\gamma(t)o_t}{\sum_{t=1}^{T}\gamma(t)}$$

其中，$\gamma(t)$ 为第 t 帧语音属于该状态的概率

可以改写为：$\hat{\mu} = (1-\beta)\mu + \beta\mu_{SD}$

表明MAP算法估计结果为SD模型参数与SI模型参数的线性组合，但加权系数随训练语音的变化而变化。当没有训练语音时，估计结果即为SI模型参数，可以防止过训练现象；当训练语音增加时，β 取值加大，SD模型参数在结果中的比重加大，使系统性能对目标说话人逐步提高；当训练语音很多时，$\beta \to 1$，系统渐进地逼近于SD系统。

自适应算法

MAP算法优缺点：

- 优点：用基于最大后验概率准则，理论上来说最优，适用于小词汇集或者较多自适应语音数据。
- 缺点：未出现过的语音的模型无法实现自适应，自适应速度缓慢。

  缺陷本质是未考虑到语音模型之间的空间相关性，由此提出很多改进的MAP算法。

自适应算法

基于线性回归预测的MAP算法

算法假设不同语音模型之间的关系可以用线性函数表示，过程为：利用SI系统的训练语音库统计出不同语音的模型参数间的线性关系，在自适应时对于未出现的语音的模型，用已出现的语音的自适应结果及线性关系预测其自适应结果

$$\lambda_i = \sum_{j=1}^{M} \alpha_{ij} \lambda_j$$

自适应算法

❑ 矢量场平滑MAP算法

该算法的基本假设是：不同语音模型自适应前后的变化量是一个连续函数，因此我们可以用已出现语音模型自适应前后的变化量预测相邻的未出现语音的模型的变化量，从而获得未出现语音模型的自适应结果。

$$\Delta\lambda_i = \sum_{j=1}^{M} \alpha_{ij} \Delta\lambda_j \qquad \lambda_i^{'} = \lambda_i + \Delta\lambda_i$$

自适应算法

❑ 基于变换的算法

基本假设：不同说话人的发音空间之间存在某种影射关系，相近语音的SI系统语音空间与目标说话人语音空间的变换关系也是相近的。因此，可以利用训练语音中出现过的语音统计出这一变换关系，对未出现的语音的模型用该变换实现从SI系统到目标说话人语音空间的映射，从而完成自适应的目标。

自适应算法

基于变换算法的优缺点

- 优点：充分利用语音模型空间相关性，适合大词汇量快速自适应。
- 缺点：SI系统与新用户的变换关系复杂，不能用简单的关系表示；当自适应语音较多时，效果不如MAP算法好，不能渐进逼近SD结果。

自适应算法

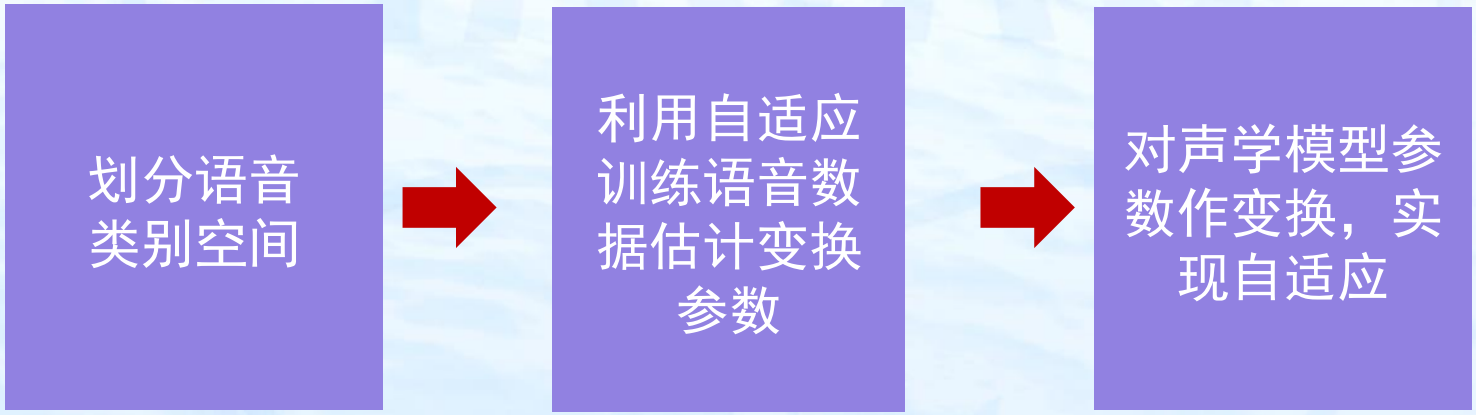
- **几种基于变换的方法：**

  - 最大似然线性回归（Maximum Likelihood Linear Regression, MLLR）算法
  - 最大后验概率线性回归（MAP Linear Regression, MAPLR）算法
  - 非线性变换算法

自适应算法

- 最大似然线性回归（MLLR）算法

算法流程



自适应算法

- 最大后验概率线性回归（MAP Linear Regression, MAPLR）算法

ML估计对于变换参数的取值没有任何限制，仅依赖于自适应数据和原来的声学模型。当自适应数据很少时，很可能使得到的变换参数破坏原来声学空间的潜在结构， 从而使自适应后的模型的性能变差。为了解决这个问题，有研究者采用MAP准则来估计线性变换参数，提出了MAPLR。

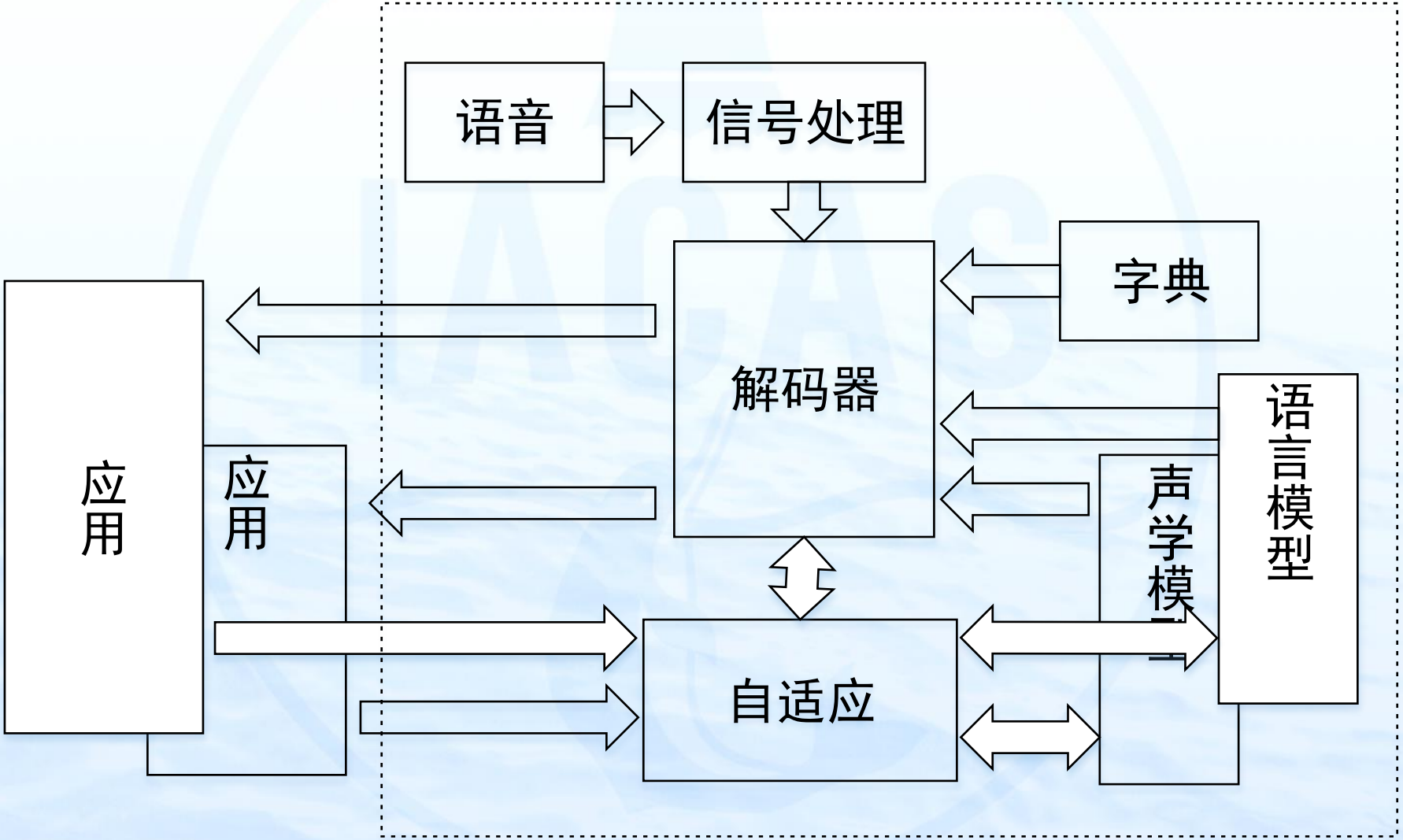
MAPLR 在估计变换参数时加入了先验信息， 用于限制线性变换参数的取值，从而保证自适应后的模型参数不会产生较大的偏差。

自适应算法

- 基于线性变换的自适应方法扩展
  - Iterative MLLR
  - 基于置信度（confidence score）的自适应
  - 基于lattice的自适应
  - 区分性线性变换:

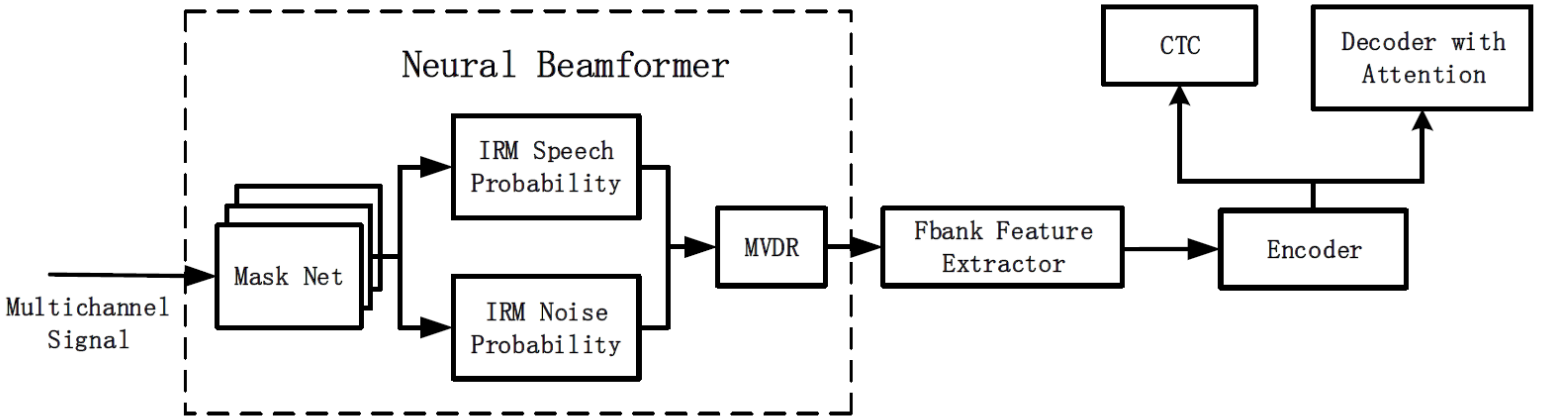
    H-criterion估计线性变换、最大互信息线性回归（MMI Linear Regression, MMILR）、最小音素错误线性回归（MPE Linear Regression, MPELR）、最小分类错误线性回归（MCE Linear Regression, MCELR）、最小词分类错误线性回归（MWCE Linear Regression, MWCELR）、软分类边缘估计线性回归（SME Linear Regression, SMELR）

自适应技术的发展趋势



端到端远场语音识别

- 基于掩蔽估计波束形成的端到端远场语音识别框图



基于时频掩蔽的波束形成　Fbank特征提取　端到端语音识别

梯度从语音识别后端一直传递到前端

端到端远场语音识别

中国科学院声学研究所
Institute of Acoustics, CAS

基于掩蔽估计波束形成的端到端远场语音识别算法

- 掩蔽网络
  - 信号

$$Z_c^S = BLSTM^S(X_c) \qquad m_{t,c}^S = sigmoid(W^S Z_{t,c}^S + b^S)$$

  - 噪声

$$Z_c^N = BLSTM^N(X_c) \qquad m_{t,c}^N = sigmoid(W^N Z_{t,c}^N + b^N)$$

- 功率谱矩阵

$$\Phi_f^S = \frac{1}{\sum_{t=1}^{T} m_{t,f}^S} \sum_{t=1}^{T} m_{t,f}^S X_{t,f} X_{t,f}^H \qquad \Phi_f^N = \frac{1}{\sum_{t=1}^{T} m_{t,f}^N} \sum_{t=1}^{T} m_{t,f}^N X_{t,f} X_{t,f}^H$$

- 波束形成

$$g_f = \frac{(\Phi_f^N)^{-1}\Phi_f^S}{Tr((\Phi_f^N)^{-1}\Phi_f^S)} u$$

端到端远场语音识别

- 基于掩蔽估计波束形成的端到端远场语音识别具体算法
  - 参考向量估计
    - 特征：

$$q_c = \frac{1}{T}\sum_{t=1}^{T}\{z_{t,c}^{S}, z_{t,c}^{N}\} \qquad r_c = \frac{1}{C-1}\sum_{k=1,k\neq c}^{C}\{R(\phi_{f,c,k}^{S}), \mathrm{I}(\phi_{f,c,k}^{S})\}_{f=1}^{F}$$

    - 注意力机制：

$$v_c = w^T \tanh(W_q q_c + W_r r_c + b)$$

$$u_c = \frac{\exp(\beta v_c)}{\sum_{c=1}^{C}\exp(\beta v_c)}$$

  - 声学建模

$$L = L_{CTC} + \lambda L_{Attention}$$

谢 谢！