深蓝学院
shenlanxueyuan.com

语音信号处理概述

主讲人 宋 辉

清华大学电子工程系 博士
滴滴AI Labs 语音技术部

1. 语音交互

2. 复杂的声学环境

3. 前端语音信号处理

4. 课程安排

5. 推荐阅读

1. 语音交互

语音交互（VUI）是指人与人/设备通过自然语音进行信息传递的过程。

语音交互的优势：

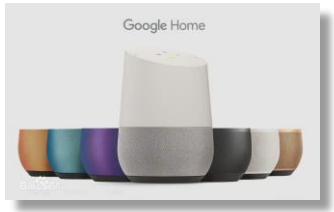
- **输入效率高。**相比键盘输入，语音输入的速度是传统输入方式的3倍以上。
- **解放双手和双眼，更安全。**例如车载场景通过语音点播音乐和导航。
- **使用门槛低。**人类本就是先有语音再有文字，对于那些无法用文字交互的人来说，语音交互会为其带来极大的便利。
- **传递更多的声学信息。**声纹、性别、年龄、情感等。

语音交互的劣势：

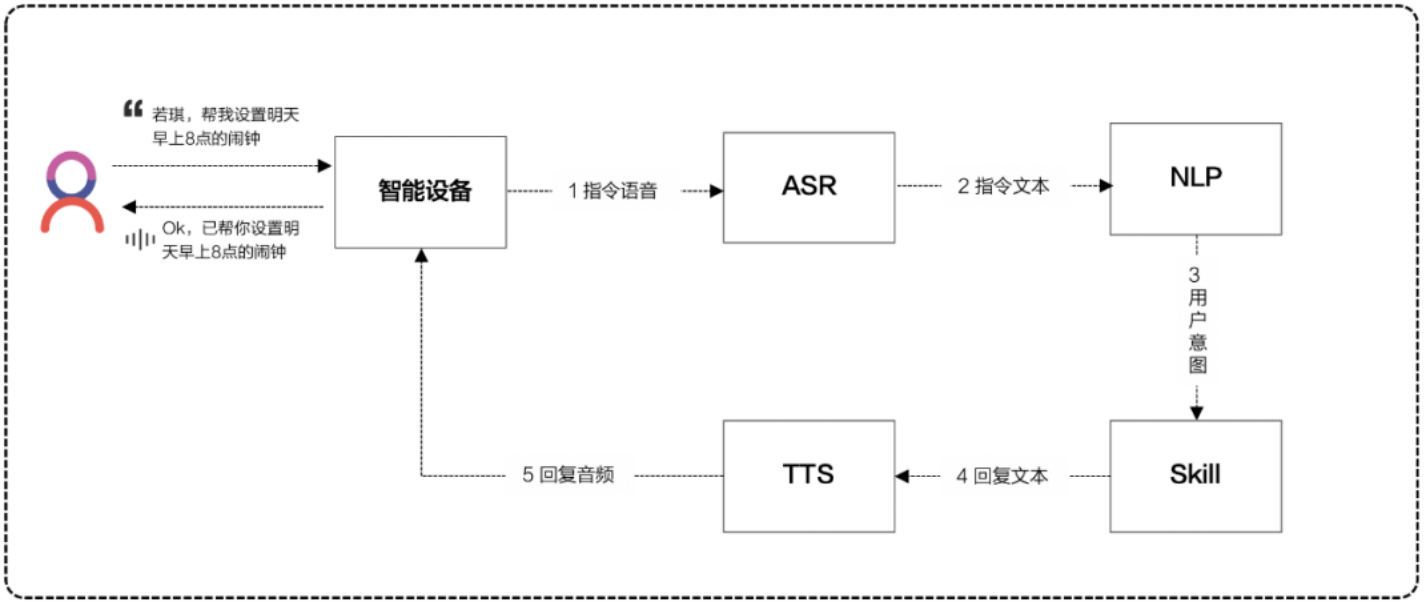
- 信息接收效率低、**复杂的声学环境、**心理负担。

人机语音交互

- 1952年，贝尔实验室，阿拉伯数字识别系统Audrey
- 1962年，IBM-Shoebox
- …… ……
- **2011年，iphone4s，Siri问世**
- 2014年，win8，Cortana
- **2014年，Amazon发布echo音箱**
- 2016年，Google发布Google Home



人机语音交互流程


若琪，帮我设置明天早上8点的闹钟
Ok，已帮你设置明天早上8点的闹钟
智能设备
1 指令语音
ASR
2 指令文本
NLP
3 用户意图
Skill
4 回复文本
TTS
5 回复音频

语音交互适合什么场景

	家庭场景							车载场景	外出场景	办公场景	医疗场景	教育场景	出行场景
	音箱	电视	空调	灯	平板电脑	冰箱	洗衣机	汽车	耳机	电脑	病历录入	学习平板	耳机
1 加分项													
1.1 需要复杂的信息输入	2	2	2	0	2	1	1	2	2	2	2	2	2
1.2 使用对象双手或双眼被占用	0	0	0	1	0	0	0	2	2	0	2	0	2
1.3 使用对象为非文字使用者（需要文字输入）	1	2	0	0	1	0	0	1	1	0	0	2	1
1.4 需要跨短距离空间的操作	1	1	0	2	0	2	0	0	0	0	0	0	0
1.5 原信息输入的工具比较受限	0	2	1	2	0	1	2	0	0	1	1	0	1
1.6 跨意图指令输入	2	2	0	0	2	0	0	2	2	1	0	1	1
1.7 使用频次	2	2	2	2	2	1	0	2	2	2	2	2	2
1.8 原设备和声音的关联度高	2	2	0	0	2	0	0	2	2	1	0	2	2
1.9 需要声音传递额外信息	1	1	0	0	1	0	0	1	0	0	0	2	1
2 减分项													
2.1 环境私密程度低	0	0	0	0	0	0	0	0	1	2	0	1	1
2.2 环境嘈杂	0	0	0	0	0	0	0	1	1	1	0	1	1
2.3 涉及到多层次交互（屏幕可弥补）	1	0	0	0	0	0	0	0	0	0	0	0	0
2.4 涉及到多条目选择（屏幕可弥补）	1	0	0	0	0	1	0	0	0	0	0	0	0
2.5 涉及到重要/隐私信息传达（屏幕可弥补）	0	0	0	0	0	0	0	0	0	0	0	0	0
综合得分	9	14	5	7	10	4	3	11	9	4	7	9	10

总结起来就是：家里、车里、路上。

2. 复杂的声学环境

现实中的语音交互系统，无一例外的会受到各种环境不利因素的影响，极大影响了交互成功率和用户体验。

- 方向性干扰
- 环境噪声（散射噪声）
- 远讲产生的混响
- 声学回声

痛点：人和机器都听不清

2. 复杂的声学环境

一个成功的语音交互产品，意味着对语音交互的场合和使用模式无约束。

前端语音信号处理的意义：

- 面对噪声、干扰、声学回声、混响等不利因素的影响，运用信号处理、机器学习等手段，提高目标语音的信噪比或主观听觉感受，增强语音交互后续环节的稳健性。
- 让人听清：更高的信噪比，更好的主观听觉感受和可懂度，更低的处理延时。
- 让机器听清：更好的声学模型适配，更高的语音识别性能。

总结：语音信号处理的目标，是为了让人和机器更容易听清语音，让语音交互更加自然和无约束。

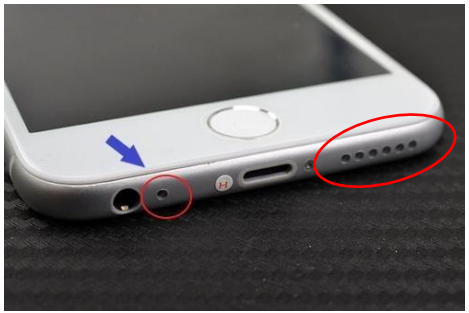
3. 前端语音信号处理用于语音交互

免提通话

电话/视频会议

3. 前端语音信号处理用于语音交互

你知道苹果手机有几个麦克风吗?

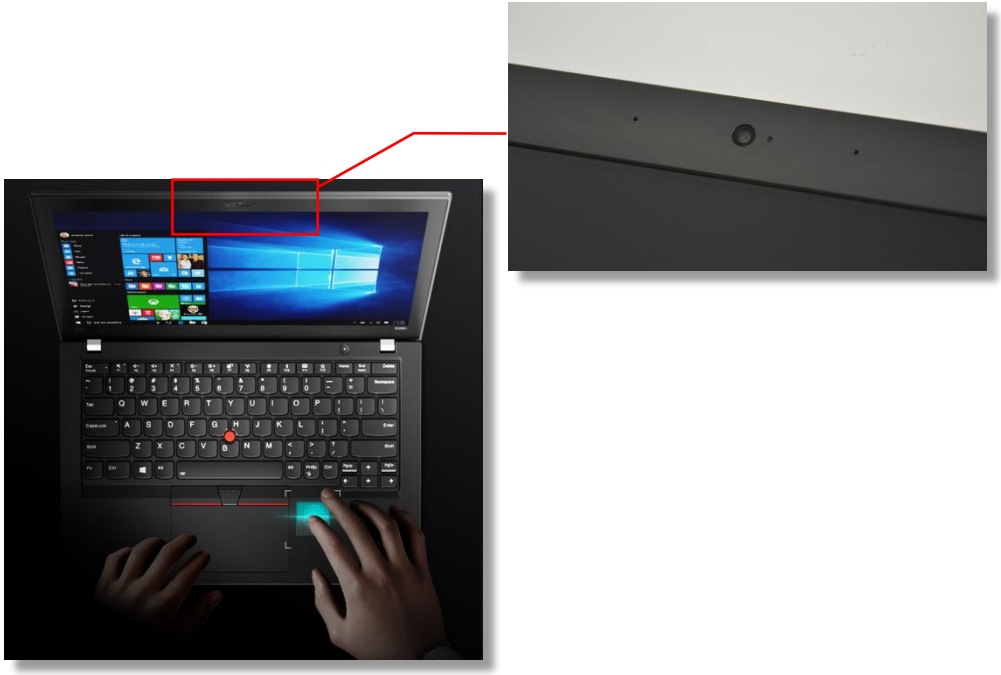


通话收音

录制视频、主动降噪

主动降噪

3. 前端语音信号处理用于语音交互



thinkpad

3. 前端语音信号处理用于语音交互

Siri

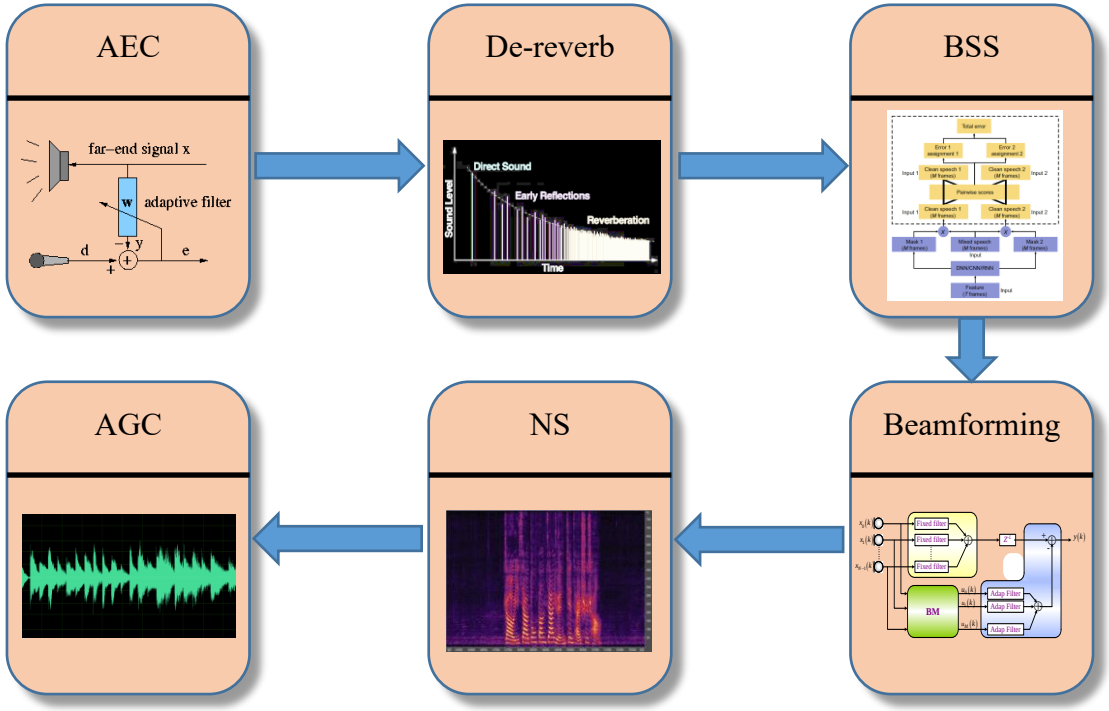
Amazon

车载

家居

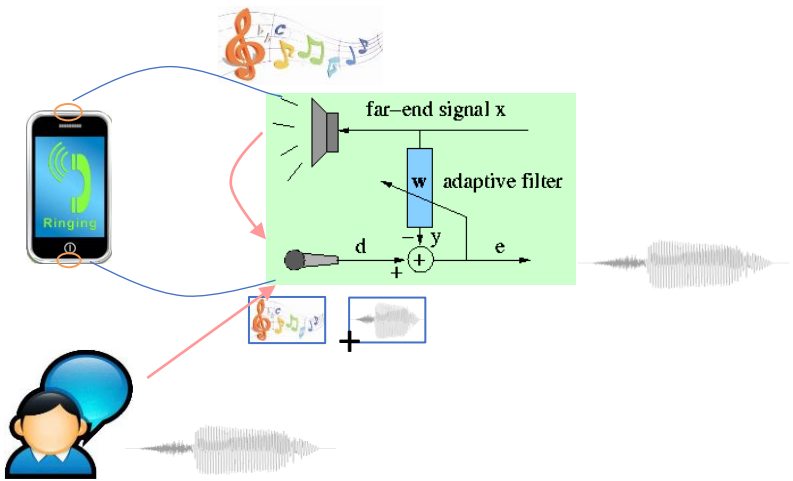
3. 前端语音信号处理的场景细分

分而治之——针对不同的干扰因素，采用不同的信号处理算法



3. 前端语音信号处理的场景细分

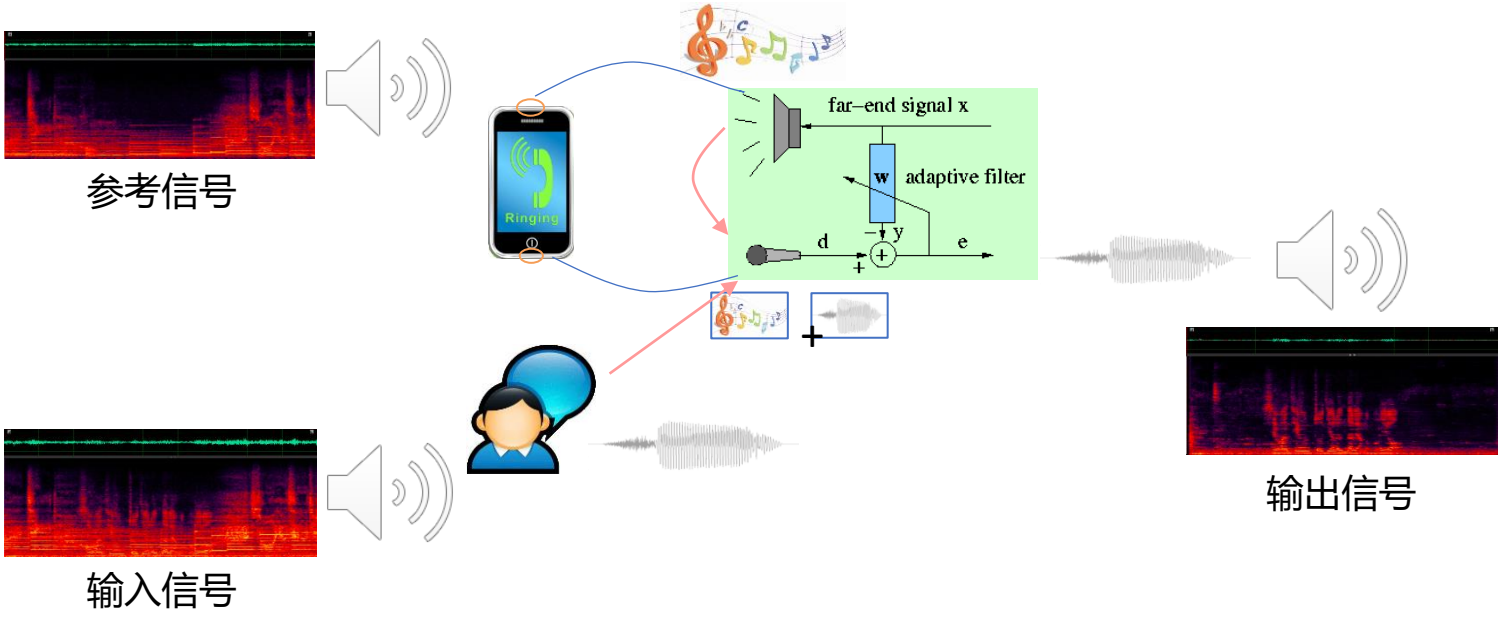
声学回声消除——消除设备自身的干扰



- 消除设备自身产生的回声干扰
- 最早应用于全双工语音通信、视频会议
- 在语音交互中起到打断唤醒的作用
- 主要模块
  - 时延估计
  - 线性回声消除
  - 双讲检测
  - 残余回声抑制

3. 前端语音信号处理的场景细分

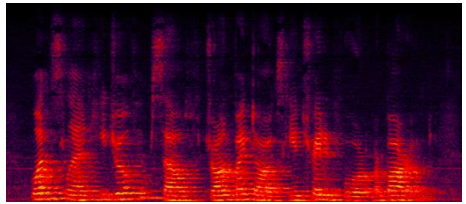
声学回声消除——音频示例



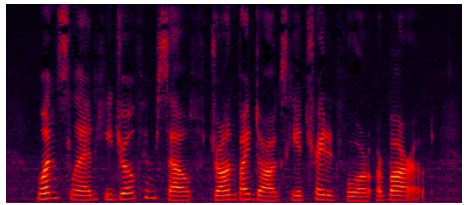
3. 前端语音信号处理的场景细分

解混响

- 盲反卷积法 [Neely and Allen，1979]
  - 估计RIR的逆滤波器
- 加权预测误差 [Takuya，2012]
  - 消除晚期混响，适用于单通道和多通道场景
- 麦克风阵列波束形成
- 深度学习用于解混响 [Han，2015]
  - 通过DAE、DNN、LSTM或者GAN，实现频谱映射



有混响



去混响

3. 前端语音信号处理的场景细分

语音分离——旨在解决“鸡尾酒会问题”

- 听觉场景分析法 [Hu and Wang，2004]

  本质上是对人的听觉特性的模拟，具体手段是二分类+监督学习

- 非负矩阵分解 [Lee and Seung，2001]

  基于统计独立假设，语音信号的稀疏性与谐波特性

- 多通道技术

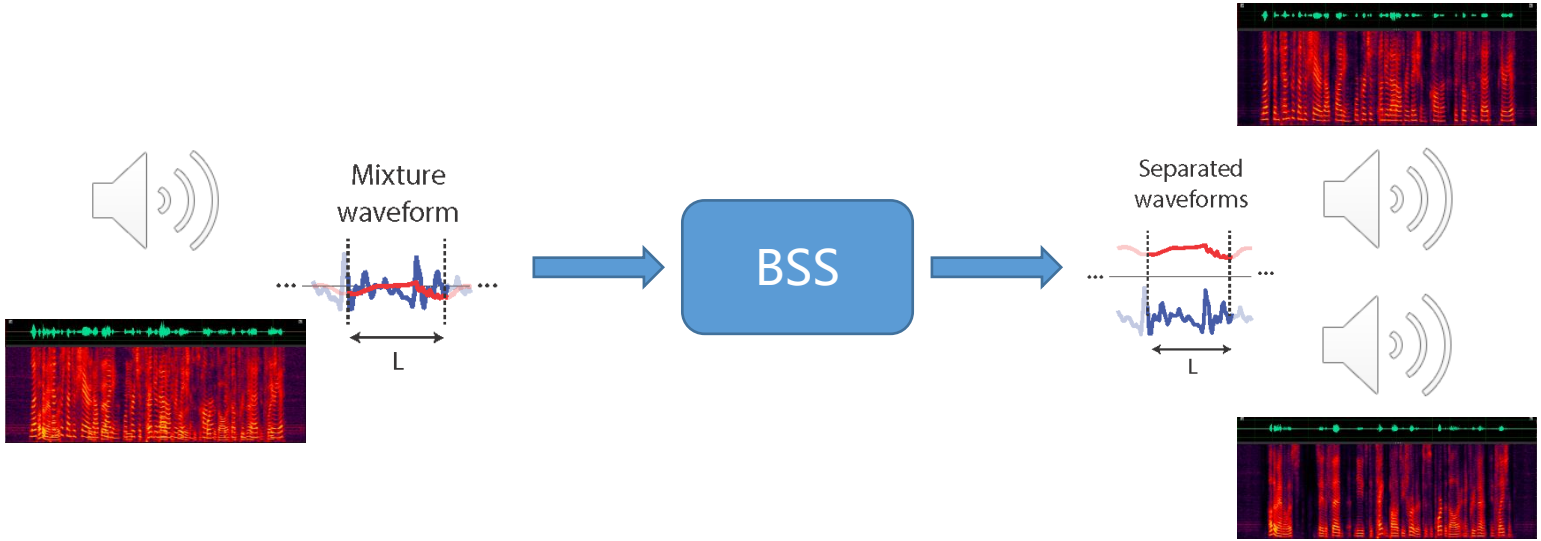
  fix beamforming，adaptive beamforming，ICA

- 基于深度学习的语音分离

  Deep clustering [Hershey, 2016]
  Deep attractor network [Luo and Chen, 2017]
  Permutation invariant training [Yu, 2017]

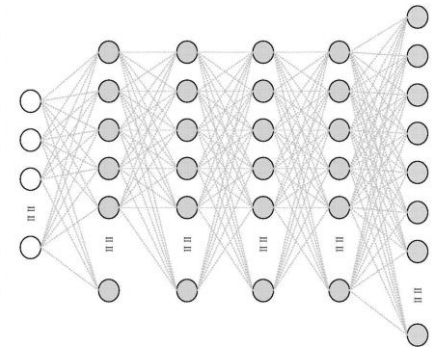
3. 前端语音信号处理的场景细分

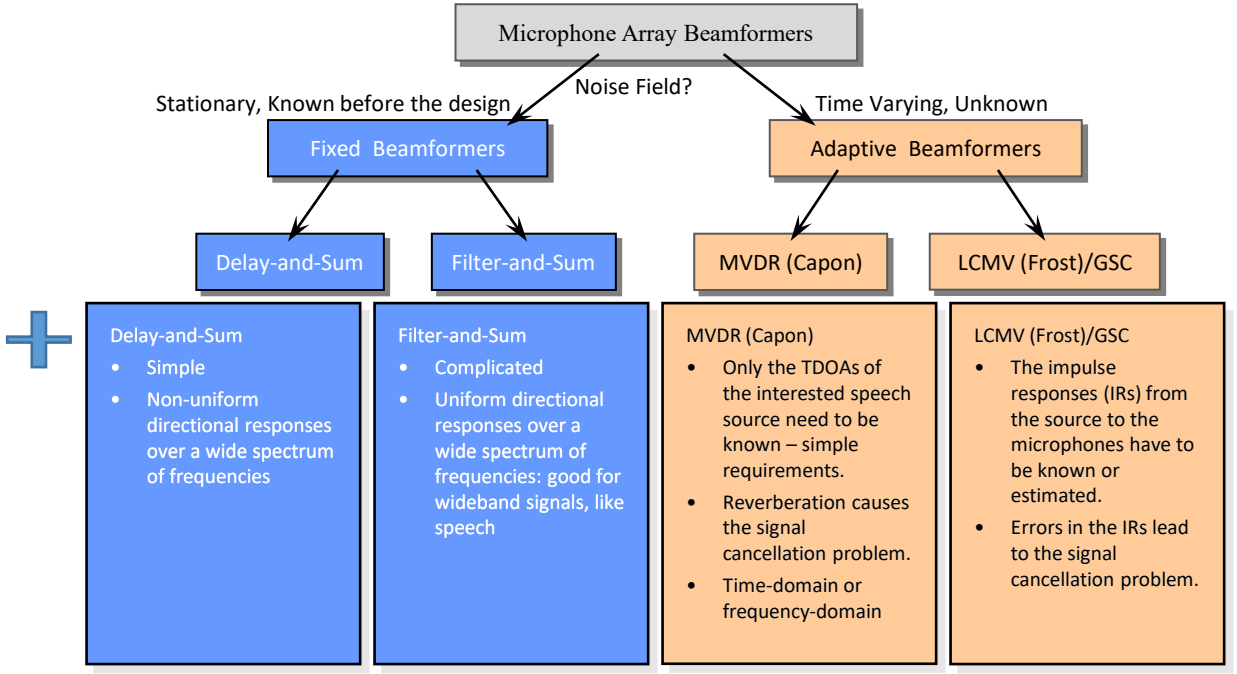
语音分离——音频示例



3. 前端语音信号处理的场景细分

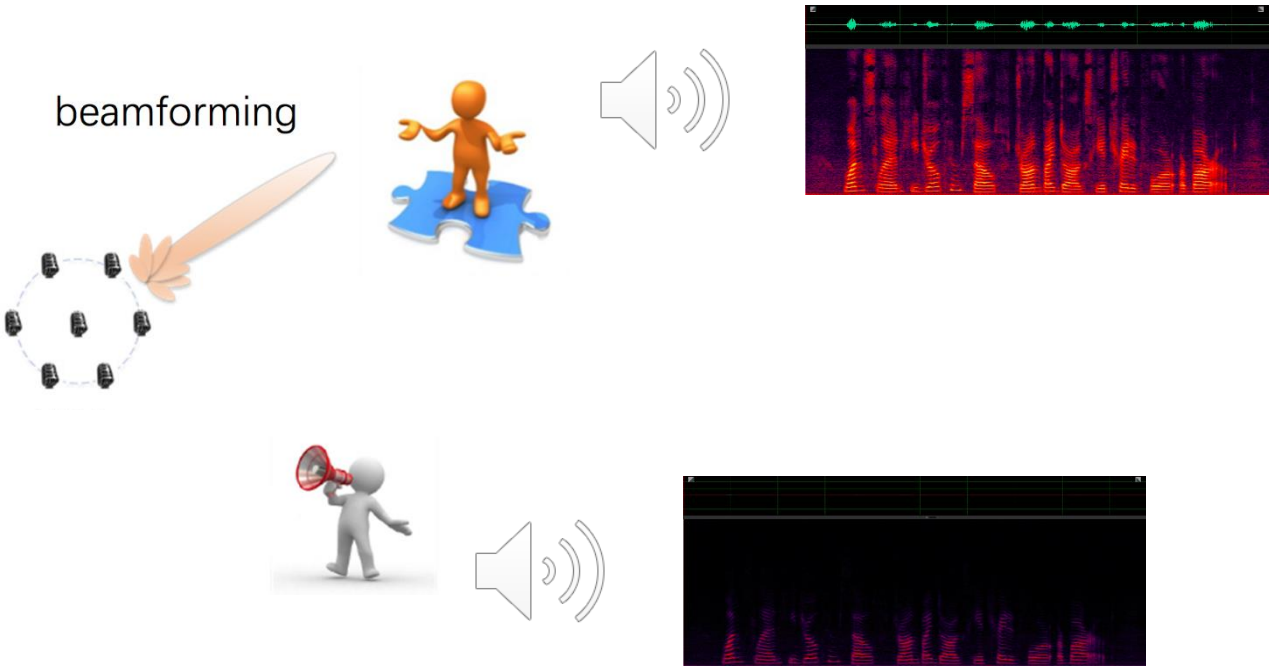
波束形成——用于多通道语音增强、信号分离、去混响、声源定位





3. 前端语音信号处理的场景细分

波束形成——音频示例



3. 前端语音信号处理的场景细分

噪声抑制——消除或抑制环境噪声，增强语音信号

- 基于统计模型的方法

  最小均方误差MMSE、最大似然估计ML、最大后验估计MAP

- 基于子空间的方法

  利用语音和噪声的不相关性，借助特征值/奇异值分解手段分解到子空间处理

- 语音增强的核心在于噪声估计

  递归平均、最小值追踪、直方图统计是比较常用的噪声估计手段

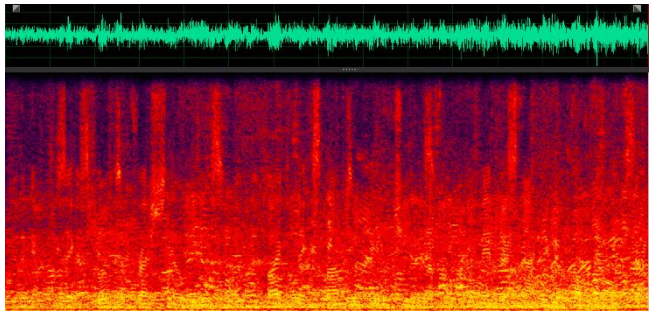
- 基于深度学习的语音增强方法

  两大类方法：masking && mapping

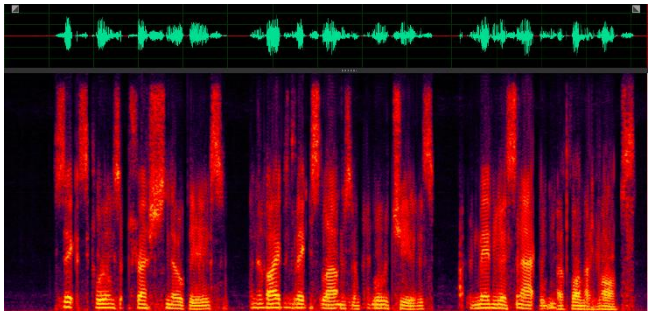
  通过DNN、CNN、RNN或者GAN，在频域或时域实现（多为频域）

3. 前端语音信号处理的场景细分

噪声抑制——音频示例



增强前



增强后

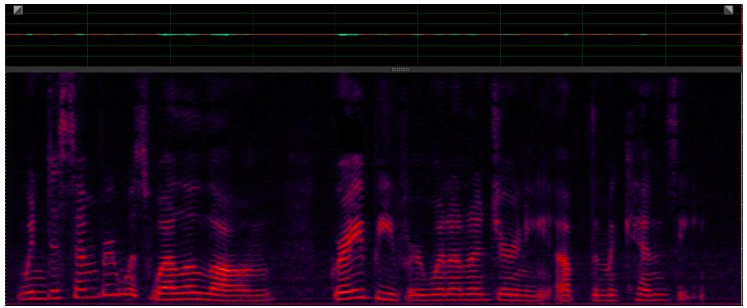
3. 前端语音信号处理的场景细分

幅度控制

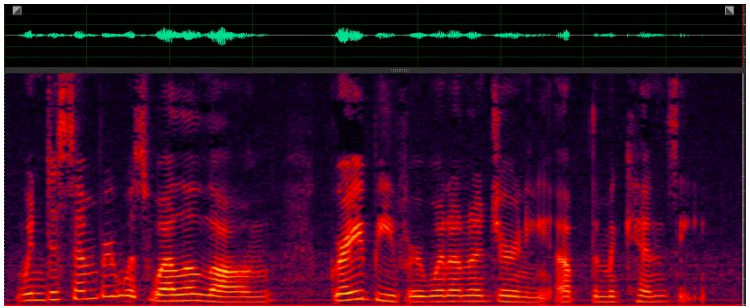
- 自动调整信号的动态范围
- 常用的两种方法
  - 动态范围控制（Dynamic Range Control）
  - 自动增益控制（Automatic Gain Control）

3. 前端语音信号处理的场景细分

幅度控制——音频示例



AGC前



AGC后

3. 前端信号处理的技术路线

传统的前端信号处理方案

- 处理依据——“规则”
  - 客观物理模型，即声音传播的物理规律
  - 语音信号的时域、频域和空域特性
- 针对不同的干扰因素，采用不同的信号处理算法加以解决
- 优化目标：
  - 抑制干扰信号，提取目标信号
- 优化准则：
  - MSE(Mean Square Error)准则

3. 前端信号处理的技术路线

信号处理与深度学习相结合的方案

- 处理依据——“规则+学习”客观物理模型
  - 语音信号的时域、频域、空域特性
  - 海量音频数据先验信息
- 既保留了声音传播的物理规律和信号本身的时域、频域、空域特性，又引入了先验数据统计建模的方法。
- 优化准则：
  - MSE准则

3. 前端信号处理的技术路线

基于深度学习的前后端联合优化方案

- 处理依据——“端到端联合建模”

  输入多通道麦克风信号，输出语音识别结果

  利用近场数据，仿真得到海量的带有各种干扰的训练数据

- 将前端信号处理与后端ASR声学模型联合建模，用一套深度学习模型完成语音增强和语音识别任务。

- 优化准则：

  识别准确率

4. 课程安排

第1章 语音信号处理概述

- 语音交互
- 复杂的声学环境
- 前端语音信号处理
- 课程安排
- 推荐阅读

第2章 数字信号处理中的几个关键概念

- 数字信号及其基本运算
- 采样定理
- 时频分析与傅里叶变换
- 实战

第3章 自适应滤波方法（一）

- LMS算法
  - 基本LMS算法
  - LMS的衍生算法
- 实战

第4章 自适应滤波方法（二）

- RLS算法
  - 基本RLS算法
  - RLS的衍生算法
- AP算法
- 实战

第5章 声学回声消除和噪声抑制技术

- 子带分解——FFTbank
- 声学回声消除AEC
- 噪声抑制NS
- 实战

4. 课程安排

第6章 阵列信号处理（一）

- 阵列信号处理的基本概念
- 几种经典的波束形成算法
  - Delay-and-sum
  - MVDR、LCMV、GSC
- 实战

第7章 阵列信号处理（二）

- 声源定位技术
  - 基于最大输出功率的可控波束形成技术
  - 高分辨率谱估计技术
  - 基于到达时间差的方法
- 波束形成——GSC算法的实际应用
- 实战

第8章 深度学习用于语音分离

- 语音分离的常用模型
- 训练目标与特征
- 单通道语音分离算法
- 实战

第9章 深度学习在语音信号处理中的其他应用

- 多通道语音分离技术
  - NN-BF算法
- End2end方法
- 实战

4. 课程特点

- 理论与实践兼顾
- 立足传统技术，辐射前沿进展
- 公式较多，重在理解
- 实战巩固

5. 推荐阅读

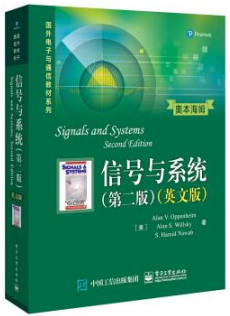
-- 奥本海姆，《信号与系统》，电子工业出版社

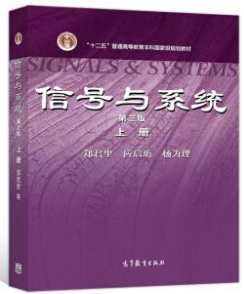
-- 奥本海姆，《离散时间信号处理》（Discrete Time Signal Processing, Third Edition）

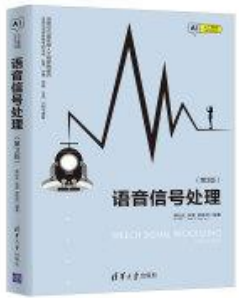
-- 郑君里，《信号与系统》，电子工业出版社，高等教育本科国家级规范教材

-- 赵力，《语音信号处理》，机械工业出版社

-- 韩纪庆，《语音信号处理》，机械工业出版社







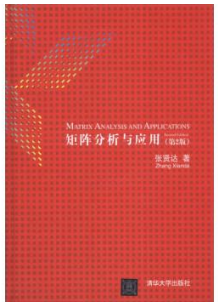
5. 推荐阅读

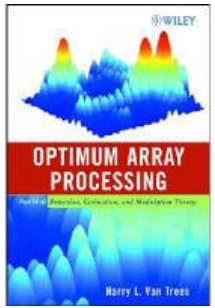
-- 张贤达，《现代信号处理》，清华大学出版社

-- 张贤达，《矩阵分析与应用》，清华大学出版社

-- Van Trees，检测、估计和调制理论（IV)《Optimum array processing》

-- Signals and Systems: an Introduction to Analog and Digital Signal Processing. 1987 Lecture. Alan V. Oppenheim





5. 推荐开源项目

-- Athena-signal

https://github.com/athena-team/athena-signal

5. 推荐开源项目

-- Python for Signal Processing

https://github.com/unpingco/Python-for-Signal-Processing

《Python for Signal Processing: Featuring IPython Notebooks》对应源码，包含信号处理12大类（采样定理、傅里叶变换、滤波器等）、随机过程15大类（高斯马尔科夫、最大似然等）

-- Speex https://www.speex.org

A Free Codec For Free Speech。专门语音压缩而设计的，包含超过9种算法：AEC、NS、VAD等，不过现在被Opus替代。

-- Google WebRTC https://webrtc.org

一个免费的开放式项目，通过简单的API为浏览器和移动应用程序提供实时通信（RTC）功能。

-- VOICEBOX: Speech Processing Toolbox for MATLAB

http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

语音处理工具箱，由MATLAB程序组成。超过100个函数，包含语音增强、ASR等在内。

感谢聆听

Thanks for Listening!