复杂声学环境下基于深度学习的语音分离

李军锋

中国科学院声学研究所

提纲

一、简介

二、训练目标

三、特征

四、增强算法

五、总结

提纲

一、简介

二、训练目标

三、特征

四、增强算法

五、总结

现实环境的听觉感知

内容：

- 语音
- 说话人
- 音乐
- 场景

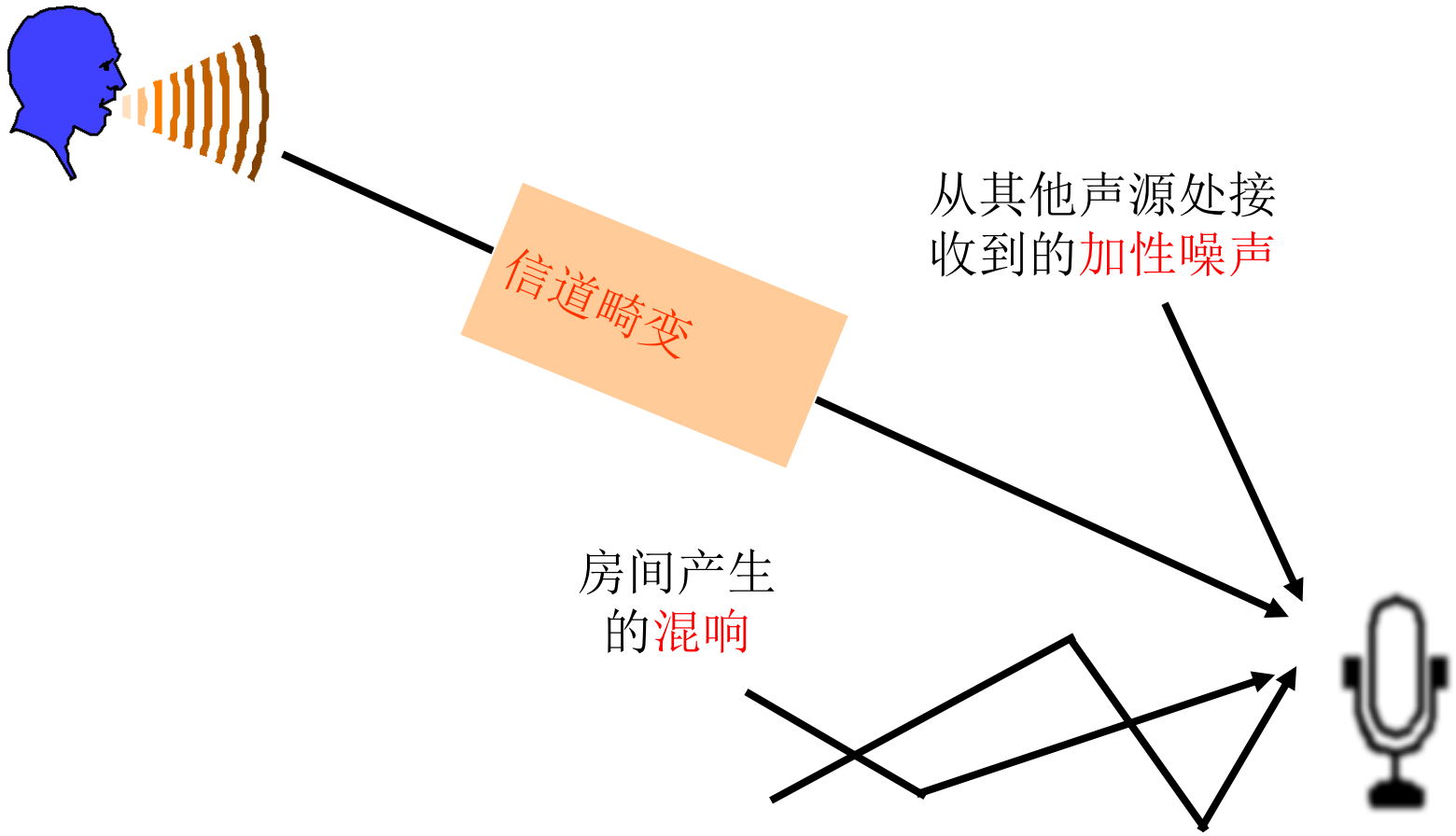
方位：

- 上下左右
- 距离

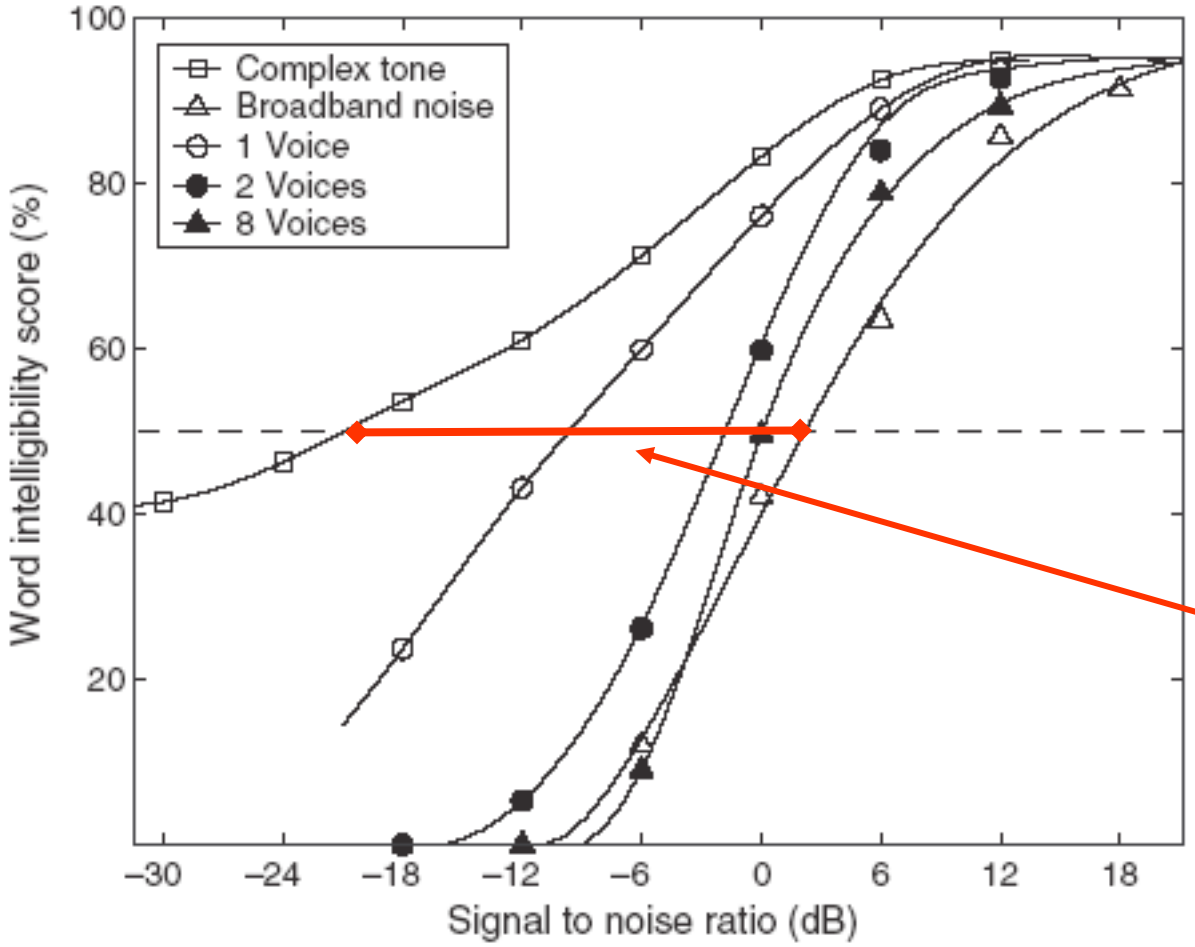
传输环境特性：

- 混响
- 加性噪声

噪声源



不同干扰强度下人类的听觉感知能力



人的听觉感知能力在不同的典型干扰源下有23dB差距!

语音分离的典型应用

- **稳健语音识别与声纹识别**
- **助听设备：**
  - 助听器
  - 人工耳蜗
- **移动通信降噪**
- **音频信息检索**

语音分离的传统方法

- **语音增强**
  - 语音与噪声的统计特性
  - 难点：噪声估计
- **基于麦克风阵列的空间滤波**
  - 波束形成（Beamforming）
    - 通过传感器阵列获取声源的方位信息，利用噪声的不相关性滤除噪声
  - 独立成分分析（ICA）
    - 寻找声源的解混矩阵
- **计算听觉场景分析 (CASA)**
  - 特征分析 (例如 基频)，建模 (例如 说话人模型，聚类模型)

语音分离的有监督学习方法

- **数据驱动方法：需要有训练集**
- **起源于CASA**
  - 时频掩码（Time-Frequency Masking）的概念使得语音分离变为一个有监督学习的问题
- **深度学习的成功应用使得有监督学习方法成为研究趋势**

深度神经网络

- **为什么需要深层网络?**
  - 随着层数的加深，微小的噪声和非线性变化分量将逐层被丢弃，特征越来越反映数据的manifold信息
  - 实际使用时，更深的网络可能带来性能的提升 (前提：如果能够被成功训练)
- **深层网络比较难训练**
  - 梯度消失问题：误差梯度从顶层往底层传播时逐渐变小
    - 解决方法：1）受限玻尔兹曼机分层预训练；2）ResNet etc.
  - 局部最优化问题：网络非凸
    - 解决方法：大数据
  - 过拟合问题：网络复杂度不易控制，容易overfit小数据
    - 解决方法：大数据

常见的DNN网络结构

- **前馈神经网络**
  - 至少有两个隐层的全连接**多层感知机**(MLP)
  - 卷积神经网络(CNN)
    - 基于subsampling 的池化（pooling）
    - 参数共享
  - 反向传播算法
- **递归神经网络**
  - 面向时间序列的反向传播算法
  - 为了避免梯度消失或梯度爆炸，LSTM使用带门限的记忆单元以便于信息（或梯度）沿着时间轴方向传播

提纲

一、简介

二、训练目标

三、特征

四、增强算法

五、总结

训练目标1：理想二值掩码（ Ideal binary mask ）

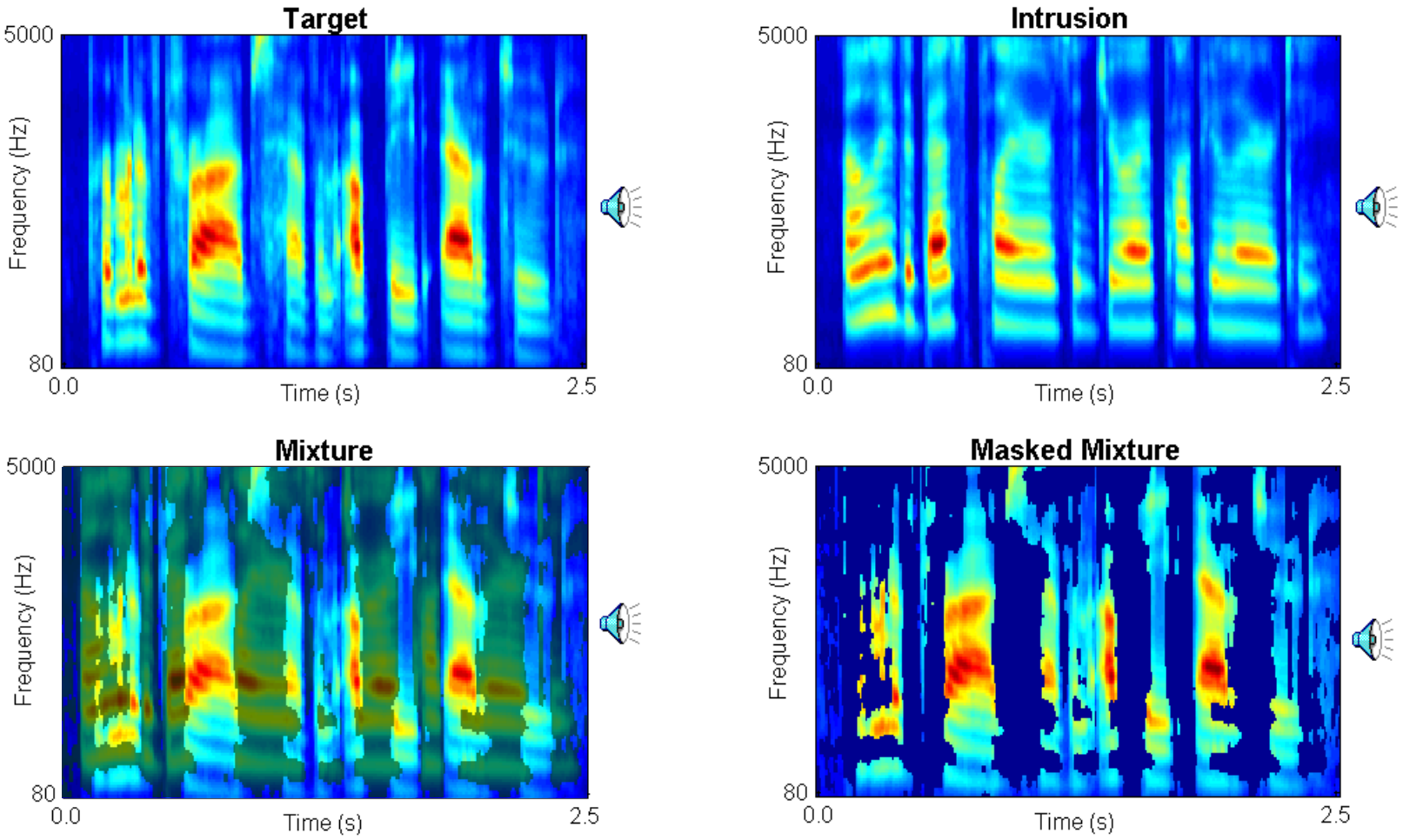
- **理想二值掩码起源于CASA (Hu & Wang'04)，受到了人类听觉掩蔽效应和听觉场景分析的启发。**
- **核心思想：在某个含噪时频谱中，保留目标声源大于背景噪声的T-F单元，将其他部分置零。**
- **理想二值掩码的定义 (IBM)**

$$IBM(t,f)=\begin{cases}1 & \text{if } SNR(t,f)\geq\theta\\ 0 & \text{otherwise}\end{cases}$$

θ: 局部信噪比门限值, 通常设置为0 DB

最优信噪比: 在某些条件下， $\theta = 0$ dB时的IBM是最优二值掩码（依据信噪比增益） (Li & Wang'09)

IBM示例


Target
5000
Frequency (Hz)
80
0.0
Time (s)
2.5
Intrusion
5000
Frequency (Hz)
80
0.0
Time (s)
2.5
Mixture
5000
Frequency (Hz)
80
0.0
Time (s)
2.5
Masked Mixture
5000
Frequency (Hz)
80
0.0
Time (s)
2.5

使用IBM作为分离目标的主观听觉测试

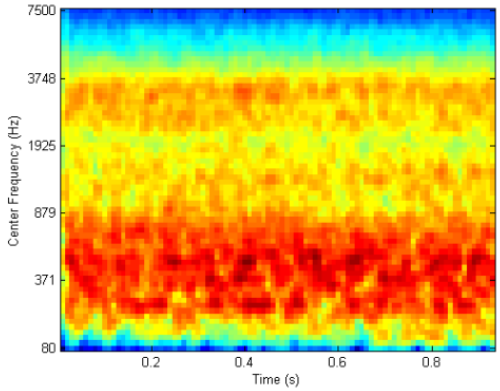
- **将IBM作为分离目标能够大幅提升人的语音可懂度**
  - 在平稳噪声的环境下，IBM能帮助听力正常的人的听觉感知能力提升7 dB 以上 (Brungart et al.’06; Li & Loizou’08; Cao et al.’11; Ahmadi et al.’13)
  - 在平稳噪声的环境下，IBM帮助有听力障碍的人的听觉感知能力提升9dB以上 (Anzalone et al.’06; Wang et al.’09)
  - 在modulated噪声环境下，IBM对人的听觉感知能力的改善要显著大于在平稳噪声下对人的听觉感知能力的改善

将IBM作为优化目标，使得语音分离问题变成了两类分类问题。

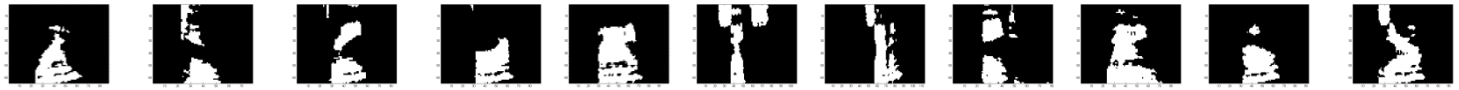
该思想打开了各种模式识别、机器学习方法在语音分离领域应用的窗口

IBM在强噪声环境下的效果

当信噪比趋近于−∞的时候，IBM仍然能给出良好的语音可懂度



Speech shaped noise



关于训练目标的研究

- **自从IBM被首次用作有监督训练的优化目标以来，提出了多种训练目标**
- **基于有监督训练的语音分离/增强的训练目标至关重要**
  - 不同的训练目标的泛化能力是不同的

训练目标2: 理想比值掩码(Ideal Ratio Mask)

- **理想比值掩码IRM (Srinivasan et al.'06; Narayanan & Wang'13; Wang et al.'14; Hummersone et al.'14)**

$$IRM(t,f) = \left(\frac{|S(t,f)|^2}{|S(t,f)|^2 + |N(t,f)|^2} \right)^{\beta} = \left(\frac{SNR(t,f)}{SNR(t,f)+1} \right)^{\beta}$$

  - S 和 N 分别表示语音和噪声
  - β 表示可调参数, 通常设为 0.5
  - 当 $\beta = 0.5$ 时, IRM 是均方根Wiener滤波器，也就是最优估计器

训练目标3：Complex Ideal Ratio Mask (cIRM)

- **该mask可以精确恢复clean speech(Williamson et al.'16)**

$$S(t,f) = cIRM * Y(t,f)$$

- 定义：

$$cIRM(t,f) = \frac{Y_r S_r + Y_i S_i}{{Y_r}^2 + {Y_i}^2} + i\frac{Y_r S_i - Y_i S_r}{{Y_r}^2 + {Y_i}^2}$$

  - 下标 r 和 i 分别表示复频域信号的实部和虚部成分
  - 因为该mask的动态范围不在[0,1]之间，所以需要限幅

其他基于掩码的训练目标

- **频域幅度掩码(spectral magnitude mask, SMM) (Wang et al.'14)**

$$SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|}$$

  - Y表示含噪语音。
- **相位敏感掩码(phase-sensitive mask) (Erdogan et al.'15)**

$$PSM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|}\cos\theta$$

  - θ 表示在该T-F unit上的含噪语音与clean语音的相位差。
  - 由于该优化目标能够将phase增强信息引入优化目标，因此其性能优于频域幅度掩码。

训练目标5：频谱映射

- **频域幅度谱 (target magnitude spectrum, TMM) (Lu et al.'13; Xu et al.'14; Han et al.'14)**

$$|S(t,f)|$$

  - 一种常见的形式是对数幅度谱
- **Gammatone频域功率谱 (Gammatone frequency target power spectrum, GF-TPS) (Wang et al.'14)**

$$S_{GF}^{2}(t,f)$$

- **这两个目标被统称做频谱映射(spectral mapping)**

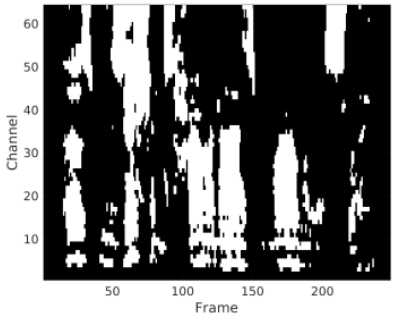
训练目标6：信号估计

- **信号估计(signal approximation)：网络输出的是IRM，损失度量的是增强语音与纯净语音之间的最小均方误差(Weninger et al.'14)**

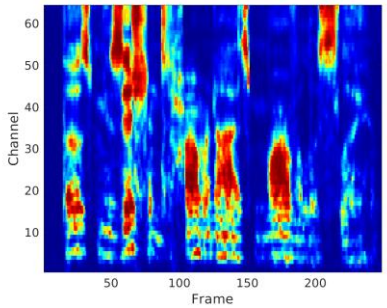
$$SA(t,f) = (RM(t,f)|Y(t,f)| - |S(t,f)|)^2$$

  - *RM*(*t*, *f*)表示网络输出的估计掩码
  - 目标函数能够最大化信噪比

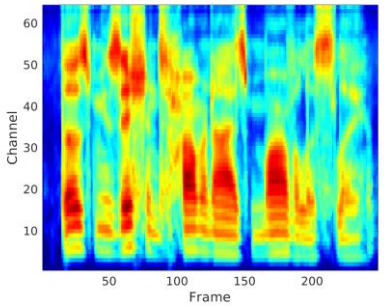
不同训练目标的比较



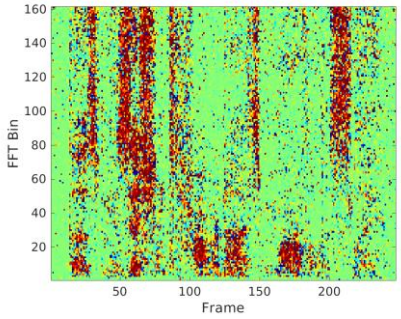
(a) 理想二值掩码



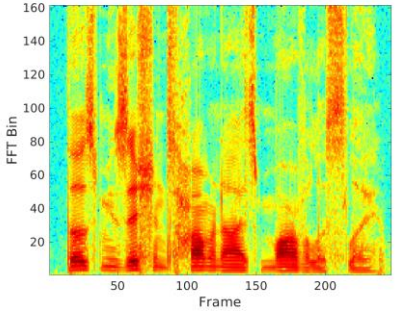
(b) 理想比值掩码



(c) Gammatone域幅度谱



(f) 相位敏感掩码



(g) 频域幅度谱

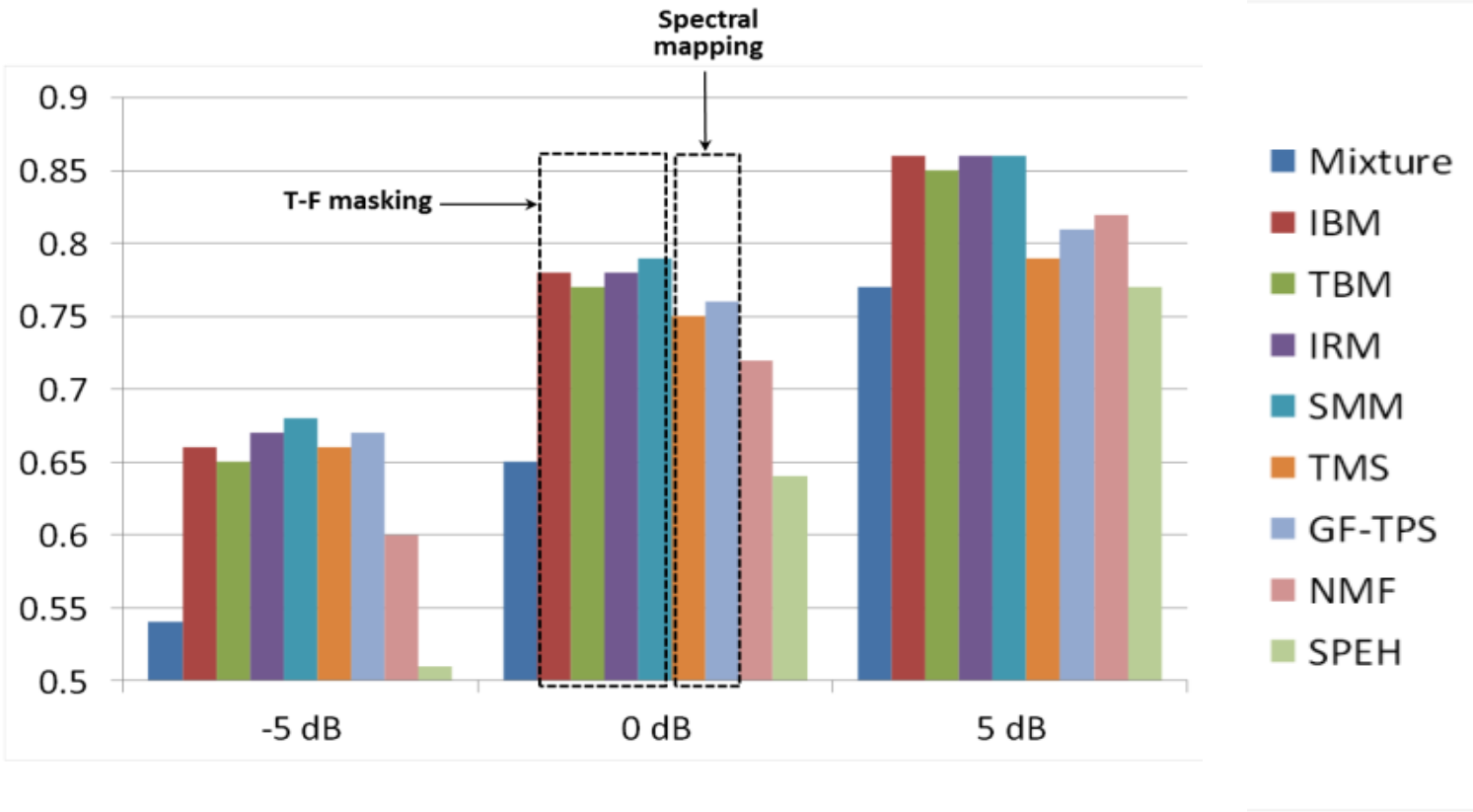
工厂噪声 -5 dB

不同训练目标的实验比较

- **Wang et al. (2014) 比较了DNN在不同训练目标下的性能**
  - T-F masking:
    - IBM, TBM, IRM, SMM
  - Spectral mapping:
    - TMS, GF-TPS
  - 传统增强算法:
    - SPEH (Hendriks et al.'10)
  - 有监督非负矩阵分解 (supervised NMF) :
    - ASNA-NMF (Virtanen et al.'13)
- **评价指标**
  - STOI: (客观) 语音可懂度
  - PESQ:语音质量

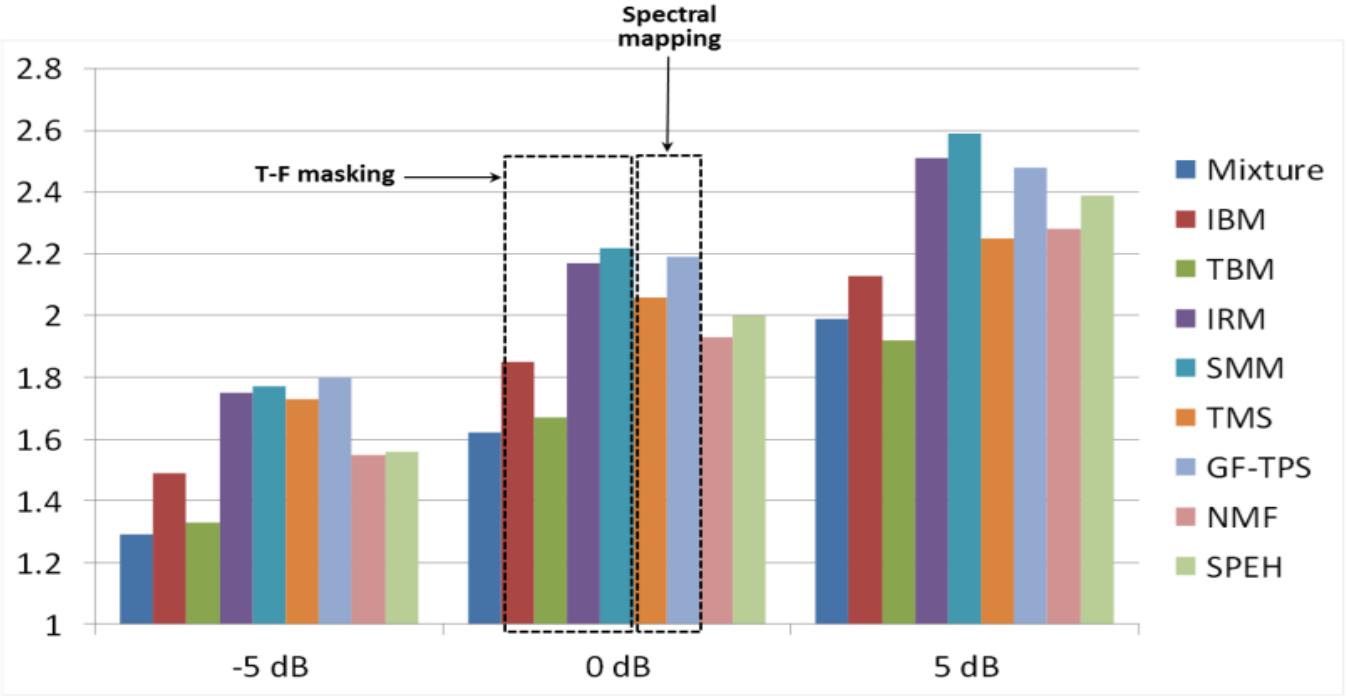
STOI比较结果

- 掩码优于频谱映射



PESQ比较结果

- 比率掩码中的 **IRM** 和频谱映射中的**GF- PTS** 能得到最好的语音质量
- 比率掩码产生的语音质量要显著优于二值掩码



提纲

一、简介

二、训练目标

三、特征

四、增强算法

五、总结

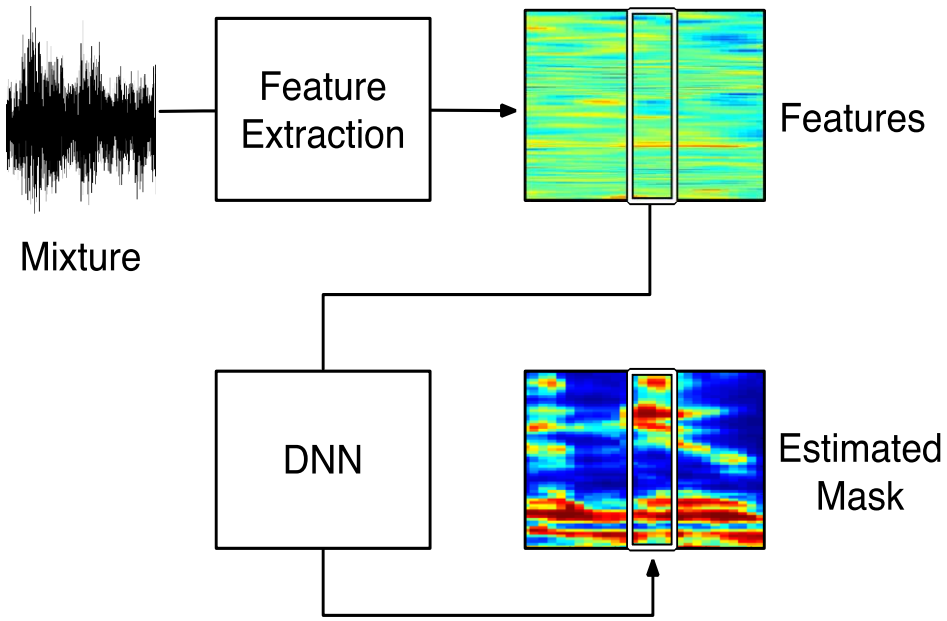
声学特征

- **早期的语音分离算法只使用少数几种特征**
  - 耳间时间延迟/耳间声压差 (ITD/ILD) (Roman et al.'03)
  - Pitch (Jin & Wang'09)
  - 短时傅里叶变换的幅度谱 (AMS) (Kim et al.'09)

- **近年来，基于有监督深度学习的语音分离算法的特征**
  - 互补特征集: AMS+RASTA-PLP+MFCC (Wang et al.'13)
  - 新特征: MRCG (multi-resolution cochleagram) (Chen et al.'14)

声学特征

- **Chen et al. (2014)和Delfarah and Wang (2017) 比较了特征在混响和加性噪声条件下的语音分离性能**
  - IRM作为训练目标
  - SNR= -5 dB



STOI的相对提升 (%)

- **Gammatone域的特征 (MRCG, GF和GFCC) 具有较强的抗噪声能力**
- **调制域特征(例如pitch, PLP等) 抗噪声能力较差**
- **不经处理的原始声学信号目前不是良好的特征表示**

Feature	Matched noise			Unmatched noise			Cochannel			Average
	Anechoic	Sim. RIRs	Rec. RIRs	Anechoic	Sim. RIRs	Rec. RIRs	Anechoic	Sim. RIRs	Rec. RIRs	
MRCG	**7.12**	**14.25**	**12.15**	**7.00**	7.28	8.99	21.25(13.00)	22.93 (13.19)	21.29 (12.81)	**12.92**
GF	6.19	13.10	11.37	6.71	7.87	8.24	22.56(11.87)	**23.95** (12.31)	22.35 (12.87)	12.71
GFCC	5.33	12.56	10.99	6.32	6.92	7.01	**23.53 (14.34)**	**23.95 (14.01)**	**22.76 (13.90)**	12.50
LOG-MEL	5.14	12.07	10.28	6.00	6.98	7.52	21.18 (13.88)	22.75 (13.54)	21.71 (13.18)	12.08
LOG-MAG	4.86	12.13	9.69	5.75	6.64	7.19	20.82 (13.84)	22.57 (13.40)	21.82 (13.55)	11.91
GFB	4.99	12.47	11.51	6.22	7.01	7.86	19.61 (13.34)	20.86 (11.97)	19.97 (11.60)	11.75
PNCC	1.74	8.88	10.76	2.18	**8.68**	**10.52**	19.97 (10.73)	19.47 (10.03)	19.35 (9.56)	10.78
MFCC	4.49	11.03	9.69	5.36	5.96	6.26	19.82 (11.98)	20.32 (11.47)	19.66 (11.54)	10.72
RAS-MFCC	2.61	10.47	9.56	3.08	6.74	7.37	18.12 (11.38)	19.07 (11.19)	17.87 (10.30)	10.44
AC-MFCC	2.89	9.63	8.89	3.31	5.61	5.91	18.66 (12.50)	18.64 (11.59)	17.73 (11.27)	9.87
PLP	3.71	10.36	9.10	4.39	5.03	5.81	16.84 (11.29)	16.73 (10.92)	15.46 (9.50)	9.46
SSF-II	3.41	8.57	8.68	4.18	5.45	6.00	16.76 (10.07)	17.72 (9.18)	18.07 (8.93)	9.09
SSF-I	3.31	8.35	8.53	4.09	5.17	5.77	16.25 (10.44)	17.70 (9.40)	18.04 (9.35)	8.97
RASTA-PLP	1.79	7.27	8.56	1.97	6.62	7.92	11.03 (6.76)	10.96 (6.06)	10.27 (6.28)	7.46
PITCH	2.35	4.62	4.79	3.36	3.36	4.61	19.71 (9.37)	17.82 (8.45)	16.87 (6.72)	7.03
GFMC	-0.68	7.05	5.00	-0.54	4.44	4.16	5.04 (-0.07)	6.01 (0.33)	4.97 (0.28)	4.40
WAV	0.94	2.32	2.68	0.02	0.99	1.63	11.62 (4.81)	11.92 (6.25)	10.54 (1.05)	3.89
AMS	0.31	0.30	-1.38	0.19	-2.99	-3.40	11.73 (5.96)	10.97 (6.76)	10.20 (4.90)	1.71
PAC-MFCC	0.00	-0.33	-0.82	0.18	-0.92	-0.67	0.95 (0.15)	1.25 (0.26)	1.17 (0.09)	-0.17

提纲

一、简介

二、训练目标

三、特征

四、算法

五、总结

第四部分：有监督语音分离算法

单通道语音分离算法

- 分离语音与非语音（语音增强）
- 多说话人语音分离
- 去混响

固定阵列的多通道语音分离算法

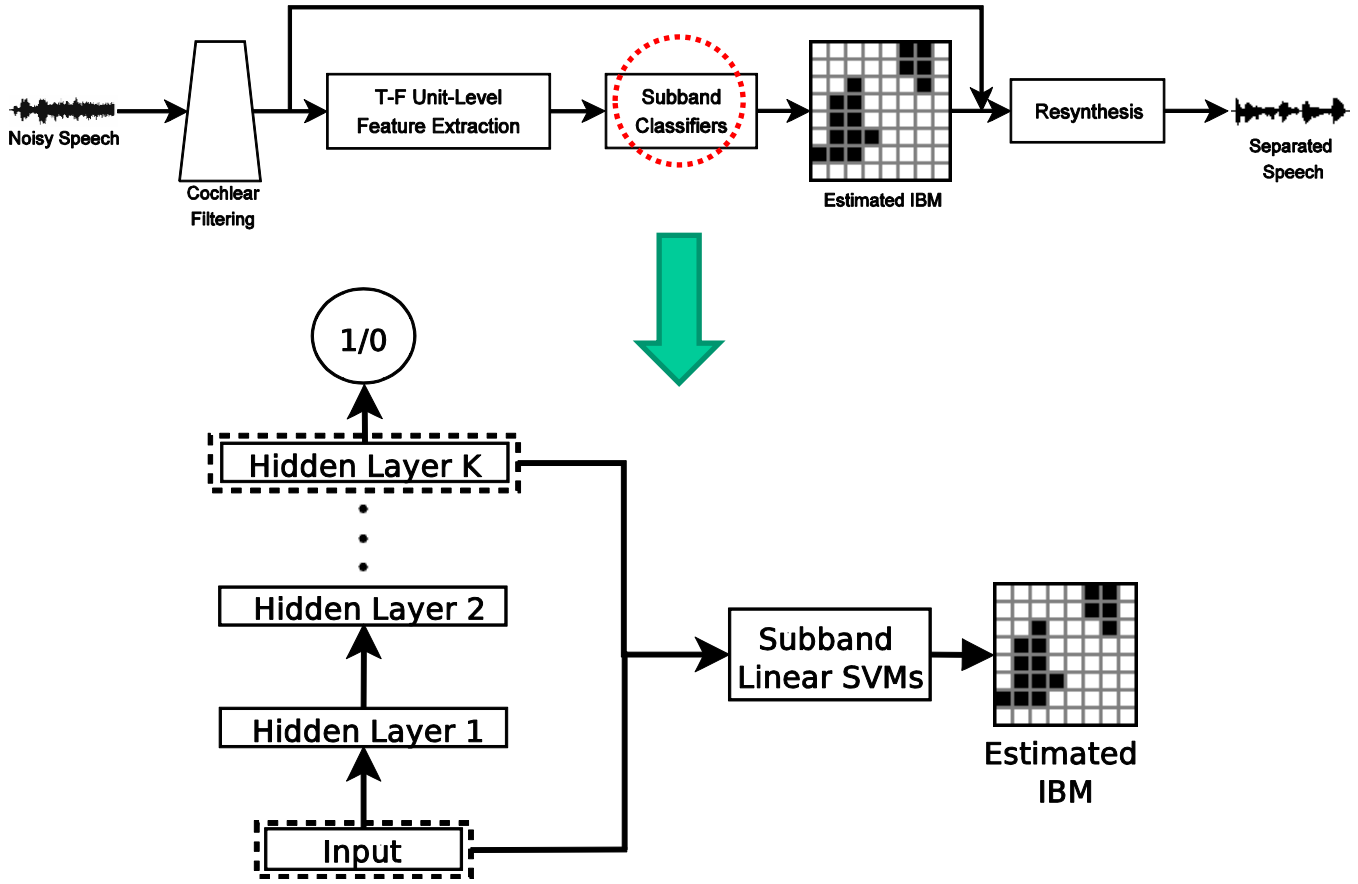
- 基于空间特征的语音分离
- 基于掩码的波束形成

随机阵列的多通道语音分离算法

- 深度Ad-Hoc波束形成

首个基于深度学习的语音分离算法

- **Y. Wang & Wang (2013) 首次将DNN应用于语音分离问题**

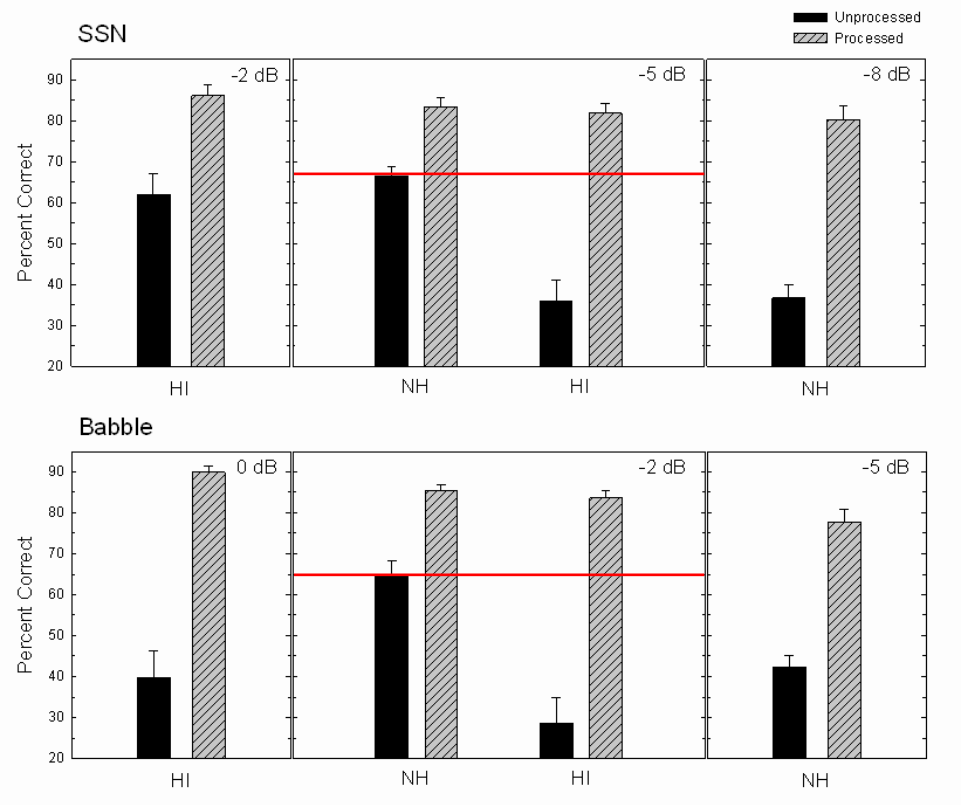


首个关于DNN的泛化能力的实验

- **训练数据集：**
  - 从IEEE数据集中选取了200 句话
  - 与100 种环境噪声在 0 dB下混合
  - 大约17小时时长
- **测试数据集：**
  - 20 个未知说话人
  - 20种未知噪声类型上进行测试
  - 测试信噪比为0 dB
- **基于DNN的分类器取得了当时state-of-the-art的性能**

主观语音可懂度

- **Healy et al. (2013) 采用主观评测的方法测试了基于DNN的语音分离性能**
- **能改善无听力障碍的人的听力，也能改善有听力障碍的人的听力**
- **有听力障碍的人+增强后的语音>听力正常的人+原始含噪语音**



泛化能力问题：大规模多条件训练

- **尽管上述实验结果超过了传统语音分离算法，但是一个突出问题在于训练和测试不匹配时性能会出现显著下降**
  - 说话人、说话内容等不匹配
  - 噪声类型不匹配
- **该问题可以通过DNN的大规模多条件训练加以解决 (Chen et al.'16)**
  - 训练数据集：
    - 560句IEEE语句
    - 10,000 个噪声段（125个小时）
    - 构造了640,000句含噪语音段（380个小时）
  - 测试语句和噪声都不同于训练集

-2 dB环境下的STOI实验结果

	Babble	Cafeteria	Factory	Babble2	平均
未处理的语音	0.612	0.596	0.611	0.611	0.608
100种噪声	0.683	0.704	0.750	0.688	0.706
10000种噪声	0.792	**0.783**	**0.807**	**0.786**	**0.792**
训练测试匹配	**0.833**	0.770	0.802	0.762	**0.792**

- 实验结论：大规模多条件训练≈训练测试匹配

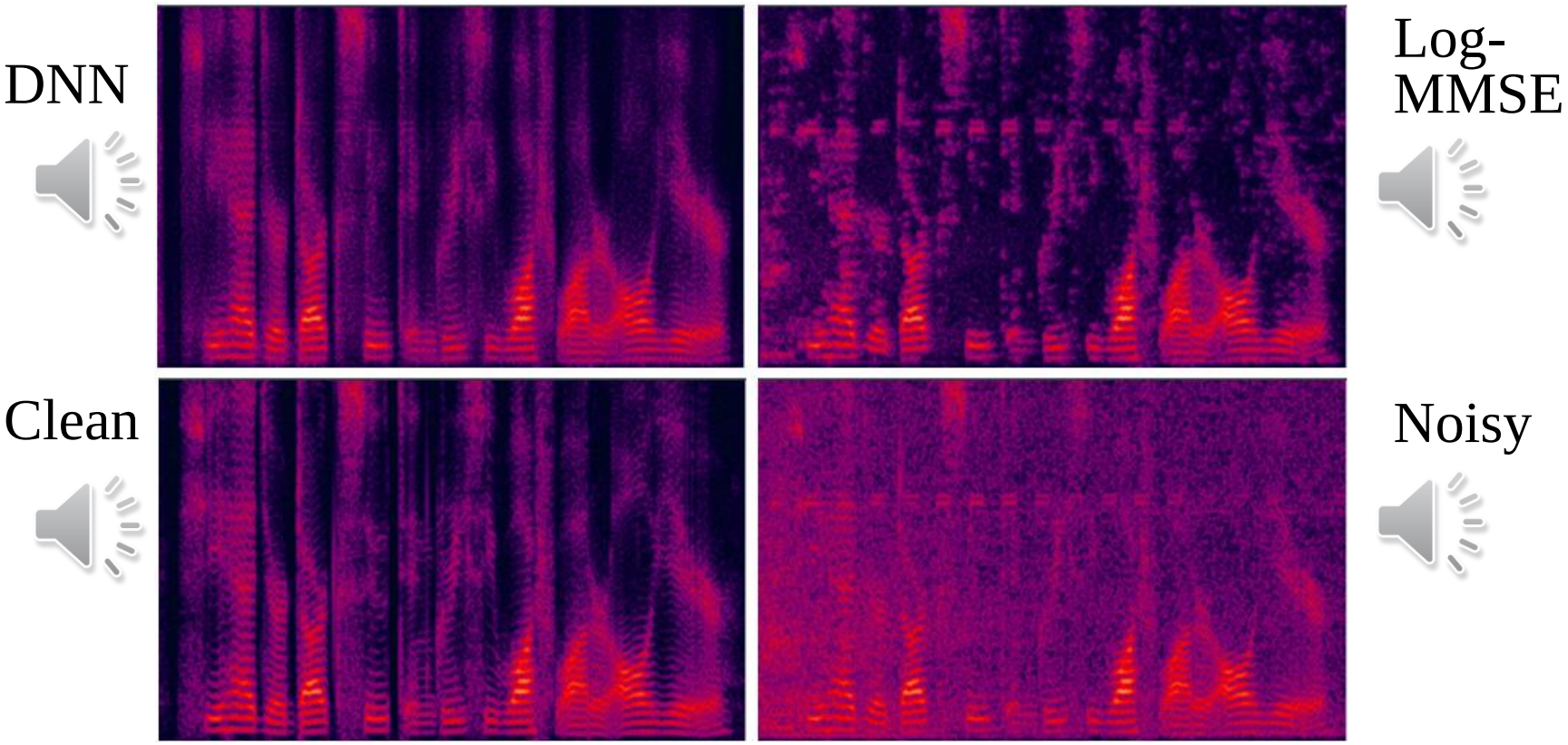
基于DNN的spectral mapping方法

- **Xu et al. (2014)提出了基于DNN的spectral mapping增强方法**
  - 使用RBM预训练DNN
  - Mapping: 含噪语音的对数功率谱→纯净语音的对数功率谱

	Noisy	L-MMSE	SNN	DNN_1	DNN_2	DNN_3	DNN_4
SNR20	2.99	3.32	3.48	3.46	3.59	**3.60**	3.59
SNR15	2.65	2.99	3.26	3.24	3.35	**3.36**	**3.36**
SNR10	2.32	2.65	2.99	2.97	3.08	**3.10**	3.09
SNR5	1.98	2.30	2.68	2.65	2.76	**2.78**	**2.78**
SNR0	1.65	1.93	2.32	2.29	2.38	**2.41**	**2.41**
SNR-5	1.38	1.55	1.92	1.89	1.95	**1.97**	**1.97**
Ave	2.16	2.46	2.78	2.75	2.85	**2.87**	**2.87**

训练测试匹配环境下的PESQ结果　　DNN的下标表示隐藏层的数量

Xu et al. (2014) 的演示



Street噪声 10 dB

第四部分：有监督语音分离算法

单通道语音分离算法

- 分离语音与非语音（语音增强）
- 多说话人语音分离
- 去混响

固定阵列的多通道语音分离算法

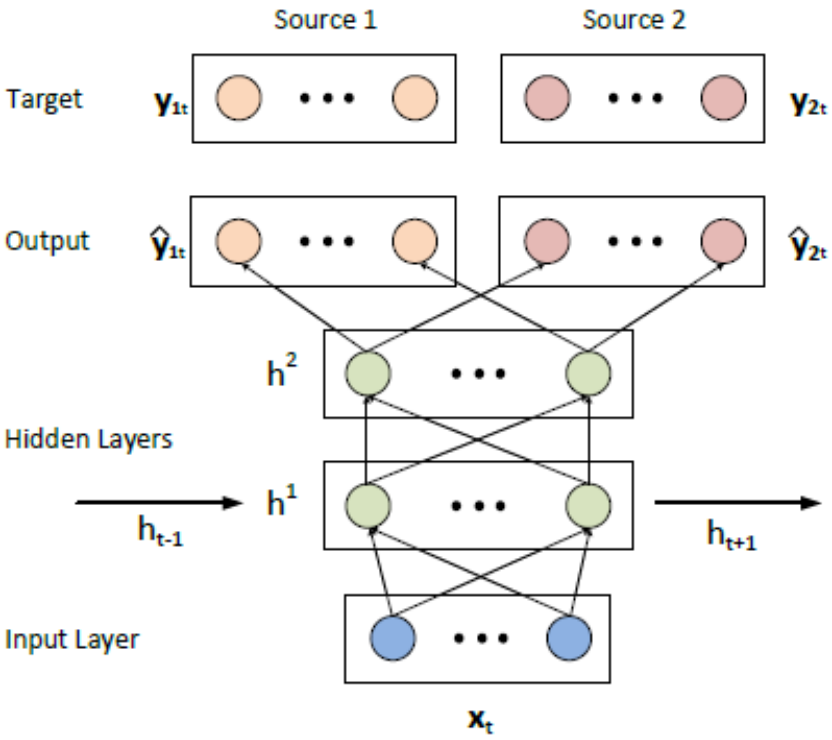
- 基于空间特征的语音分离
- 基于掩码的波束形成

随机阵列的多通道语音分离算法

- 深度Ad-Hoc波束形成

基于DNN的说话人分离

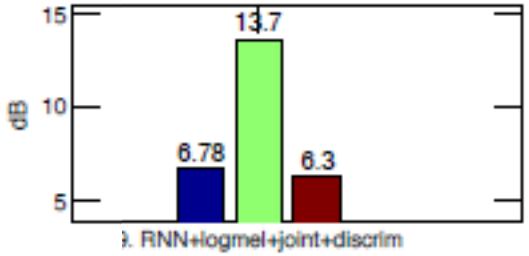
- **Huang et al. (2014; 2015) 提出了基于DNN的说话人分离算法**
  - DNN: 混合信号的时频谱→分离信号的时频谱
  - 使用T-F masking作为优化目标
    - IBM或者IRM



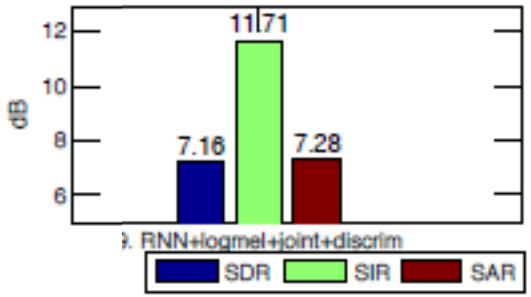
Huang et al. 实验结果

- **DNN与RNN的实验结果接近**
  - 同非负矩阵分解(NMF)相比，有4-5 dB 的SIR提升

二值掩码



比率掩码



混合信号

分离后的说话人1

原始信号说话人 1

分离后的说话人2

原始信号说话人 2

模型对目标说话人的依赖性问题

- **Huang et al.’s 的说话人分离算法是talker-dependent的，即训练和测试的两个说话人是相同的**
  - 注：训练和测试的语句内容可以是不同的

- **说话人分离算法可以分为三类**
  - Talker dependent：训练和测试的说话人完全相同
  - **Target dependent：训练和测试的目标说话人**
  - **Talker independent：训练和测试的说话人完全不同**

Target-dependent说话人分离算法

- **通过在训练阶段增加大量干扰说话人可以成功训练Target-dependent的说话人分离模型(Du et al.'14; Zhang & Wang'16)**

Talker-independent说话人分离算法

- **问题：**
  - 早期的Talker-independent 说话人分离算法可以看成是无监督聚类问题 (Bach & Jordan'06; Hu & Wang'13)，难以充分利用有监督学习中的鉴别性先验信息
  - DNN不能够通过构造许多成对的说话人得到成功训练

- **Deep clustering (Hershey et al.'16) 是首个基于DNN的talk-independent说话人分离算法，核心思想是将有监督DNN和无监督聚类结合**

Deep clustering

- **给定一句话的T-F unit的IBM $\boldsymbol{Y}$，其伴随矩阵可以定义为**

$$\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{Y}^T$$

  - $\boldsymbol{Y}$ 是IBM矩阵：其中当第i个unit属于第c个说话人，则$Y_{i,c} = 1$，否则$Y_{i,c} = 0$
  - 如果$\boldsymbol{Y}$的第i个unit和第j个unit属于同一个说话人，则 $A_{i,j} = 1$，否则$A_{i,j} = 0$
- **Deep clustering的训练目标是$\boldsymbol{A}$ 矩阵**
- **为了得到$\boldsymbol{A}$ 的估计，DNN为每一个T-F unit 学习一个嵌入式表示，使得训练风险最小**

Deep clustering

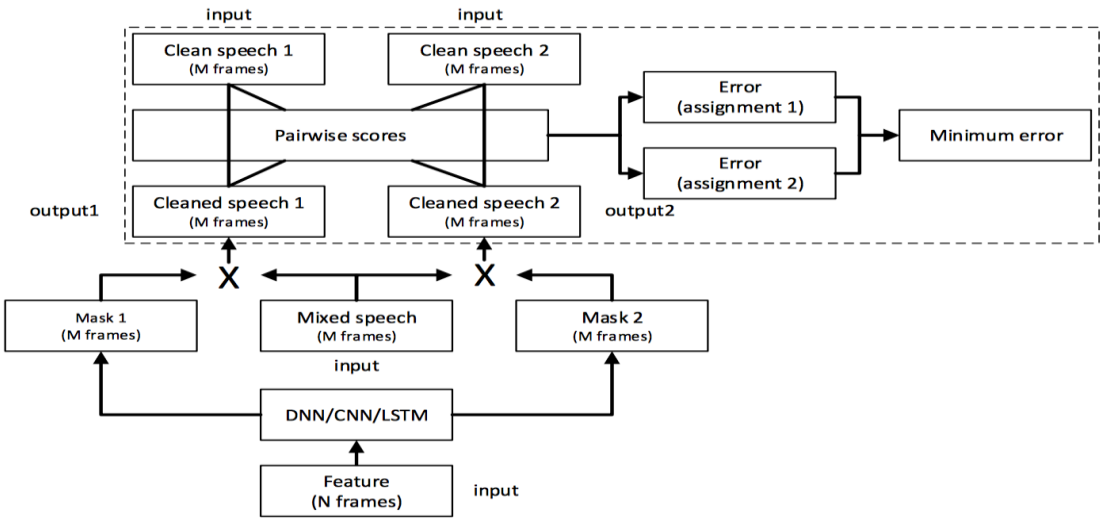
- **DNN 的训练目标**

$$C_{\boldsymbol{Y}}(\boldsymbol{V}) = \left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\right\|_F^2 = \left\|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{Y}\boldsymbol{Y}^T\right\|_F^2$$

  - $\boldsymbol{V}$ 是T-F units的嵌入式表示矩阵，其每一行表示一个T-F unit的嵌入式表示向量
  - $\|\cdot\|_F^2$ 表示squared Frobenius norm
- **在测试阶段，首先用DNN产生每个T-F unit的嵌入式表示，然后用K-means 算法对T-F units的嵌入式表示做聚类**

Permutation invariant training (PIT)

- **PIT算法(Kolbak et al.'17) 原理：**
  - 在训练阶段，speaker A, speaker B → AB? BA?
  - AB → MSE1, BA → MSE2
  - MSE = min(MSE1,MSE2)



两种PIT

- **帧级的PIT (tPIT): 对每一帧都计算排列问题，因此需要追踪说话人(speaker tracing)，将帧级别的分离结果串列起来**

- **语句级别PIT (uPIT): 对整句混叠的语音只计算一次排列，不需要追踪说话人**

基于CASA的方法

- **Deep clustering 和 PIT 的缺陷**
  - 对于deep clustering，难以区分能量相似的T-F units
  - tPIT比uPIT 的性能好(尤其是对同性说话人的情况), 但是tPIT需要追踪说话人，有很高的复杂度

- **Liu & Wang (2018)结合PIT和deep clustering，提出基于CASA的方法**
  - 首先将每帧数据分离成多个频域表示
  - 然后在时间轴上将不同频域表示的数据分配到特定说话人

Talker-independent语音分离方法实验比较

SDR 实验结果 (dB)

	Same Gender	Different Gender	Overall
Deep clustering++ (Isik et al., 2016)	9.4	12.0	10.8
PIT (Kolbæk et al., 2017)	7.5	12.2	10.0
CASA (Liu and Wang, 2018)	10.3	12.6	11.5

两个男性说话人混叠语音 

说话人1 -uPIT 

说话人2 - uPIT 

说话人1 – DC++ 

说话人2 – DC++ 

说话人1 - CASA 

说话人2 - CASA 

说话人1 - clean 

说话人2 - clean 

第四部分：有监督语音分离算法

单通道语音分离算法

- 分离语音与非语音（语音增强）
- 多说话人语音分离
- 去混响

固定阵列的多通道语音分离算法

- 基于空间特征的语音分离
- 基于掩码的波束形成

随机阵列的多通道语音分离算法

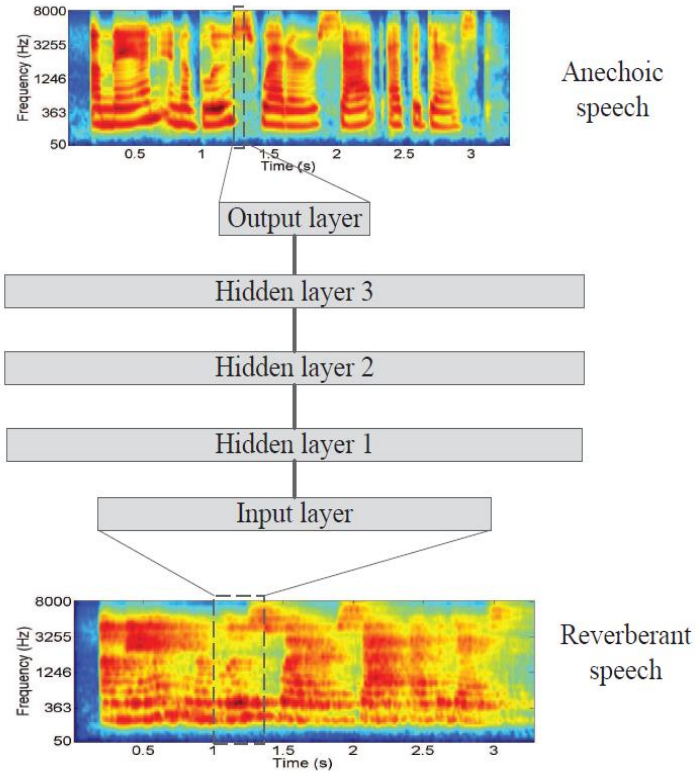
- 深度Ad-Hoc波束形成

去混响

- **混响是语音信号处理非常困难的问题，特别是存在加性噪声的情况**
  - 语音通信
  - 语音识别
  - 说话人识别
- **已有的工作**
  - Inverse filtering (Avendano & Hermansky'96; Wu & Wang'06)
  - 二值掩码 (Roman & Woodruff'13; Hazrati et al.'13)

基于DNN的去混响算法

- **Han et al.'14 首次提出使用DNN学习冲击响应的逆过程**
  - DNN: 带混响的语音谱→纯净语音的语音谱
  - 在STFT幅度谱和cochleagram特征上表现一样良好
  - Han et al.'15 提出了同时去混响和环境噪声

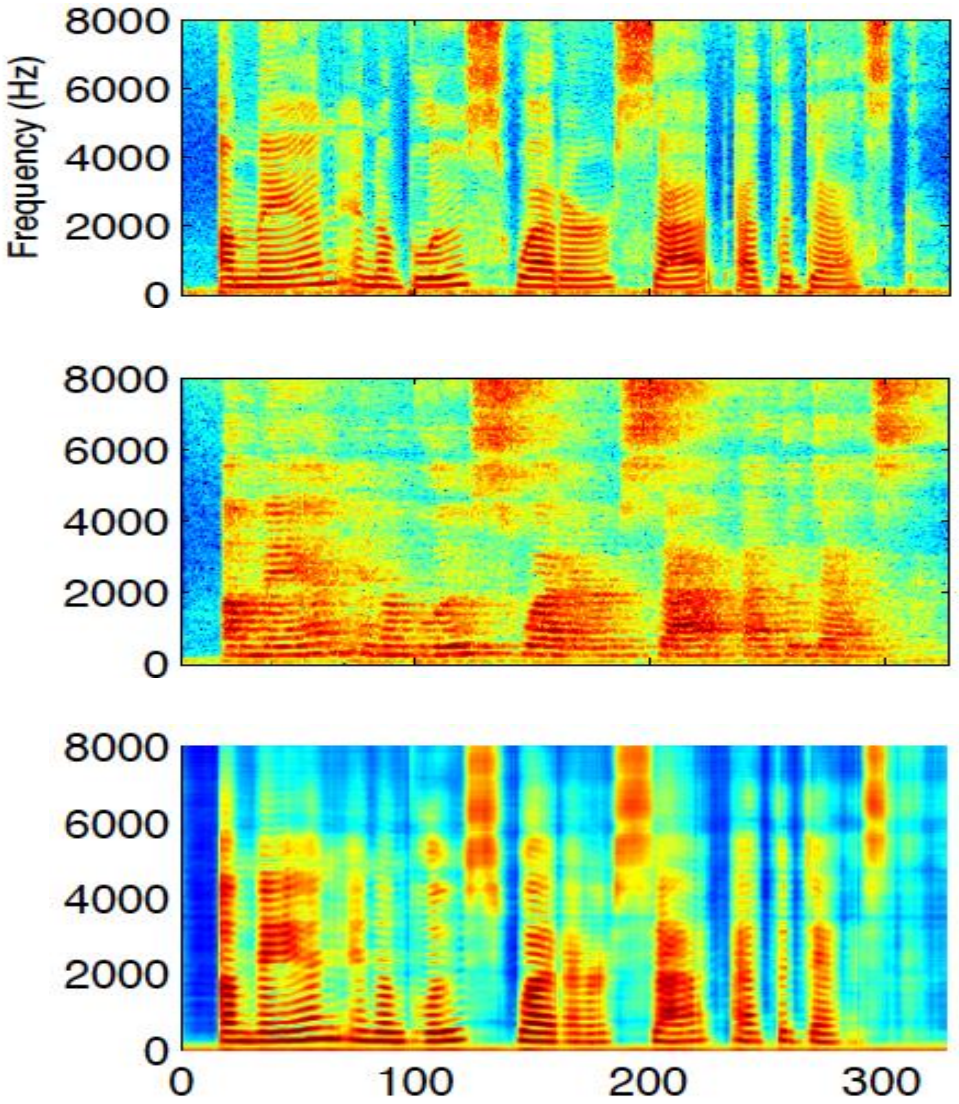


示意图

纯净语音

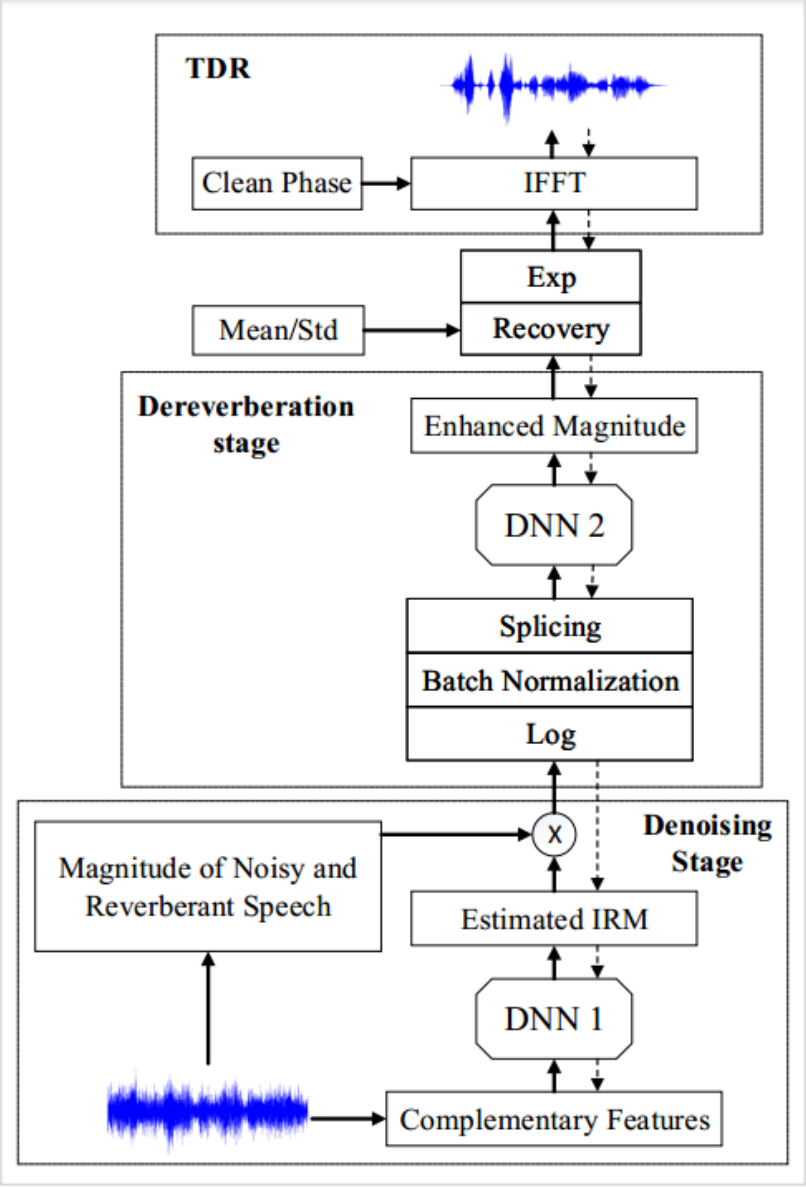
含混响的语音
(T_{60} = 0.6 s)

去混响后的语音



Two-stage model算法

- **环境噪声和混响是两种不同类型的噪声**
  - 环境噪声是加性噪声
  - 混响是语音信号与空间冲激响应的卷积，是频域上的乘性噪声
- **Zhao et al. (2017) 针对这两种噪声的不同性质提出了two-stage model**



实验结果

- **混响时间为(0.3 s, 0.6 s, 0.9 s)、信噪比等级为(-6 dB, 0 dB, 6 dB)、以及四种噪声环境(babble, SSN, DLIVING, PCAFETER)下的平均结果**

	STOI (in %)	PESQ
原始语音	61.8	1.22
去加性噪声	77.7	1.81
Two-stage model	82.6	2.08

- **Demo (T60 = 0.6 s, babble 噪声, SNR=3 dB)**


原始语音


去加性噪声


Two-stage model

第四部分：有监督语音分离算法

单通道语音分离算法

- 分离语音与非语音（语音增强）
- 多说话人语音分离
- 去混响

固定阵列的多通道语音分离算法

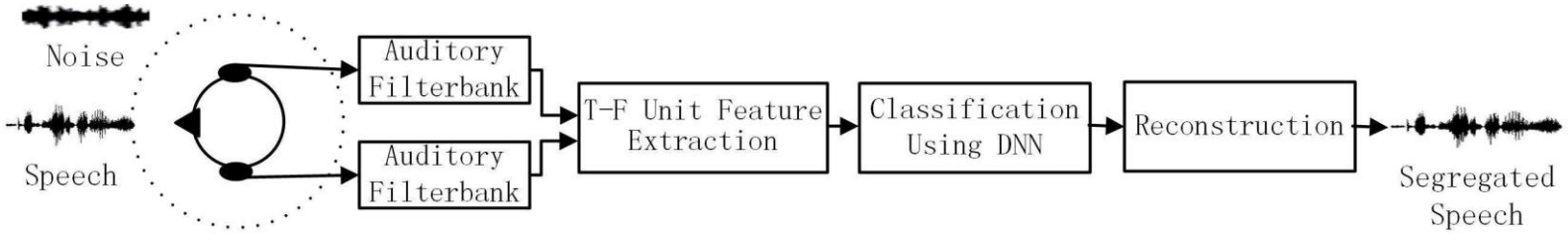
- 基于空间特征的语音分离
- 基于掩码的波束形成

随机阵列的多通道语音分离算法

- 深度Ad-Hoc波束形成

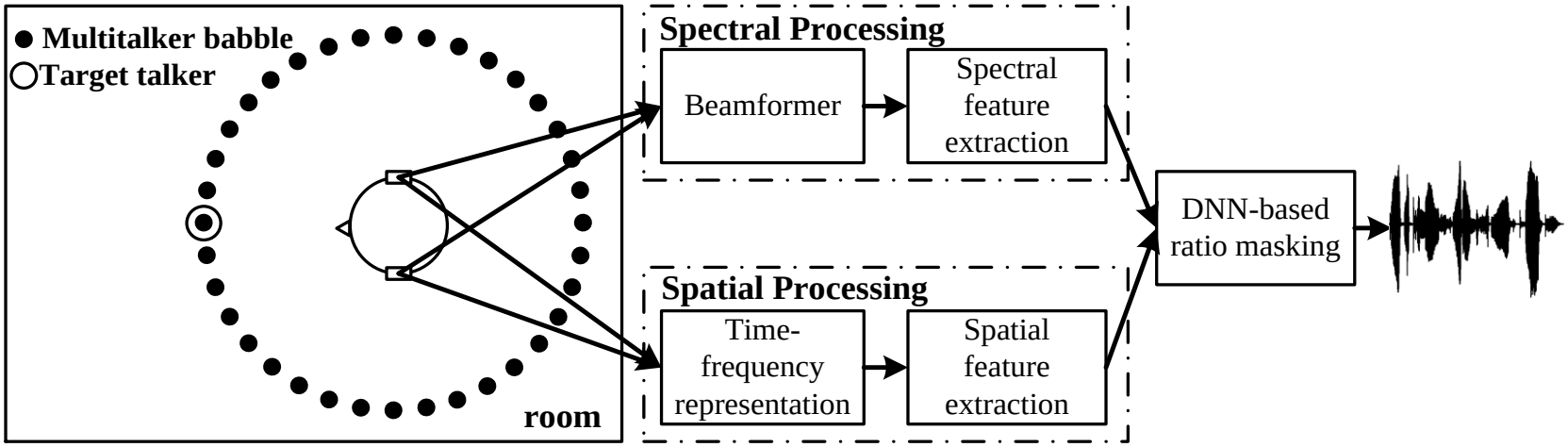
基于空间特征的语音分离算法

- **Jiang et al. (2014) 提出了使用双耳特征作为单通道DNN的输入特征，以IBM作为训练目标，用于去混响**
  - 双耳时间差 (ITD) 和双耳声压差 (ILD)



将频域特征和空间特征结合

- **Zhang & Wang (2017)将频域特征和空间特征结合，作为单通道DNN的输入**
  - 频域特征：用beamforming将多通道信号抽取为单通道信号，然后从单通道输出信号中抽取声学特征
  - 空间特征：ITD、ILD
  - 该方法显著超过传统beamformer



第四部分：有监督语音分离算法

单通道语音分离算法

- 分离语音与非语音（语音增强）
- 多说话人语音分离
- 去混响

固定阵列的多通道语音分离算法

- 基于空间特征的语音分离
- 基于掩码的波束形成

随机阵列的多通道语音分离算法

- 深度Ad-Hoc波束形成

基于掩码的波束形成

- **麦克风阵列的接收信号可以表示为**

$$\mathbf{y}(t,f) = \mathbf{c}(f)s(t,f) + \mathbf{n}(t,f)$$
$$= \mathbf{x}(t,f) + \mathbf{n}(t,f)$$

  - $\mathbf{y}(t,f)$ 和$\mathbf{n}(t,f)$ 表示在t时刻和第f个频段上阵列接收到的信号向量和噪声向量
  - $s(t,f)$: 语音源
  - $\mathbf{c}(f)$: 空间传递函数
  - $\mathbf{c}(f)s(t,f)$: 接收到的语音信号

基于掩码的波束形成

- **MVDR滤波器旨在保持DOA方向能量不变的同时，最小化噪声的输出功率**

$$\mathbf{w}_{\text{opt}} = \underset{\mathbf{w}}{\text{argmin}}\{\mathbf{w}^H \mathbf{\Phi}_n \mathbf{w}\}, \qquad \text{subject to } \mathbf{w}^H \mathbf{c} = 1$$

$$\mathbf{w}_{\text{opt}} = \frac{\mathbf{\Phi}_n^{-1}\mathbf{c}}{\mathbf{c}^H \mathbf{\Phi}_n^{-1}\mathbf{c}}$$

  - H 表示共轭转置
  - $\mathbf{\Phi}_n$表示噪声的空间协方差矩阵
  - $\mathbf{c}$是纯净语音（的估计）的空间协方差矩阵$\mathbf{\Phi}_x$的最大特征向量
- **MVDR滤波:**

$$\tilde{s}(t) = \mathbf{w}_{\text{opt}}^H \mathbf{y}(t)$$

基于掩码的波束形成

- 在假设语音源和噪声是不相关的条件下, 存在

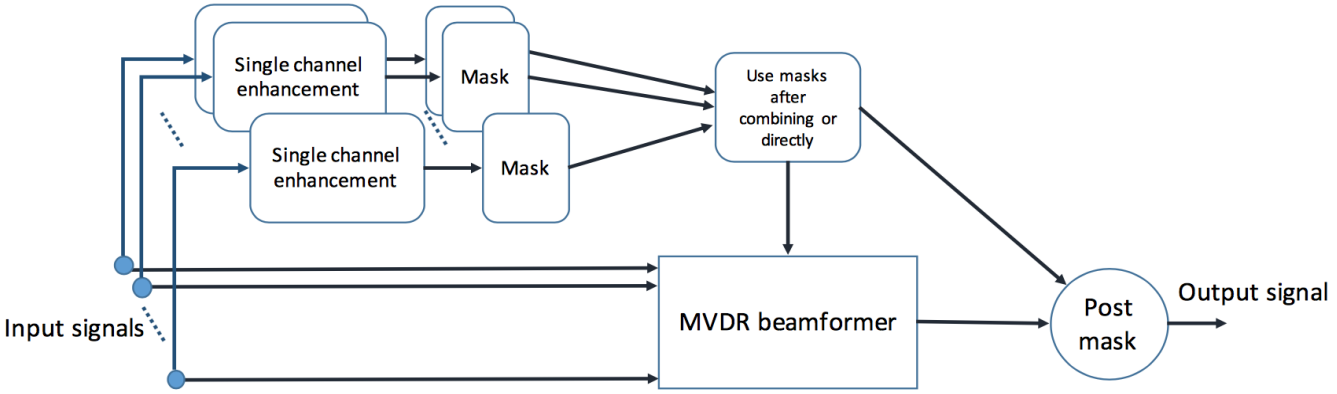
$$\boldsymbol{\Phi}_x = \boldsymbol{\Phi}_y - \boldsymbol{\Phi}_n$$

- 由上可知$\boldsymbol{\Phi}_n \rightarrow \boldsymbol{\Phi}_x \rightarrow \mathbf{c}(f)$， $\boldsymbol{\Phi}_n$ 的求解是关键

- 问题：
  - 在有混响和强噪声的环境里， $\boldsymbol{\Phi}_n$的估计是不够准确的

基于掩码的波束形成

- **Heymann et al.'16; Higuchi et al.'16提出基于DNN的MVDR beamforming**
  - 使用基于DNN的T-F masking的方法进行单通道噪声估计
  - 计算噪声的空间协方差矩阵$\boldsymbol{\Phi}_n$
  - 使用MVDR 进行beamforming



- **基于DNN的MVDR在CHiME-3 and CHiME-4噪声环境下的语音识别率显著优于传统MVDR**

第四部分：有监督语音分离算法

单通道语音分离算法

- 分离语音与非语音（语音增强）
- 多说话人语音分离
- 去混响

固定阵列的多通道语音分离算法

- 基于空间特征的语音分离
- 基于掩码的波束形成

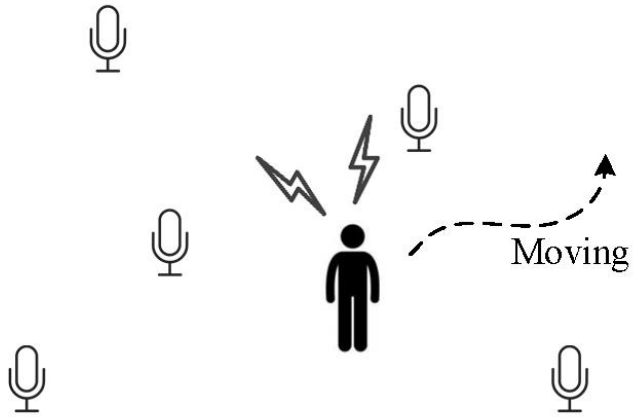
随机阵列的多通道语音分离算法

- 深度Ad-Hoc波束形成

问题描述

- **基于DNN的语音分离受设备与说话人之间的距离影响大**
- **基于DNN的固定阵列多通道算法并没有显著优于基于DNN的单通道算法**

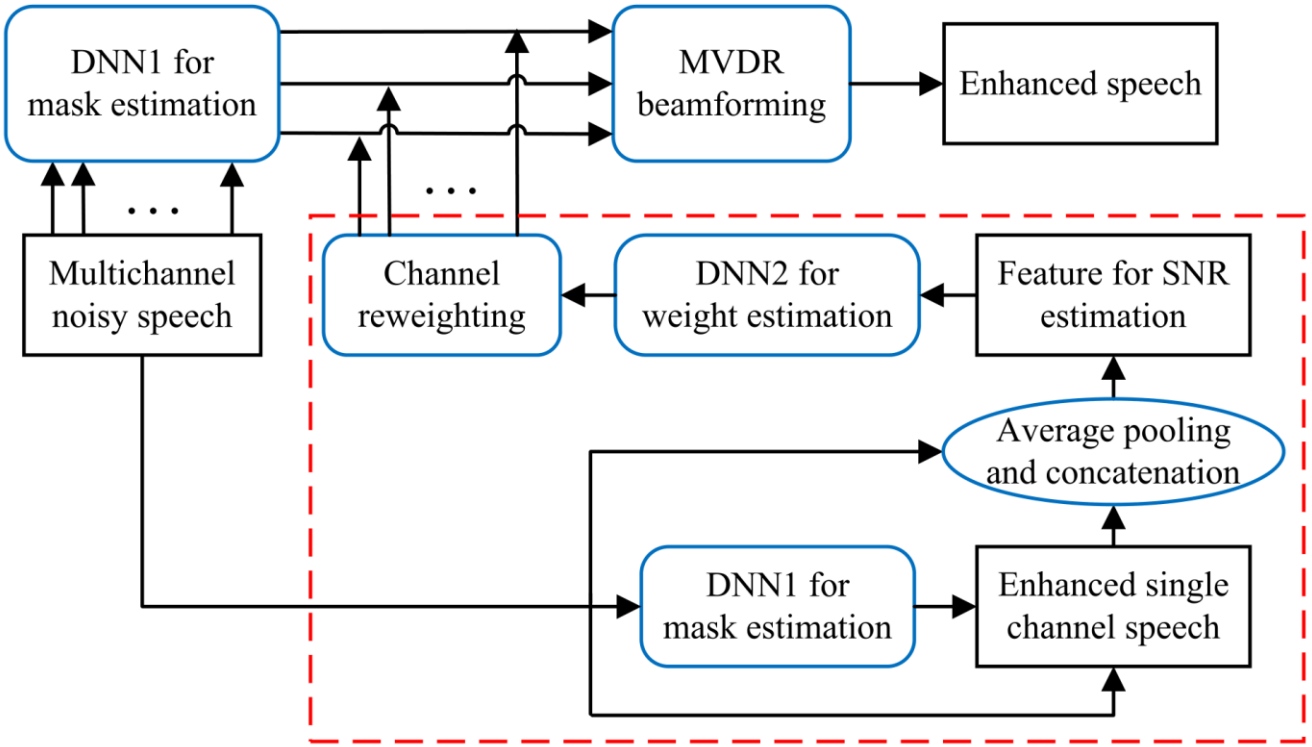
- **随机阵列(Ad-hoc microphone array, distributed microphone array)(Heusdens et al., 2012)的研究通常假设理想的噪声估计和语音检测**



深度Ad-Hoc波束形成

- **Zhang et al. 2018 提出了深度Ad-Hoc波束形成，与固定阵MVDR不同之处在于，它需要对每个信道计算权重**

$$\mathbf{y_p}(t,f) = \mathbf{p} \odot \mathbf{y}(t,f) = \mathbf{p} \odot \mathbf{c}(f)s(t,f) + \mathbf{p} \odot \mathbf{n}(t,f)$$



算法描述

- **信道权重：真实信噪比**
- **信道权重的估计：用DNN2预测每个microphone的信噪比**
  - DNN2的输入：
    - 含噪语音段average pooling后得到的矢量**a**
    - 增强语音段average pooling后得到的矢量**b**
  - DNN2的输出：
    - 预测信噪比 $p = f_{\mathrm{DNN2}}([\mathbf{a}, \mathbf{b}])$
- **MVDR算法以$\mathbf{y_p}(t, f)$ 作为输入**

实验结果

STOI

实验场景	噪声类型	原始语音	单通道算法	固定阵多通道算法	Ad-Hoc多通道算法
(2,14,8)	Babble	0.6967	0.7529	0.7680	**0.8417**
	Factory	0.7045	0.7727	0.8077	**0.8853**
(2,18,10)	Babble	0.6523	0.6990	0.6971	**0.8007**
	Factory	0.6518	0.7196	0.7560	**0.8119**

第五部分: 总结

- **基于深度学习的语音分离的核心组件：**
  - 将语音分离构造为分类问题
  - 构造掩码，作为训练目标
- **基于深度学习的语音分离的贡献：**
  - 大幅提升语音算法和系统的抗噪声能力
  - 首次在含噪声环境下显著提升了人的语音可懂度
  - 提升了beamforming的性能
- **一篇综述性文献**

Wang D.L. and Chen J. (2018): Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702-1726.

参考文献

Ahmadi, Gross, & Sinex (2013) JASA 133: 1687-1692.

Anzalone et al. (2006) Ear & Hearing 27: 480-492.

Avendano & Hermansky (1996) ICSLP: 889-892.

Bach & Jordan (2006) JMLR 7: 1963-2001.

Bronkhorst & Plomp (1992) JASA 92: 3132-3139.

Brungart et al. (2006) JASA 120: 4007-4018.

Cao et al. (2011) JASA 129: 2227-2236.

Chen et al. (2014) IEEE/ACM T-ASLP 22: 1993-2002.

Chen et al. (2016) JASA 139: 2604-2612.

Cherry (1957) On Human Communication. Wiley.

Darwin (2008) Phil Trans Roy Soc B 363: 1011-1021.

Delfarah & Wang (2017) IEEE/ACM T-ASLP 25: 1085-1094.

Dillon (2012) Hearing Aids (2nd Ed). Boomerang.

Du et al. (2014) ICSP: 65-68.

Erdogan et al. (2015) ICASSP: 708-712.

Erdogan et al. (2016) Interspeech: 1981-1985.

Gonzales & Brookes (2014) ICASSP: 7029-7033.

Han et al. (2014) ICASSP: 4628-4632.

Han et al. (2015) IEEE/ACM T-ASLP 23: 982-992.

Han & Wang (2014) IEEE/ACM T-ASLP 22: 2158-2168.

Hazrati et al. (2013) JASA 133: 1607-1614.

Healy et al. (2013) JASA 134: 3029-3038.

Helmholtz (1863) On the Sensation of Tone. Dover.

Hendriks et al. (2010) ICASSP: 4266-4269.

Hershey et al. (2016) ICASSP: 31-35.

Heymann et al. (2016) ICASSP: 196-200.

Higuchi et al. (2016) ICASSP: 5210-5214.

Hu & Wang (2004) IEEE T-NN 15: 1135-1150.

Hu & Wang (2013) IEEE T-ASLP 21: 122-131.

Huang & Lee (2013) IEEE T-ASLP 21: 99-109.

Huang et al. (2014) ICASSP: 1581-1585.

Huang et al. (2015) IEEE/ACM T-ASLP 23: 2136-2147.

Hummersone et al. (2014) In Blind Source Separation, Springer.

Isik et al. (2016) Interspeech: 545-549.

Jiang et al. (2014) IEEE/ACM T-ASLP 22: 2112-2121.

Jin & Wang (2009) IEEE T-ASLP 17: 625-638.

Kim & Stern (2012) ICASSP: 4101-4104.

Kim et al. (2009) JASA 126: 1486-1494.

参考文献

Kjems et al. (2009) JASA 126: 1415-1426.

Kolbak et al. (2017) IEEE/ACM T-ASLP: 153-167.

Li & Loizou (2008) JASA 123: 1673-1682.

Li & Wang (2009) Speech Comm. 51: 230-239.

Liu & Wang (2018) ICASSP: 5399-5403.

Loizou & Kim (2011) IEEE T-ASLP 19: 47-56.

Lu et al. (2013) Interspeech: 555-559.

Narayanan & Wang (2013) ICASSP: 7092-7096.

Papadopoulos et al. (2016) IEEE/ACM T-ASLP 24: 2495-2506.

Pertila & Cekar (2017) ICASSP: 6125-6129.

Roman & Woodruff (2013) JASA 133: 1707-1717.

Roman et al. (2003) JASA 114: 2236-2252.

Shao et al. (2008) ICASSP: 1589-1592.

Srinivasan et al. (2006) Speech Comm. 48: 1486-1501.

Virtanen et al. (2013) IEEE T-ASLP 21: 2277-2289.

Wang (March 2017) IEEE Spectrum: 32-37.

Wang & Brown, Ed. (2006) Computational Auditory Scene Analysis. Wiley & IEEE Press.

Wang & Chen (2018) IEEE/ACM T-ASLP 26: 1702-1726.

Wang et al. (2008) JASA 124: 2303-2307.

Wang et al. (2009) JASA 125: 2336-2347.

WangY & Wang (2013) IEEE T-ASLP 21: 1381-1390.

WangY et al. (2013) IEEE T-ASLP 21: 270-279.

WangY et al. (2014) IEEE/ACM T-ASLP 22: 1849-1858.

WangY & Wang (2015) ICASSP: 4390-4394.

Wenninger et al. (2014) GlobalSIP MLASP Symp.

Williamson et al. (2016) IEEE/ACM T-ASLP 24: 483-492.

Wu & Wang (2006) IEEE T-ASLP 14: 774-784.

Xu et al. (2014) IEEE Sig. Proc. Lett. 21: 65-68.

Yilmaz & Rickard (2004) IEEE T-SP 52: 1830-1847.

Zhang & Wang (2016) IEEE/ACM T-ASLP 24: 967-977.

Zhang & Wang (2017) IEEE/ACM T-ASLP 25: 1075-1084.

Zhang et al. (2017) ICASSP: 276-280.

Zhang & Wu (2013) IEEE T-ASLP 21: 697-710.

Zhao et al. (2017) ICASSP: 5580-5584.