Report

Understanding of Speech Separation and Overlap Speech diarization

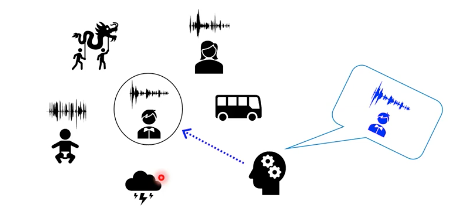
**Zhaohui Yin**

**Tue, Dec 01, 2020**

0. 基础知识学习

0.1 语音分离

参考视频：李宏毅教学：https://www.bilibili.com/video/BV1RE411g7rQ?p=12,Task 3.



**Speech Separation**：1）Speech Enhancement（降噪、声源分离）；2）Speaker Separation（单个麦克风）

|  |  |
| --- | --- |
|  |  |
| (a) Speech Enhancement | (b) Speaker Separation |

Figure 1: 语音分离分类

1. Training Data：直接进行混合即可。



Figure 2: WSJ0-2数据

2. Evalution（Scale Invariant signal-to-distortion ratio）[SI-SDR]:

|  |  |
| --- | --- |
|  |  |
| （a） | （b） |

Figure 2: 评估方法：a）SI-SDR的计算方法，b）改进的精度评估

3.Permutation Issue：如何处置输出先后的问题。

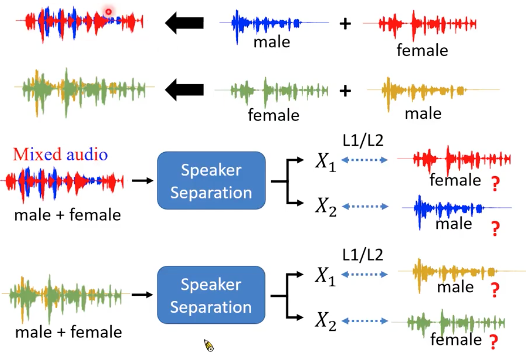
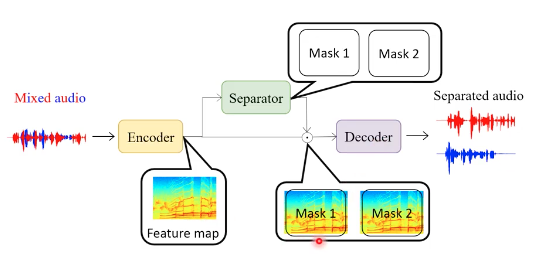


Figure 3: 存在输出顺序的问题

4.模型细节：

|  |  |  |
| --- | --- | --- |
| 模型 | 结构 | 描述 |
| Deep Clastering（？？？） | Ideal Binary Mask（IBM） | 1. 在没有真实标签的时候，很难得到mask矩阵；  2. 并非真的end-to-end，k-mean值聚类是不需要训练的；  3. 确实是可以工作的；  4. 存在问题：训练的时候输入和输出是什么？ |
| Permutation Invariant Training (PIT) |  | 1. 开始的时候训练是随机的，但是慢慢会找到合适的排序方法，最后会收敛    2. 训练过程中会慢慢收敛 |

6. 模型架构详解：TasNet - Time-Domain Audio Separation Network (Luomet al., TASLP'19)



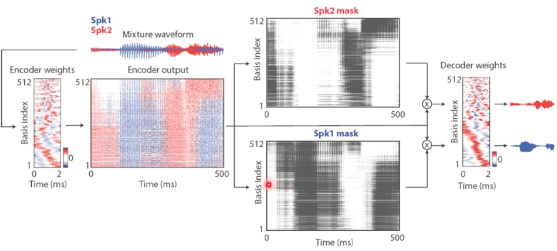


Figure 4: network架构

|  |  |  |
| --- | --- | --- |
| 模块名称 | 架构 | 过程变量实例 |
| Encoder  **（WSJ0-2mix）** |  |  |
| Separator |  | 1. 1-D 卷积神经网络  2. sigmoid函数是指结果区间 [0,1]  3. 不需要额外的限制。 |
| Decoder |  |  |

7. 仍然存在的问题：1）Unknown number of Speakers；2）Multiple Microphones；3）visual information

|  |  |  |
| --- | --- | --- |
|  |  |  |

0.2 重叠音检测

GitHub源码学习链接：https://github.com/pyannote/pyannote-audio

模型复现：

|  |  |
| --- | --- |
| 工作目录 | CPU2, /home/yinzhaohui/ |
| 环境变量 | /file2/Work/yinzhaohui/.bashrc |
| 数据源 |  |
| 方法原理 |  |
| 实验结果 |  |
| 困难 | 1. 如何在本地pycharm连接远程服务器使用GPU |

1. 11月份尚未弄懂的技术手段/专业名词

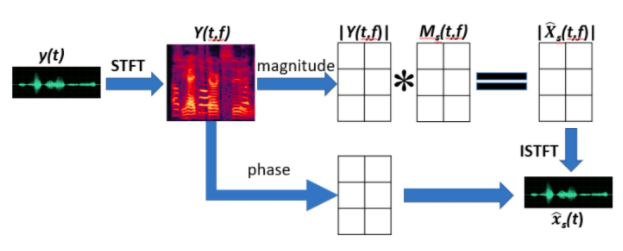


Figure 1：语音增强和分离流程图

|  |  |  |
| --- | --- | --- |
| 所属领域 | 缩写 | 描述 |
| 语音增强 | STFT | 短时傅立叶变换（Short-Time Fourier Transformation,STFT），时间域→频率域， |
| EM | 数据挖掘算法，https://zhuanlan.zhihu.com/p/40991784 |
| GSS | CHiME5中提出的一种语音增强系统。可以很好的用来生成说话人的Time-frequency masks。 |
| Beamformer | 波束成形 |
| PLDA | PLDA 是一种信号补偿算法，用来处理说话人和信道的变形, ivector从音频中提取了说话人和信道的特征, 当用PLDA时,进行了两次降维操作: 一次是提取ivector时,从2048的GMM降到了400-600的ivector维度,第二次就是用PLDA模型. |
| AHC | ？？？ |

2.文献粗读

1.1 kanda (2019), CHiME5挑战赛，实际场景技术综合

简述：

Q：CHiME挑战赛中如何讲重叠语音的起始时间和终止时间给提取出来？

A：发声开始和结束的时间都是有的，在这个特定场景下不需要使用重叠音检测。但是实际使用过程中，没有先验知识的情况下，重叠音检测是基础性工作。

Q1：去混响用什么方法？

Q2：训练数据少，数据增强用什么方法？

Q3：语音分离用什么方法？

A：利用SMM模型可以直接确定说话人的time-frequency masks。GSS可以稳定降低1.5%的词误率。

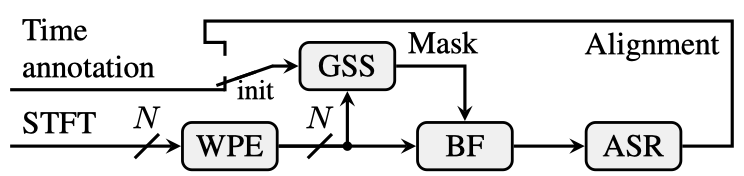
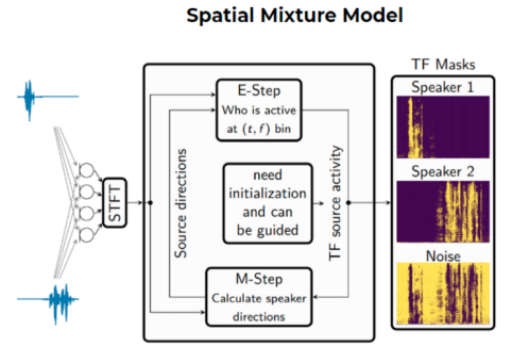
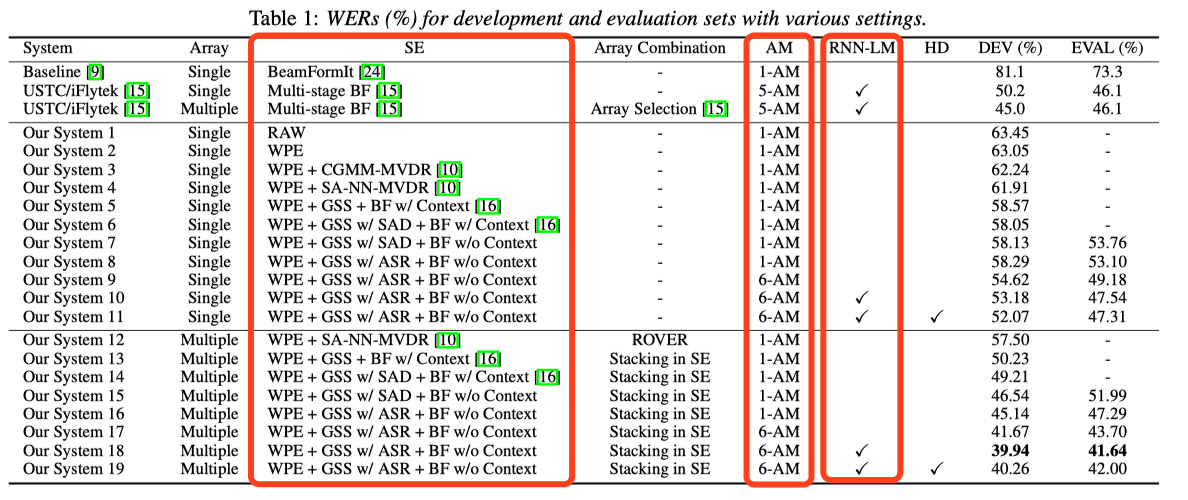


Figure 1：语音增强模块

Q：语音识别用什么方法？

A：声学模型+语言模型



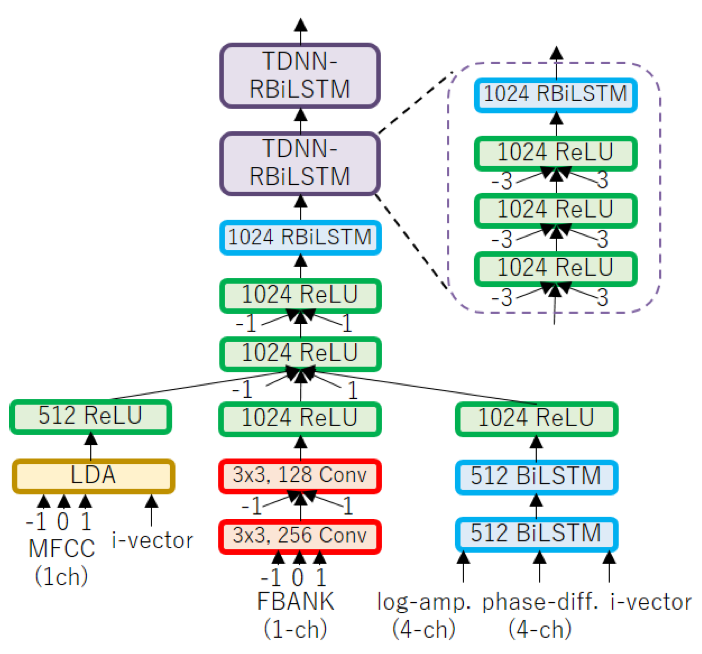


Figure 2：CNN-TDNN-RBiLSTM声学模型