基于深度学习的说话人日志技术

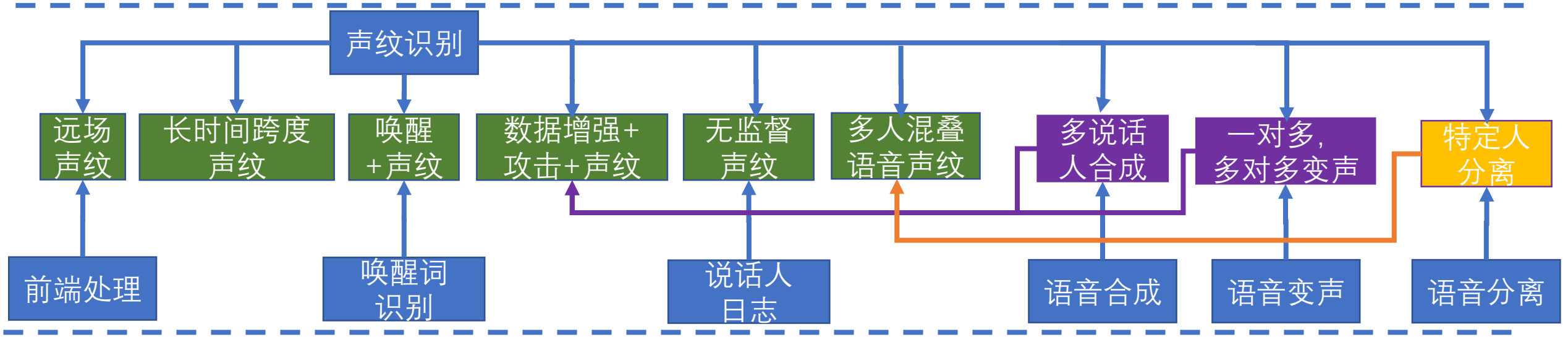
李明
Speech and Multimodal Intelligent Information Processing (SMIIP) Lab
昆山杜克大学大数据研究中心
ming.li369@dukekunshan.edu.cn

部分PPT素材来自于林庆健硕士论文答辩

昆山杜克大学

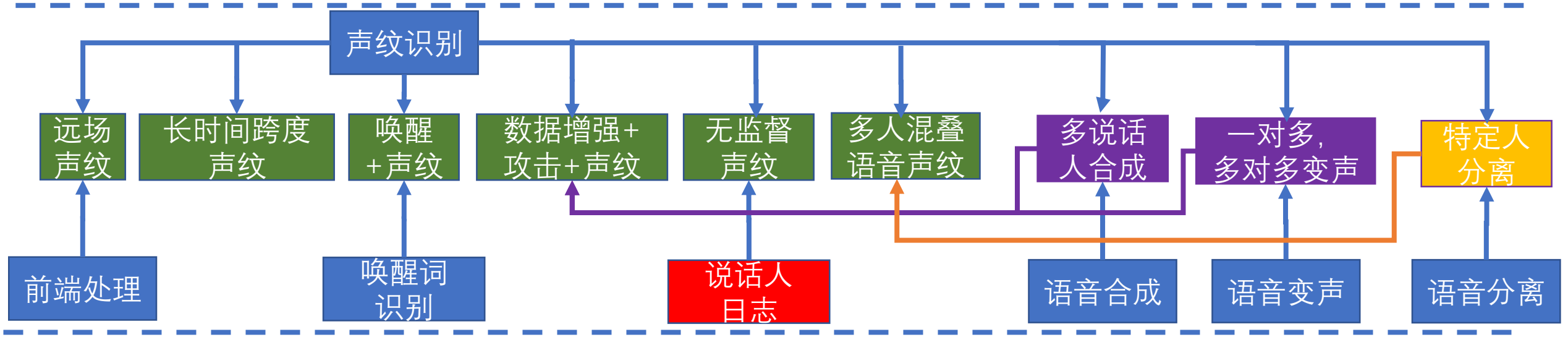
昆山杜克大学语音及多模态智能信息处理实验室在声纹识别领域的研究思路：

面向实际应用中的真实复杂场景，有针对性的研发一系列具备高可靠性和产品落地潜力的声纹识别前沿关键技术。具体地，这里所指的复杂场景包含使用场景的复杂声学环境；交互过程的复杂人声信号，安全方面的复杂攻击手段，识别对象的复杂声学特性等方面。同时我们也探索如何把高可靠性声纹编码与其他语音关联任务进一步有机结合，在多说话人语音合成，多目标说话人语音变声，特定人语音分离，特定人语音唤醒，复杂信道说话人日志等新兴交叉任务上，展开集成创新，提高系统的准确率和鲁棒性。我们也在积极参与开源数据，代码，举办国际评测，组织特殊议题和期刊专刊。
2020 Interspeech Far-field Speaker Verification challenge, 2021 ISCSLP Personal Voice Trigger Challenge, HIMIA, AISHELL3, etc.



昆山杜克大学语音及多模态智能信息处理实验室在声纹识别领域的研究思路：

面向实际应用中的真实复杂场景，有针对性的研发一系列具备高可靠性和产品落地潜力的声纹识别前沿关键技术。具体地，这里所指的复杂场景包含使用场景的复杂声学环境；交互过程的复杂人声信号，安全方面的复杂攻击手段，识别对象的复杂声学特性等方面。同时我们也探索如何把高可靠性声纹编码与其他语音关联任务进一步有机结合，在多说话人语音合成，多目标说话人语音变声，特定人语音分离，特定人语音唤醒，复杂信道说话人日志等新兴交叉任务上，展开集成创新，提高系统的准确率和鲁棒性。我们也在积极参与开源数据，代码，举办国际评测，组织特殊议题和期刊专刊。
2020 Interspeech Far-field Speaker Verification challenge, 2021 ISCSLP Personal Voice Trigger Challenge, HIMIA, AISHELL3, etc.

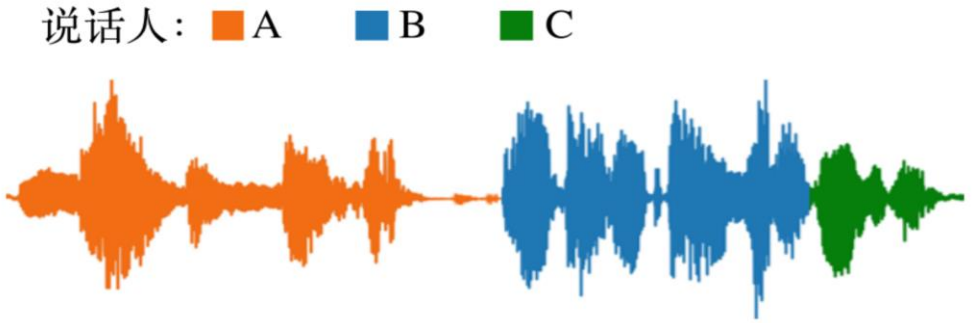


目录

目录

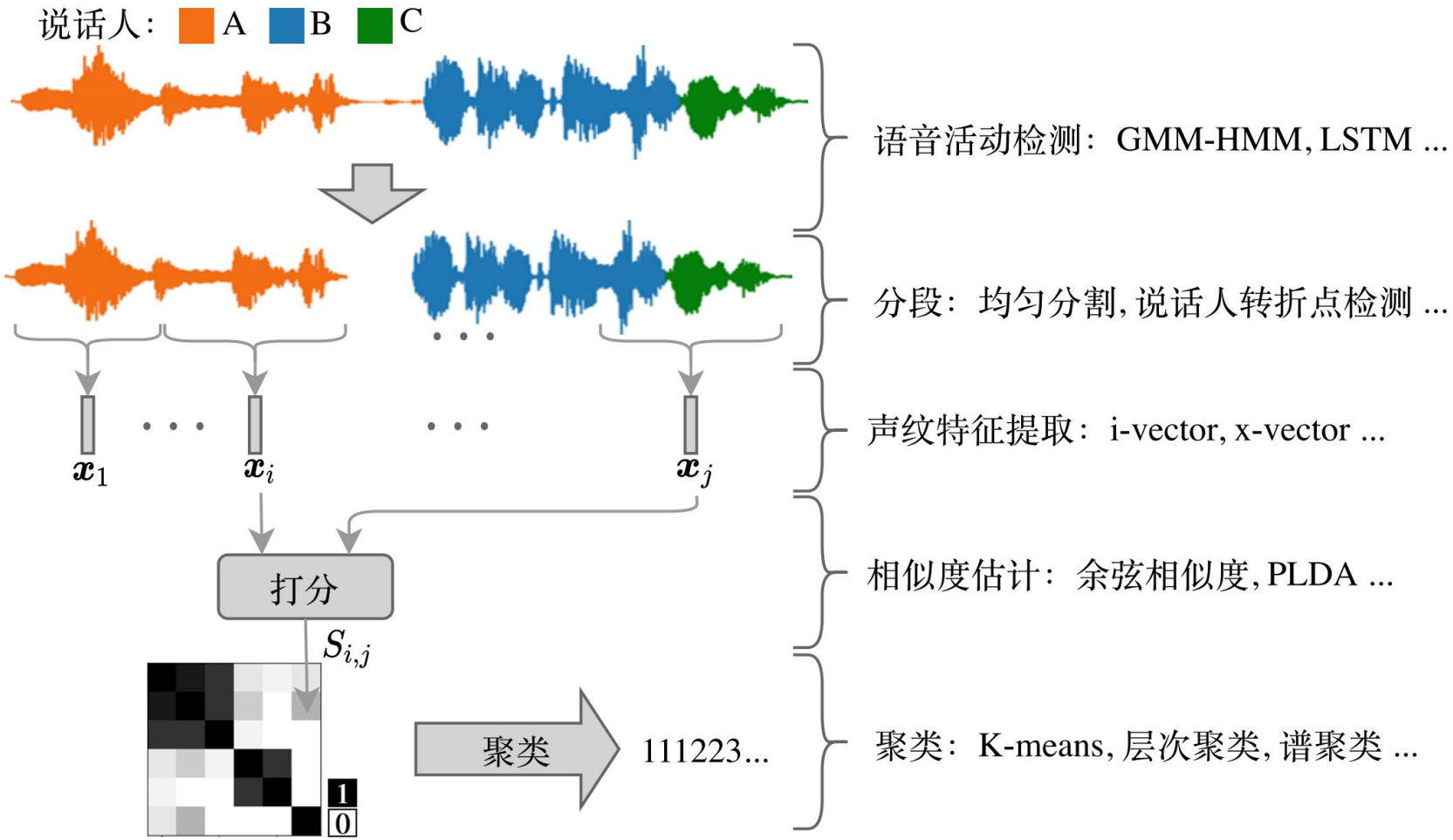
- **综述**
- 语音活动检测 (VAD)
- 声纹提取
- 相似度估计&聚类
- 端到端说话人日志

- **问题描述**
  - 识别出音频中每个说话人发声的起止时间戳，解决“谁在什么时候说话”(Who spoke when)的问题。



- **应用前景**
  - 音频归档，将会议内容按发言人身份进行整理。
  - 作为多人语音识别的前端，提高识别正确率（录音笔，会议场景等）。
  - 语音搜索引擎，对影视作品中的特定说话人按语音进行检索。
  - 等等

- **当前主流的模块化系统**


说话人：A B C
x_1
x_i
x_j
打分
$S_{i,j}$
聚类
111223...
1
0
语音活动检测：GMM-HMM, LSTM ...
分段：均匀分割, 说话人转折点检测 ...
声纹特征提取：i-vector, x-vector ...
相似度估计：余弦相似度, PLDA ...
聚类：K-means, 层次聚类, 谱聚类 ...

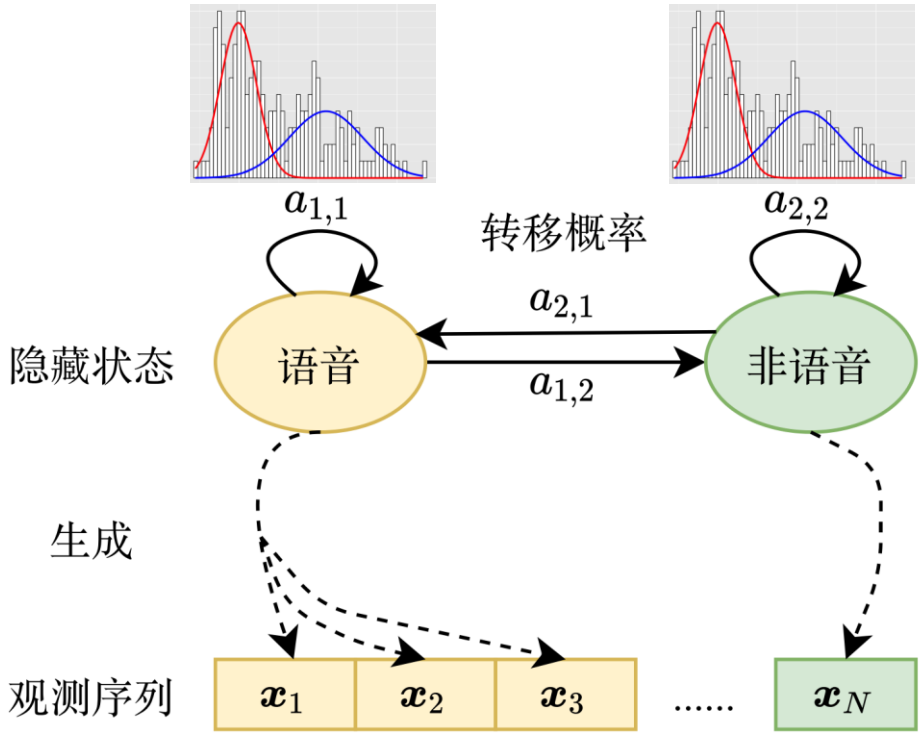
- **所汇报的实验结果主要使用的数据集**
  - 说话人日志数据

数据集	时长	说话人数	场景
AMI	100h	4~5	会议
ICSI	70h	3~10	会议
DIHARD2019	40h	1~10	11个复杂场景

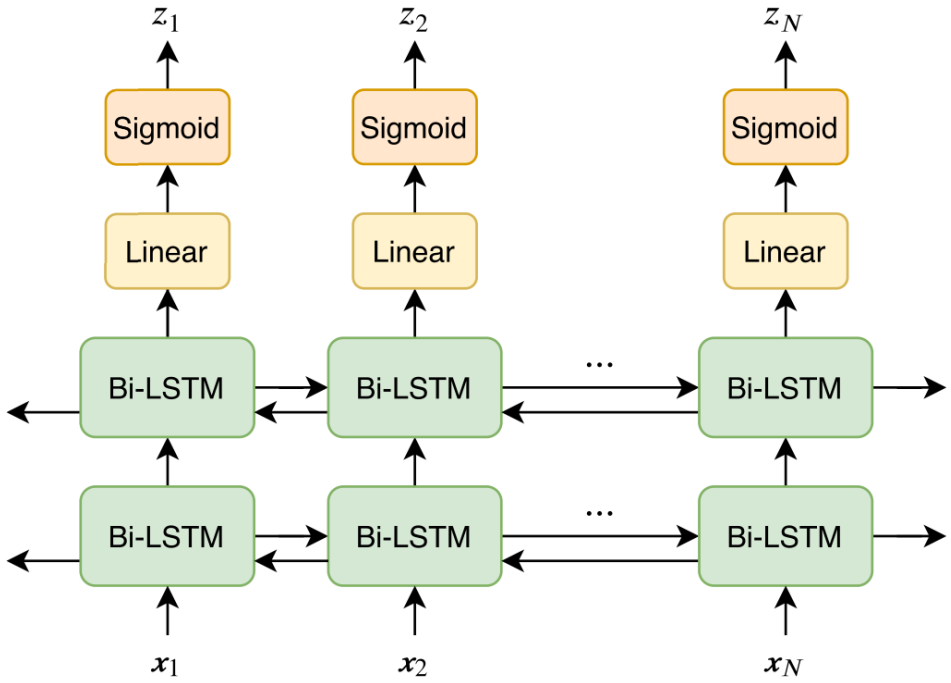
  - 声纹数据
    - Voxceleb1&2
  - 噪声数据
    - MUSAN

目录

- **高斯混合模型-隐马尔科夫模型 （GMM-HMM）**
  - 训练：
    - 为隐藏状态（语音/非语音）分别建立基于GMM的统计模型，统计两者之间的转移概率。
  - 测试：
    - 基于观测序列，采用维特比算法推断概率最大的隐藏状态序列。

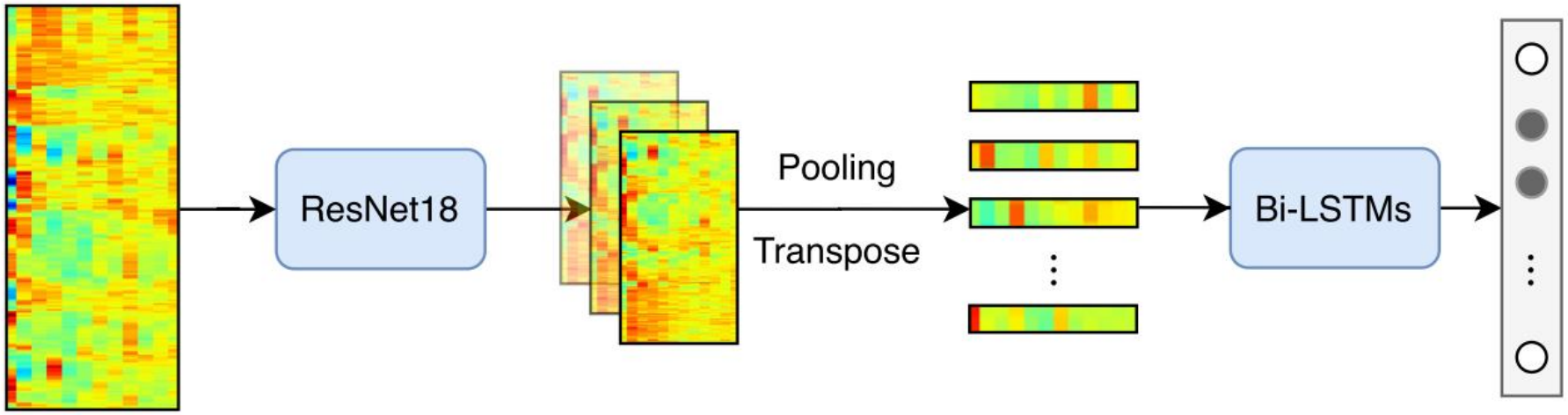


- **长短时记忆网络（LSTM）**
  - 输入为 fbank 特征，输出为介于 0 到 1 之间的浮点数，表示当前帧为语音的概率，其中 1 表示语音，0 表示非语音。
  - 优点：强大的时序分析能力。
  - 缺点：参数较多，不适合搭建较深的网络，对特征的信息抽取能力不强。



基于Bi-LSTM的语音活动检测模型

- **改进的 ResNet-LSTM 语音活动检测模型**
  - ResNet前端：对fbank谱进行信息抽取和池化，强化特征的信息表达能力。
  - LSTM后端：对 $F_1, F_2, F_3, \ldots$ 进行时序分析，得到语音活动检测结果。



[1] Qingjian Lin, Weicheng Cai, Lin Yang, Junjie Wang, Jun Zhang, Ming Li(*), “DIHARD II is Still Hard: Experimental Results and Discussions”, Odyssey 2020.
[2] Qingjian Lin, Tingle Li and Ming Li(*), “The DKU Speech Activity Detection and Speaker Identification Systems for Fearless Steps Challenge Phase-02”, Interspeech 2020

- **实验结果**
  - AMI数据集
    - 训练集，开发集，测试集均由AMI数据集切分得到，训练集与测试集领域匹配。

Dataset	Model	FAR(%)	MDR(%)	DCF(%)
AMI_dev	GMM-HMM	28.62	6.26	11.85
	LSTM	19.2	3.03	7.07
	ResNet-LSTM	**17.5**	**2.37**	**6.15**
AMI_test	GMM-HMM	36.29	6.13	13.67
	LSTM	20.8	2.25	6.89
	ResNet-LSTM	**18.1**	**2.24**	**6.21**

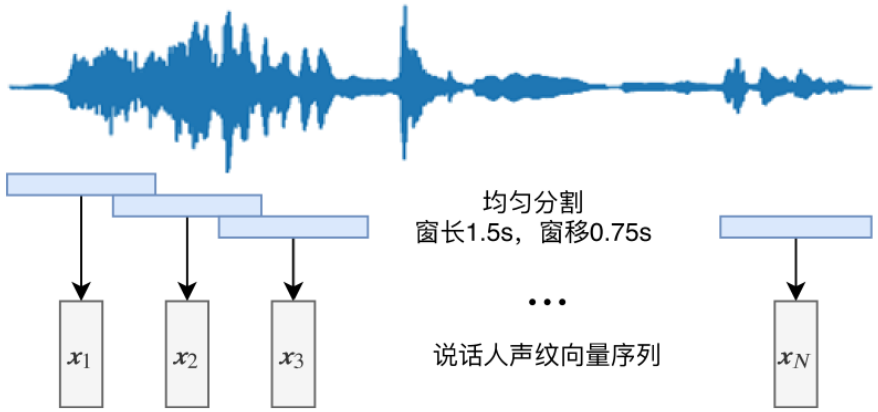
  - DIHARD2019数据集
    - 训练集：AMI + ICSI，开发集：DIHARD2019_dev，用于adapt，测试集：DIHARD2019_test，训练集与测试集领域不匹配。

Dataset	Model	FAR(%)	MDR(%)	DCF(%)
DIHARD2019_test	GMM-HMM	56.42	9.35	21.12
	LSTM	51.16	25.54	31.95
	ResNet-LSTM	52.92	14.18	23.87
DIAHRD2019_test (adapt)	GMM-HMM	41.42	8.45	12.46
	LSTM	22.61	5.32	9.65
	ResNet-LSTM	**16.25**	**5.24**	**7.99**

目录

- 综述
- 语音活动检测 (VAD)
- **声纹提取**
- 相似度估计&聚类
- 端到端说话人日志

- **语音分割**
  - 采用均匀分割，窗长1.5s，窗移0.75s。每段切片可近似看做只包含一个说话人。
- **说话人声纹特征提取**
  - 给定单一说话人的音频，提取能够表征说话人身份信息的固定维度向量。



- **3种不同的声纹特征**
  - **i-vector**
    - 将GMM生成的超向量投影到包含说话人信息和信道信息的总体差异空间中，得到表征目标说话人身份的声纹向量，即 i-vector [1]。
  - **x-vector**
    - 使用时延神经网络TDNN提取帧级别特征，计算时间维度的均值和标准差统计量，将帧特征转化为句子级别的特征，即x-vector [2] 。
  - **Deep ResNet**
    - 使用深度残差神经网络ResNet提取帧级别特征并计算统计量，即得到embedding。其具有更深的网络结构，对特征的抽取和表达能力也更强 [3,4]。

[1] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. IEEE TASLP, 19(4), 788-798.
[2] Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S. (2017) Deep Neural Network Embeddings for Text-Independent Speaker Verification. Proc. Interspeech 2017, 999-1003, DOI: 10.21437/Interspeech.2017-620.
[3] Weicheng Cai, Jinkun Chen and Ming Li (2018). Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System. Proc. Odyssey 2018, DOI: 10.21437/Odyssey.2018-11.
[4] Weicheng Cai, Jinkun Chen, Jun Zhang, Ming Li(*), Variable-length Data Loader and Utterance-level Aggregation for Speaker and Language Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020

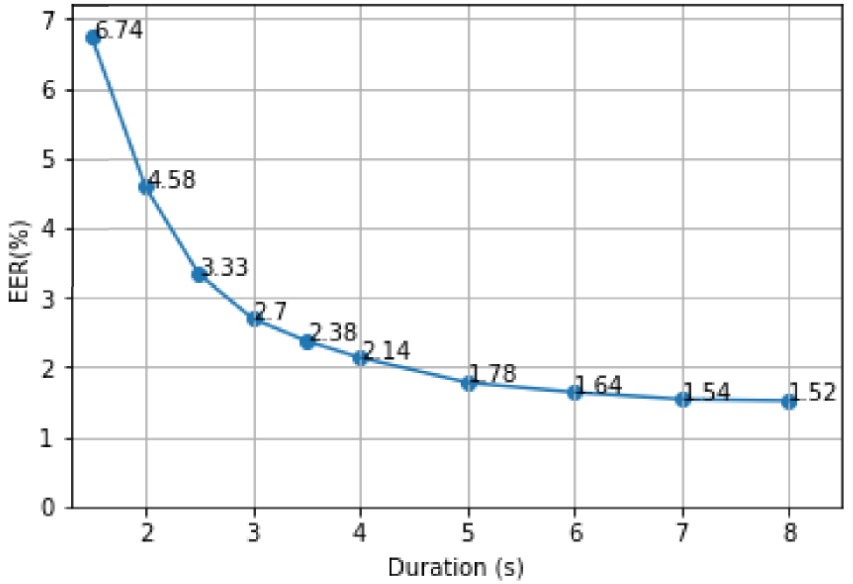
- **实验结果**

不同声纹特征之间的结果对比

表 3-3 i-vector、x-vector 和 Deep ResNet vector 等错误率与参数量对比

Dataset	Model	EER(%)	Parameters
Voxceleb1_test	i-vector	5.33	5.5M
	x-vector	3.22	8M
	Deep ResNet vector	**1.51**	**5.2M**

音频时长对声纹性能的影响



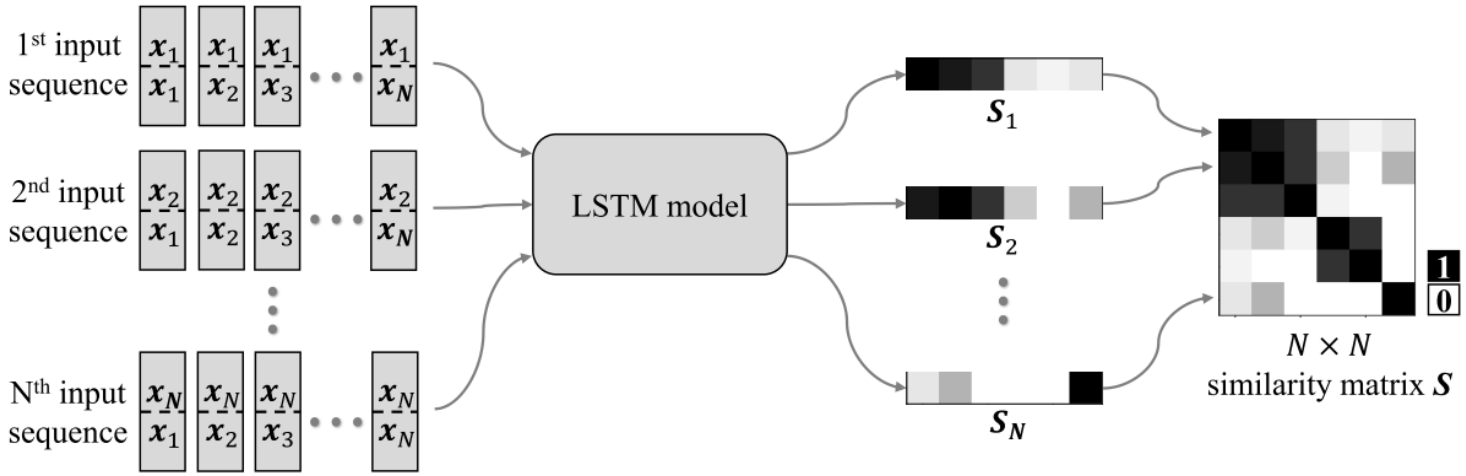
将音频截断至不同时长的情况下 Deep ResNet vector 的等错误率

高鲁棒性短时文本无关声纹仍然需要更多的研究

目录

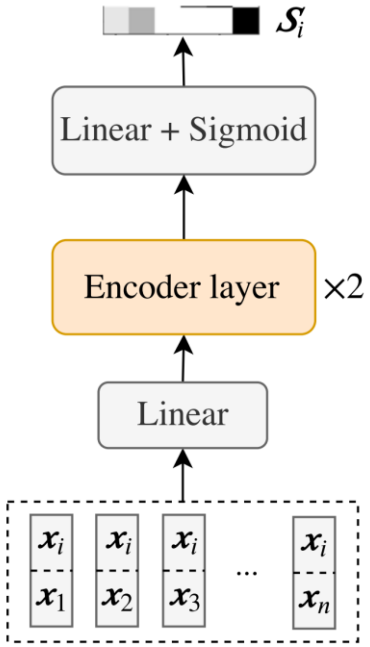
- **相似度估计**
  - **线性概率判别分析（PLDA）**
    - **i-vector，x-vector**
  - **余弦距离**
    - **Deep ResNet embedding**

- **相似度估计**
  - **基于LSTM的相似度估计**

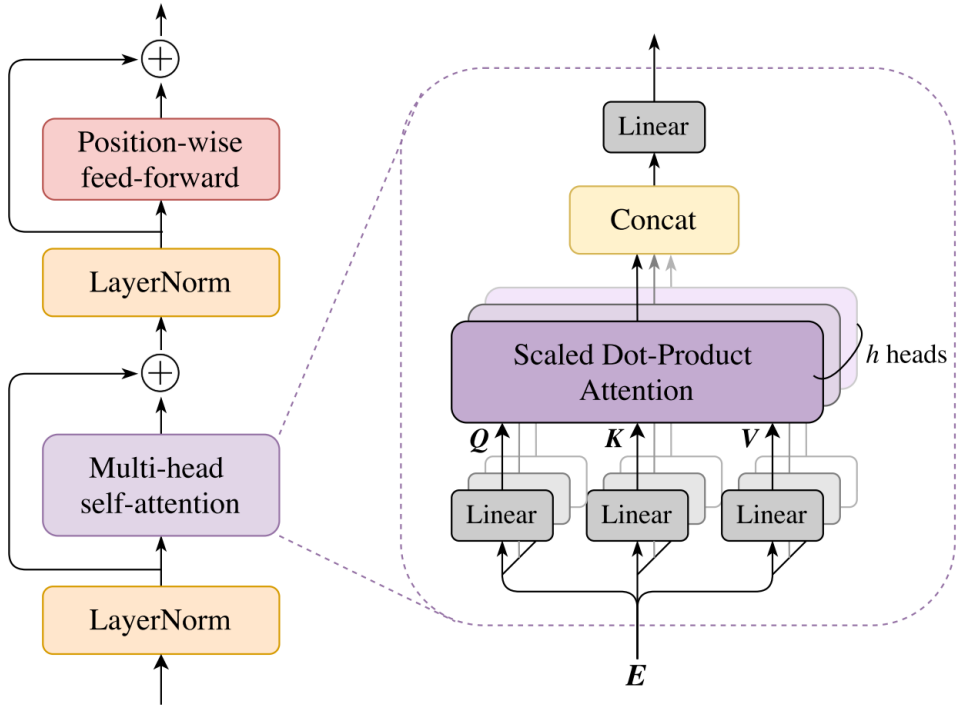


[1] Qingjian Lin, Ruiqing Yin, Ming Li(*), Hervé Bredin and Claude Barras, “LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarzation”, Interspeech 2019.

- **相似度估计**
  - **基于Attention的相似度估计 (vector to sequence)**



基于 Attention 的向量-序列打分模型

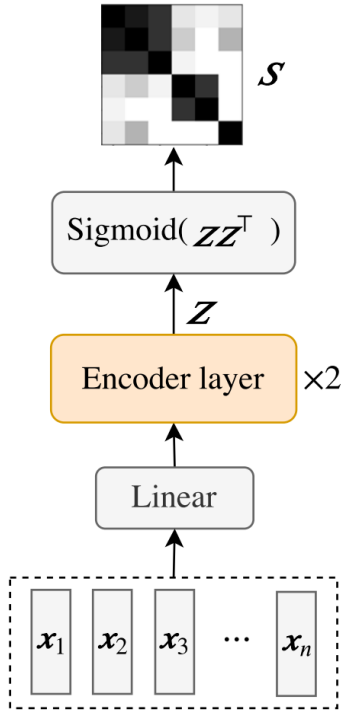


(a) Encoder layer (b) Multi-head self-attention

编码层结构

[1] Qingjian Lin, Yu Hou and Ming Li(*), “Self-Attentive Similarity Measurement Strategies in Speaker Diarization”, Interspeech 2020.

- **相似度估计**
  - **基于Attention的相似度估计 (sequence to sequence)**
    - 优点：一次构建完整的相似度矩阵，高效
    - 缺点：网络需要表达的信息量较多，因此性能可能不如向量-序列打分模型。



基于 Attention 的序列-序列打分模型

[1] Qingjian Lin, Yu Hou and Ming Li(*), “Self-Attentive Similarity Measurement Strategies in Speaker Diarization”, Interspeech 2020.

- **聚类算法**
  - **层次聚类** （Agglomerative Hierarchical Clustering，AHC）
    - 将每个样本初始化为单独的类，在迭代过程中不断合并子类。
  - **谱聚类**
    - 将相似度矩阵视为无向连通图。通过切断低权重的边，分割无向图为多个互不连通的子图。

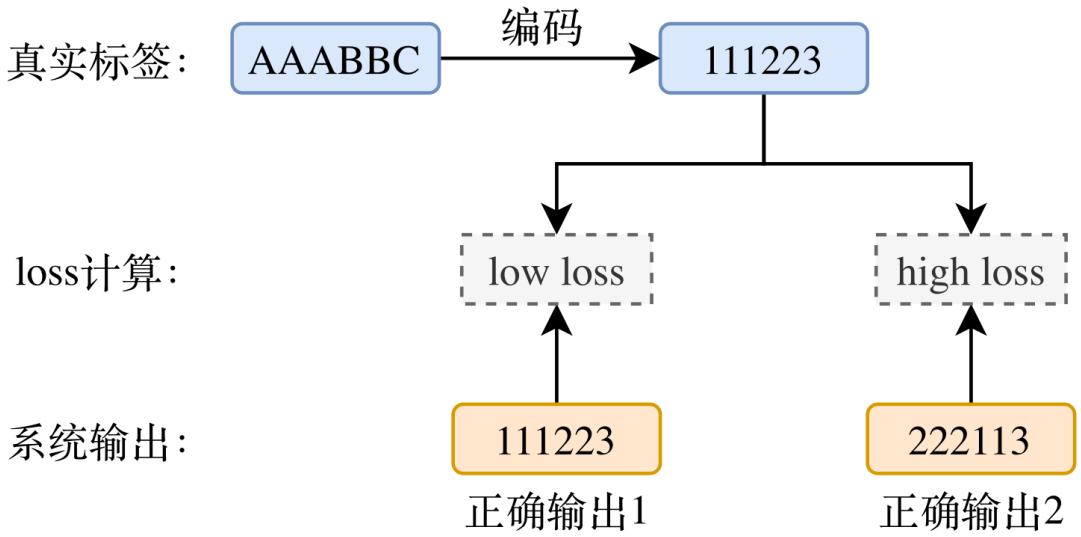
- **实验及结果**
  - 数据集
    - 训练集：AMI+ICSI
    - 测试集：DIHARD2019
  - 结果
    - Att-v2s > LSTM > Att-s2s > PLDA

Evaluation on DIHARD II corpus. Results are reported with and without domain adaptation by the Dev Set.

Model	+VB	Dev		Eval		Eval + adaptation		Time cost (Eval)
		DER(%)	JER(%)	DER(%)	JER(%)	DER(%)	JER(%)	
LSTM	×	19.65	49.60	20.57	50.25	19.72	46.49	67 min
	√	19.48	49.21	19.98	49.42	19.26	45.91	-
Att-v2s	×	**19.07**	**47.43**	**20.15**	**47.84**	**18.98**	43.20	148 min
	√	**18.76**	**46.77**	**19.46**	**47.01**	**18.44**	42.52	-
Att-s2s	×	19.39	48.42	21.46	48.71	21.45	**43.19**	24 s
	√	19.16	47.99	20.78	47.92	20.12	**41.73**	-
PLDA	×	23.48	57.17	-	-	23.73	56.84	51 s
DIHARD II winner system						18.42	44.58	
DIHARD II official baseline						25.99	59.51	

目录

- **模块化说话人日志系统的缺陷：**
  - 流程复杂
  - 不同子模块之间相互关联
  - 难以解决混叠语音的检测问题
- **端到端系统的挑战：**
  - 说话人歧义
  - 说话人数难以确定
  - 损失函数计算复杂度高



说话人歧义问题

- **损失函数**
  - **PIT Loss [1]**

$$J^{\text{PIT}} = \frac{1}{TN} \min_{\phi \in \text{perm}(N)} \text{BCE}(\boldsymbol{Z}, \boldsymbol{Y}^{\phi}).$$

  - 缺点：时间复杂度为$\mathcal{O}(N!)$。

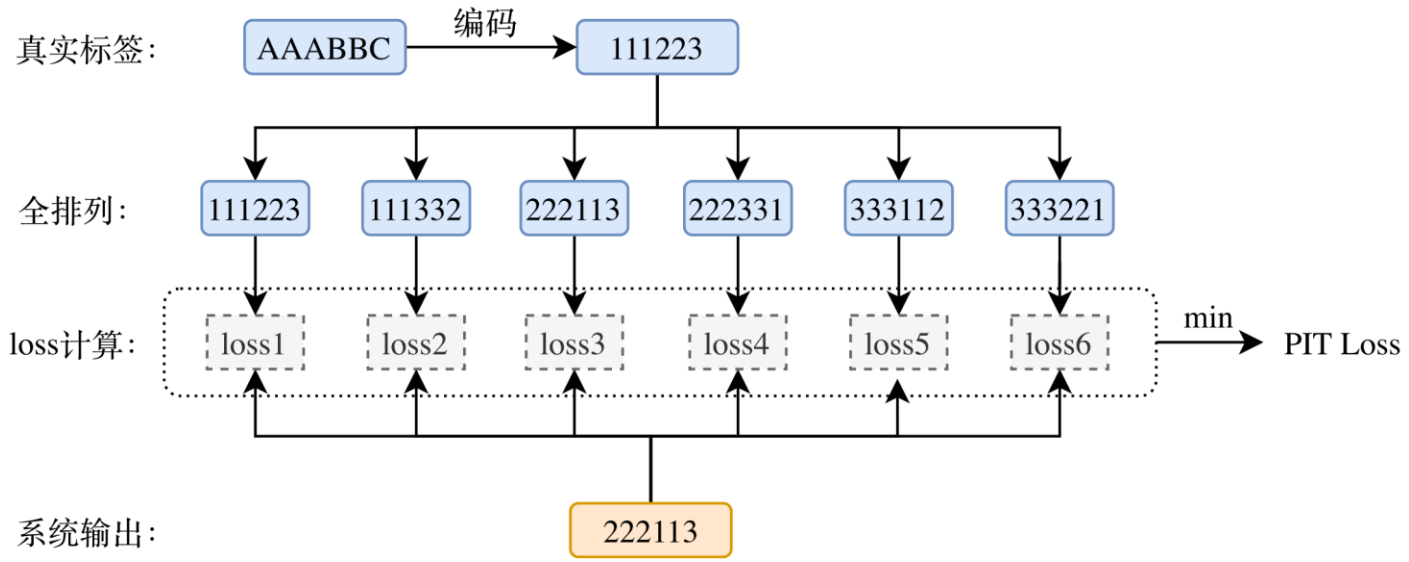
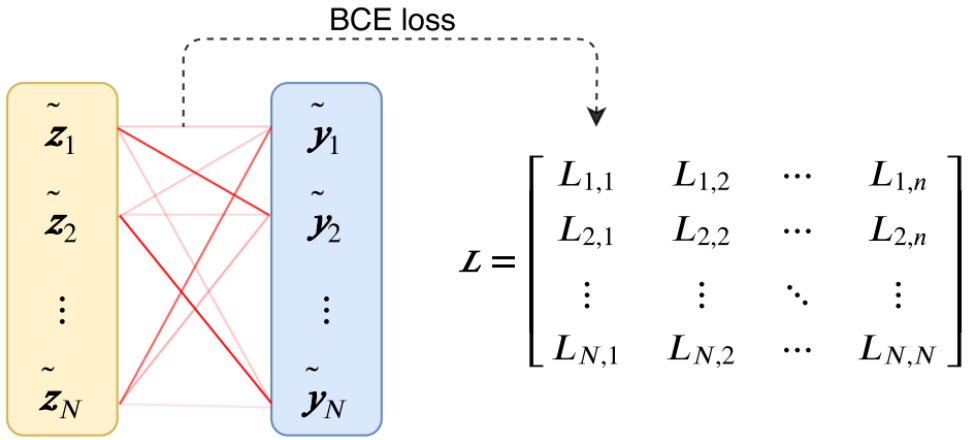


图 5-3 PIT 损失函数计算

[1] Yu, D., Kolbæk, M., Tan, Z. H., & Jensen, J. (2017, March). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In 2017 IEEE ICASSP (pp. 241-245). IEEE.

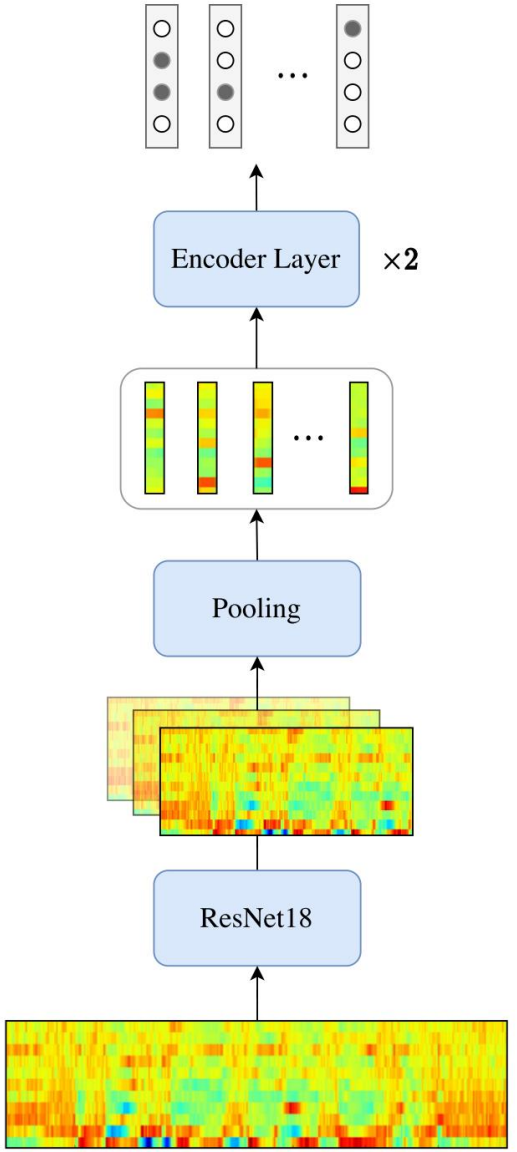
- **损失函数**
  - **OPTM Loss**
    - 将loss的计算过程看做任务分配问题，可使用匈牙利算法，时间复杂度为$\mathcal{O}(N^3)$。



匈牙利算法

[1] Qingjian Lin, Tingle Li, Lin Yang, Junjie Wang, Ming Li(*), “Optimal Mapping Loss: A Faster Loss for End-to-End Speaker Diarization”, Odyssey 2020.

- **网络结构**
  - ResNet前端
  - 池化层
  - 编码层后端



- **实验结果**
  - 训练集：AMI_train
  - 测试集：AMI_test

Train dataset	System	DER(%)			
		30s	1min	3min	5min
	Baseline	39.06 22.1+12.9+4.1	38.50 21.5+5.5+11.5	36.42 20.9+6.0+9.5	35.64 20.9+6.0+8.8
AMI_train	EESD	24.70 14.1+4.6+6.1	26.25 14.2+4.2+7.8	30.12 14.8+3.6+11.7	32.77 15.3+3.6+13.9
	EESD_PT	23.16 11.9+5.5+5.8	25.07 12.3+4.9+7.8	28.51 13.0+4.3+11.2	31.21 13.7+4.3+13.2
	EESD_FT	22.87 11.6+5.4+5.8	24.75 11.9+5.0+7.8	**28.14** 12.8+4.4+10.9	**30.29** 13.1+4.3+12.9
AMI_train +aug	EESD	22.97 12.2+4.9+5.9	25.32 12.7+4.6+8.0	29.47 13.5+4.1+11.9	31.37 14.0+3.6+13.8
	EESD_PT	23.03 11.0+5.9+6.2	24.99 10.9+5.7+8.4	29.46 11.6+5.1+12.7	31.99 11.7+5.2+15.1
	EESD_FT	**22.19** 10.4+5.7+6.2	**24.31** 10.3+5.5+8.5	28.23 11.3+5.0+12.0	30.60 12.0+4.6+14.0

- **实验结果**
  - 训练集：由voxceleb数据集模拟而成的训练数据，并加入MUSAN噪声以提高模型鲁棒性。
  - 开发集：在DIHARD2019_dev进行finetune。
  - 测试集：DIHARD2019_test。

DER (%) on DIHARD 2019 test

DIHARD II Baseline[1]	Best pre-is2019-deadline[2]	Best post-is2019-deadline[3]	SA-EEND + EDA[4]	本次汇报中的端到端方法
40.86	35.10	27.11	32.59	33.69

[1]G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, “Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD chal- lenge,” in INTERSPEECH, 2018, pp. 2808–2812.
[2] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, A. Avdeeva, A. Gorlanov, and A. Kozlov, “Speaker diarization with deep speaker embeddings for DIHARD Challenge II,” in IN- TERSPEECH, 2019, pp. 1003–1007.
[3] F. Landini, S. Wang, M. Diez, L. Burget, P. Matejka, K.Zmolkova, L. Mosner, A. Silnova, O. Plchot, O. Novotny, H. Zeinali, and S. Rohdin, “BUT system for the Second DIHARD Speech Diarization Challenge,” in ICASSP, 2020, pp. 6529–6533.
[4] Horiguchi S, Fujita Y, Watanabe S, Xue Y, Nagamatsu K. End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors, in arXiv preprint arXiv:2005.09921. 2020.

谢谢大家！

- ming.li369@dukekunshan.edu.cn

  https://scholars.duke.edu/person/MingLi

说话人日志方向近期国际评测成绩：
2019 Dihard2评测task1&2　第二名
2020 Voxceleb SRC评测task4 第三名

感谢说话人日志方向同学们的辛苦努力和卓越成果：
汪维清（在读）　林庆健（20年6月已毕业）

感谢此说话人日志项目科研合作方：
联想 AI LAB

Work reported represents collaborative efforts with many students, colleagues and collaborators!