Acoustic-aware Training for Multi-genre Speaker Recognition

Zhenyu Zhou

Instructors: Lantian Li

2022.11.04

目录

1 简介

Introduction

信号特点：语音信号具有“形简意丰”的特点，即在一段语音信号中可能融合了包括内容信息，说话人信息，环境信息在内的各种信息。

研究难点：对于说话人识别任务来说，场景信息是极大的干扰因素。但是在实际应用中，尤其是多场景说话人识别下，由于描述说话人的表征向量中冗杂了场景信息，造成了性能指标极度下降。

研究现状：当前主流的多场景说话人识别方法是基于数据驱动的多场景训练，即在训练时就丢入一个说话人的多场景语音数据。但是这种纯粹靠输入数据的分布特征来提取特征，难免会掺杂部分的场景信息，从而难以”纯化”说话人表征。

改进方法：本项目研究基于音频事件表征的多场景训练方法。加入预训练的音频事件表征，并以此为先验知识辅助模型训练，将盲目型训练变成知识型训练，从而更好地表征多场景说话人模型，提高在多场景下的识别性能。

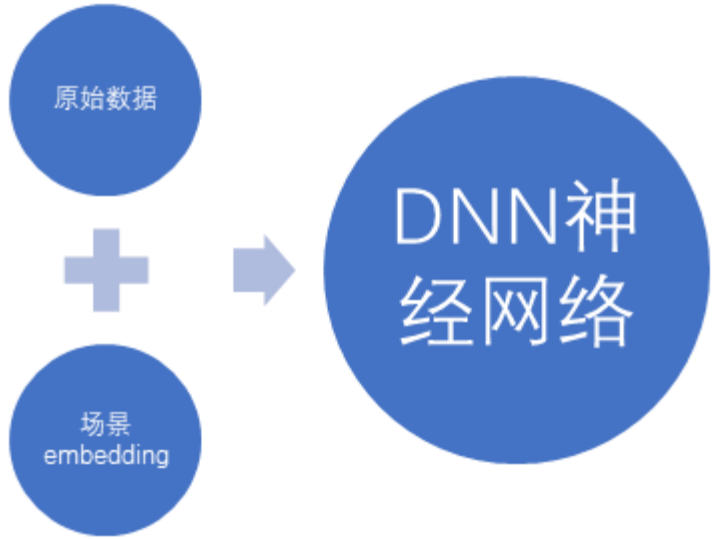
2 动机

Motivation

在一段语音数据中，声纹信息和场景信息是强耦合的，而这种耦合关系会极大的影响识别的准确率。

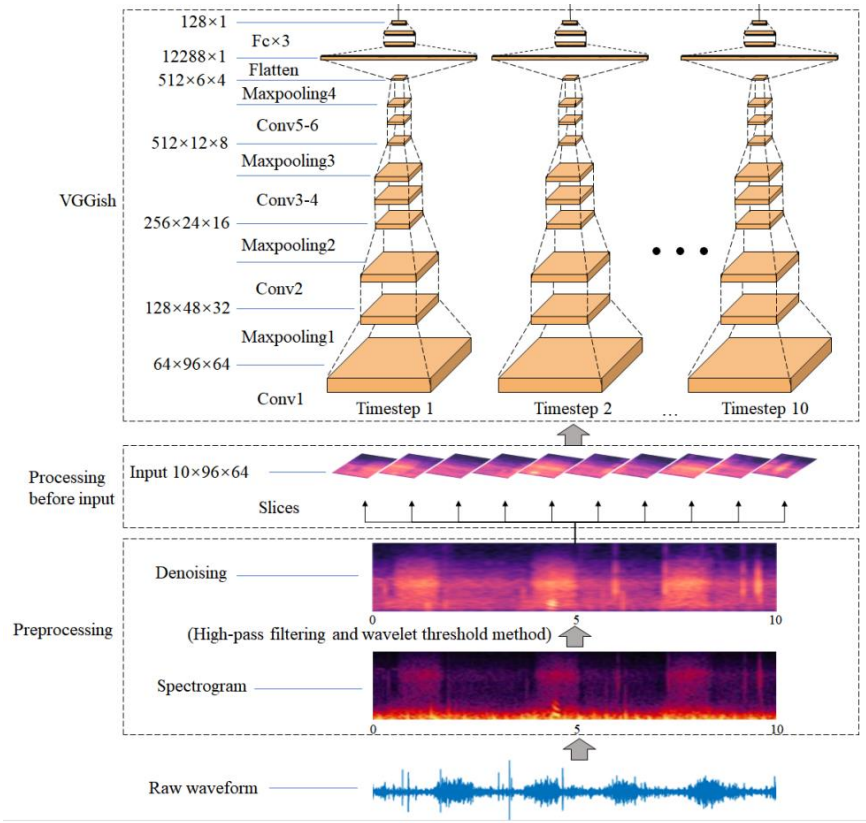
直觉上，场景信息和说话人信息可以看作是一对彼此独立的变量。因此，可以考虑采用条件学习（Conditional learning）的方式，把场景信息作为训练的辅助条件，来增强对深度说话人模型的学习。

具体地，对于一条语音信号，首先得到描述其场景信息的场景表征（Genre embedding），然后将该场景表征拼接到原始语音特征上，然后作为神经网络模型的输入。



3 场景表征的获取

Acoustic embedding




```
# Architectural constants.
NUM_FRAMES = 96  # Frames in input mel-spectrogram patch.
NUM_BANDS = 64  # Frequency bands in input mel-spectrogram patch.
EMBEDDING_SIZE = 128  # Size of embedding layer.

# Hyperparameters used in feature and example generation.
SAMPLE_RATE = 16000
STFT_WINDOW_LENGTH_SECONDS = 0.025    # 帧长
STFT_HOP_LENGTH_SECONDS = 0.010       #帧移
NUM_MEL_BINS = NUM_BANDS
MEL_MIN_HZ = 125
MEL_MAX_HZ = 7500
LOG_OFFSET = 0.01  # Offset used for stabilized log of input mel-spectrogram.
EXAMPLE_WINDOW_SECONDS = 0.96  # Each example contains 96 10ms frames
EXAMPLE_HOP_SECONDS = 0.96     # with zero overlap.
```

场景表征 (Acoustic embedding) 的获取采用的是在 Google AudioSet 数据集上预训练得到的 VGGish 模型。以 VGGish 模型为预训练模型，将音频原始输入特征转化成一个具有语义信息的低维音频事件表征向量。

左图是 VGGish 的网络架构。其中 VGGish 模型的输入是固定的，即 96 Frames * 64 Mel-spectrum。举例来说，对于一条 10s 的音频，首先按照 0.96s (96 frames * 0.01s hop length) 为最小单位，分成 10 个 segments，然后每个 segment 得到一个 128 维的 acoustic embedding，最后的输出便是 [10, 128] 的 tensor。最后将每一帧的 embedding 求均值得到句子级的 acoustic embedding。

4 可视化分析

Visualization analysis

Research background:

Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement	9.37	20.17	15.63	10.18	6.61	31.33	18.52	-	25.26	25.16	8.42	17.10
Drama	22.35	11.70	21.37	22.37	14.35	25.88	8.46	31.82	32.78	29.31	16.48	20.22
Entertainment	14.11	19.61	7.31	9.70	12.45	18.40	14.09	9.78	27.53	12.07	10.87	11.90
Interview	20.15	21.72	11.46	6.98	11.04	18.77	12.62	7.84	28.92	10.72	11.91	12.40
Live Broadcast	8.50	17.38	14.40	13.50	5.42	18.52	11.48	22.99	22.34	16.99	9.76	6.00
Movie	29.21	25.00	15.07	19.13	12.57	11.47	12.46	14.29	29.42	14.51	25.38	18.29
Play	9.09	16.84	14.91	19.29	15.09	24.67	11.56	5.88	28.25	11.82	18.18	12.65
Recitation	-	20.00	23.58	8.61	9.24	20.00	3.69	16.55	34.72	4.79	33.33	9.71
Singing	24.91	27.56	31.49	28.87	21.11	29.00	18.83	36.51	20.86	21.86	24.07	18.76
Speech	29.03	25.27	12.37	10.41	12.66	20.83	10.66	2.91	24.95	3.21	36.54	5.18
Vlog	5.26	19.73	26.75	10.45	10.38	15.07	9.68	20.00	25.40	36.63	5.31	5.32

EER (%) 37.0 20.0 3.0

Fig. 8. Cross-genre tests with the x-vector system. The lightness of the color corresponds to the numerical value of the EER (%).

在多场景的说话人确认任务中，会存在注册-测试场景失配问题，即当注册场景和测试场景不同时，识别错误率会很高。不同场景对于说话人声纹信息的干扰程度不尽相同。

为了消除这种识别误差，滤除场景特征对于声纹特征的影响，我们提出了加入场景表征的 acoustic-aware training 方法。通过对不同场景表征的降维可视化实验，可以定性地分析出不同场景之间的分布特性（近还是远离）。对于分布相近的场景，注册-测试失配问题并不大；而对于分布较远的场景，则可以加入场景表征，完成辅助训练和预测，提升识别性能。

从图中可以看出，singing 场景与 interview、entertainment、live_broadcast 三个场景会容易出现注册-测试失配问题。

参考文献：Li L, Liu R, Kang J, et al. CN-Celeb: multi-genre speaker recognition[J]. Speech Communication, 2022, 137: 77-91.

4 可视化分析

Visualization analysis

Technical comparison

• PCA

PCA降维方式用于提取数据的主要特征分量，主要思想是**将n维特征映射到互相正交的k维特征上，这些k维特征是在原有n维特征的基础上重新构造出来的**。PCA降维可以滤除掉部分噪声和冗余信息的干扰，但是PCA的可视化效果并不理想，聚类效果不明显。

• T-SNE

T-SNE降维能够将高维数据降低到2维或3维，在进行数据可视化时，能得到很好的结果。**主要思想是计算高维状态下，不同样本点之间的距离关系；同时在映射到低维时，保证相同样本点之间的距离要和高维下的距离关系尽可能相同**。这种相似性的描述，可以使用KL散度来定义loss function。

• UMAP

UMAP相比较于T-SNE可以在强调局部特征的前提下，兼顾到部分的整体信息。主要思想是，**使用KNN找到样本点的临近节点，并构建一个图。**

4 可视化分析

Visualization analysis

Experiments

通过 t-SNE，UMAP 等流形学习方法，对 Acoustic embedding 进行降维可视化，分析 speaker factor 和 genre factor 之间的分布特性。

Single-Speaker Multi-Genre (SS-MG)：验证表格中非对角线上的性能表现

青色的点代表 singing，红色的点代表 interview，蓝色的点代表 entertainment，粉色的点代表 live_broadcast。

Multi-Genre

- id00100

图1.1 PCA降维

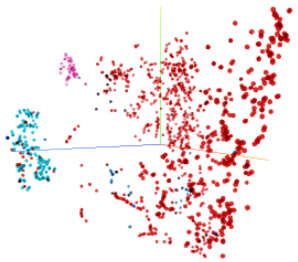


图1.2 T-SNE降维

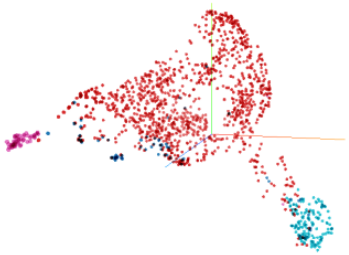


图1.3 UMAP降维

- id00392

图2.1 PCA降维

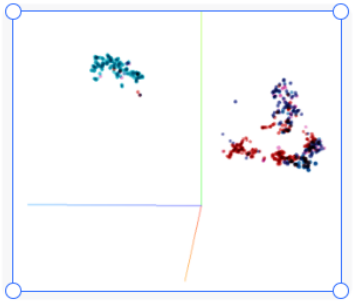


图2.2 T-SNE降维

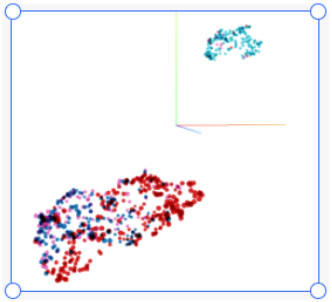


图2.3 UMAP降维

4 可视化分析

Visualization analysis

Multi-Genre

- id00500

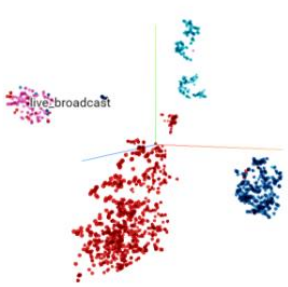


图3.1 PCA降维

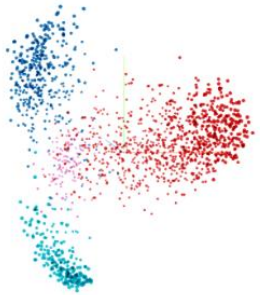


图3.2 T-SNE降维

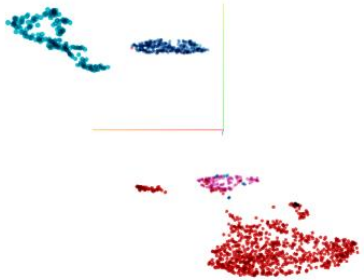


图3.3 UMAP降维

- id00523

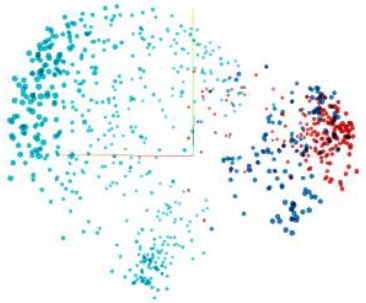


图4.1 PCA降维

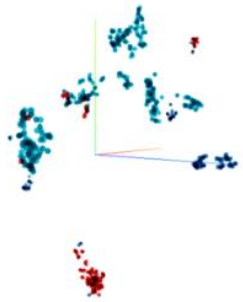


图4.2 T-SNE降维

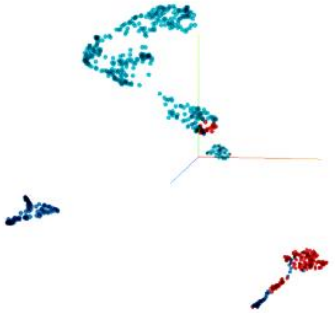


图4.3 UMAP降维

- id00702

图5.1 PCA降维

图5.2 T-SNE降维

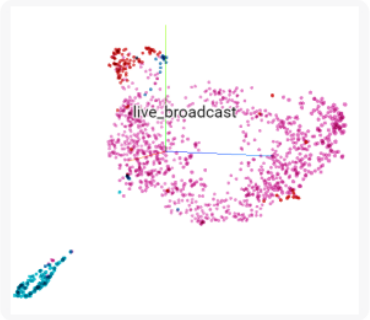


图5.3 UMAP降维

- id00976

图6.1 PCA降维

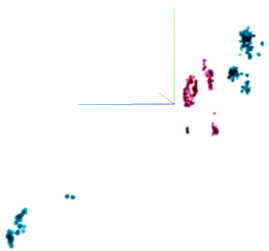


图6.2 T-SNE降维

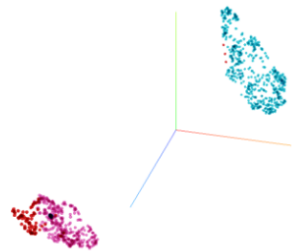


图6.3 UMAP降维

4 可视化分析

Visualization analysis

Conclusion

我们选取了六个不同的说话人，对于他们所共有的四类场景 singing、interview、entertainment 以及live_broadcast 进行 PCA、T-SNE、UMAP 三种方式进行降维可视化，可以得到一个普遍的结论，即singing 场景的 acoustic distribution（声学分布） 和其他三类场景的 acoustic distribution（声学分布）的距离较远，内聚性较差；其中 singing 和 entertainment 场景的 acoustic distribution 相差最大。其他三类场景的分布距离较近，具有一定的相似性；其中 Interview 场景和 entertainment 场景的分布距离很近，内聚效果很好，有着很高的相似性。

对于 interview、entertainment、live_broadcast 三类场景下的说话人确认来说，由于三类场景的 acoustic distribution 具有一定的相似性，场景信息的干扰较弱，当发生注册-测试场景失配时，识别性能仍相对可靠。对于 singing 和其他三类场景的说话人确认来说，场景特征对于说话人声纹特征的干扰作用较为明显，当发生场景失配时，极易造成较高的错误率。值得一提的是，acoustic embedding 分布的特性和 multi-genre 测试的结果完全吻合。这侧面意味着，acoustic embedding 中蕴含着能够描述场景特性的信息，可以作为 conditional learning 的辅助信息因子，帮助 multi-genre training，提升模型在 multi-genre 上的性能表现。

4 可视化分析

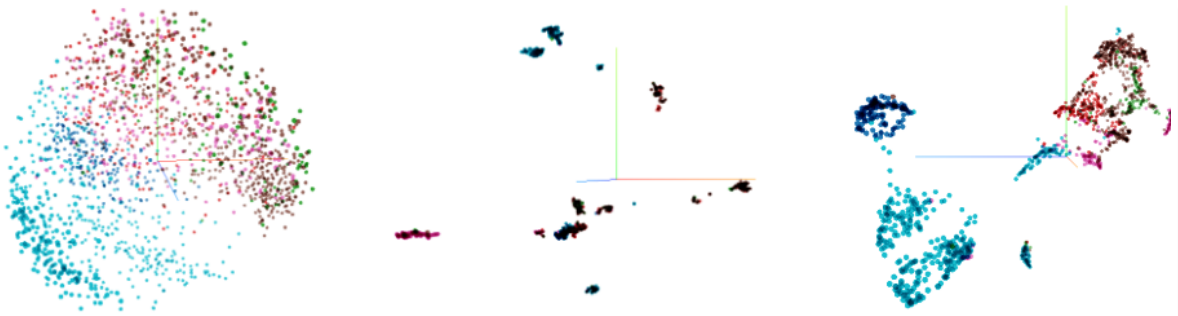
Visualization analysis

Multi-Speaker Single-Genre (MS-SG)：验证表格中对角线上的性能表现

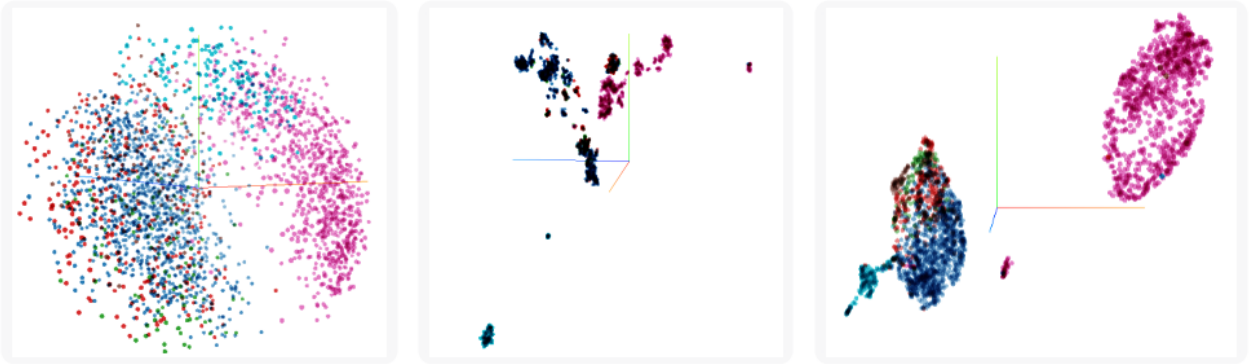
不同颜色代表不同的说话人，深蓝色代表 id00100，绿色代表 id00702，粉色代表 id00500，棕色代表id00976，红色代表 id00392，淡蓝色代表 id00523

Multi-speaker

- Singing



- Interview

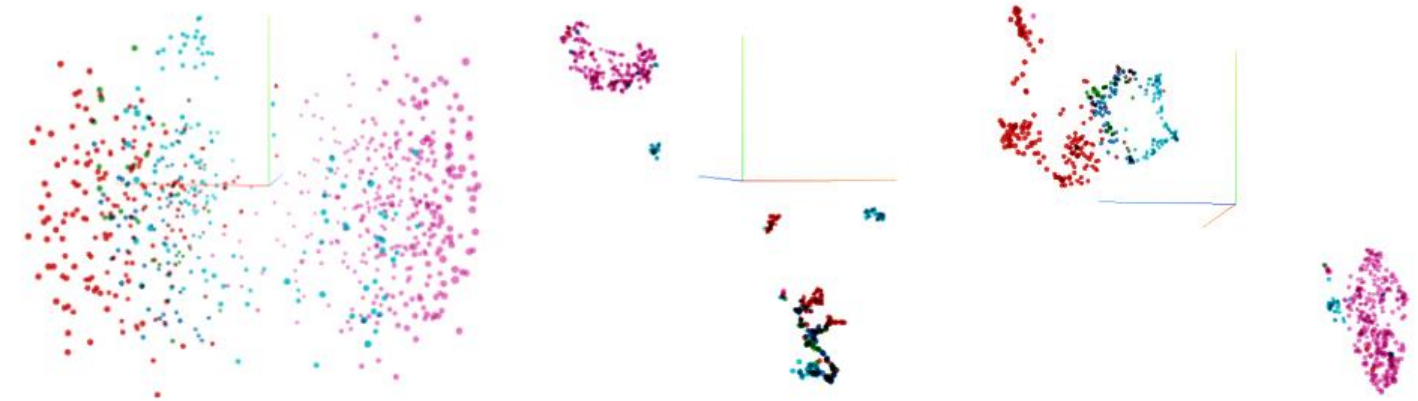


4 可视化分析

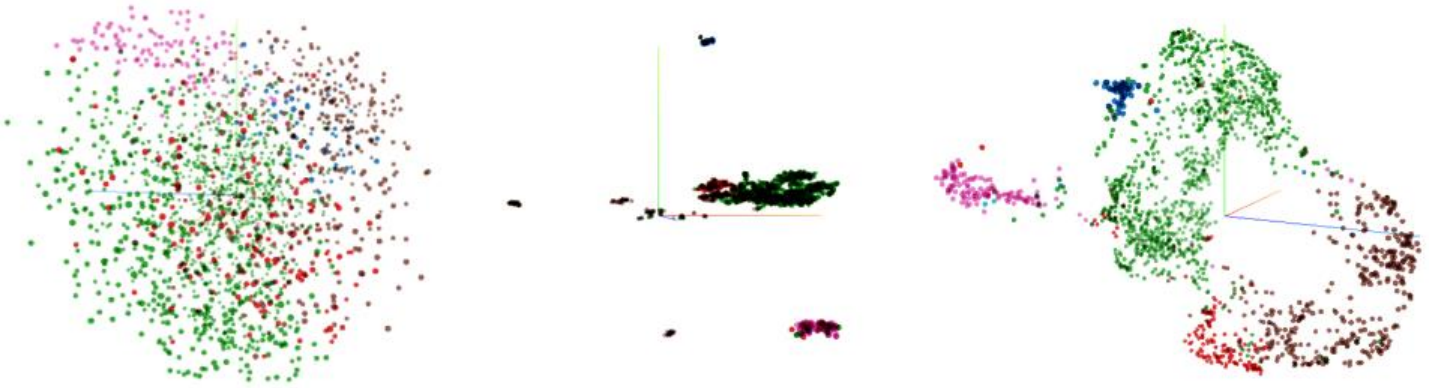
Visualization analysis

Multi-speaker :

- Entertainment



- live_broadcast



Conclusion

如上所示，live_broadcast 场景下的 acoustic distribution 具有不错区分性，entertainment 场景下的 acoustic distribution 的说话人区分性略逊于 live_broadcast，而 interview 场景下的 acoustic distribution 的说话人区分性最差的。

对照于 single-genre 的测试结果（对角线 EER 结果），live_broadcast 优于 entertainment 优于 interview。

这意味着 acoustic embedding 的说话人区分性与 speaker embedding 的说话人区分性在某种程度上具有一定的相关性。换言之，二者就说话人识别任务而言是互补的。

5 条件学习

Aware condition learning

Data

采用 CN-Celeb1 多复杂场景中文明星数据集。该数据集包含1000位说话人，覆盖11种真实场景。选择 CNC1.dev 作为训练集（800人），用于模型训练；CNC1.eval 作为测试集（200人），用于性能评测。

Vanilla-blind training：

基于 x-vector 框架，使用 ResNet34L、ECAPA-TDNN 为两种 backbones，采用 AM-Softmax 损失函数，采用 Attentive statistic pooling 池化策略；后端使用 Cosine 余弦距离进行系统打分；使用等错误率 EER 作为评价指标。

Vanilla-blind training result：

Backbone	Loss function	Pooling	Emb_dim	Backend	EER(%)	Ckpt
ResNet34L	AM-Softmax	ASP	256	Cosine	15.736	epoch=94_train_loss=0.11.ckpt
ECAPA-TDNN	AM-Softmax	ASP	256	Cosine	16.412	epoch=71_train_loss=0.15.ckpt

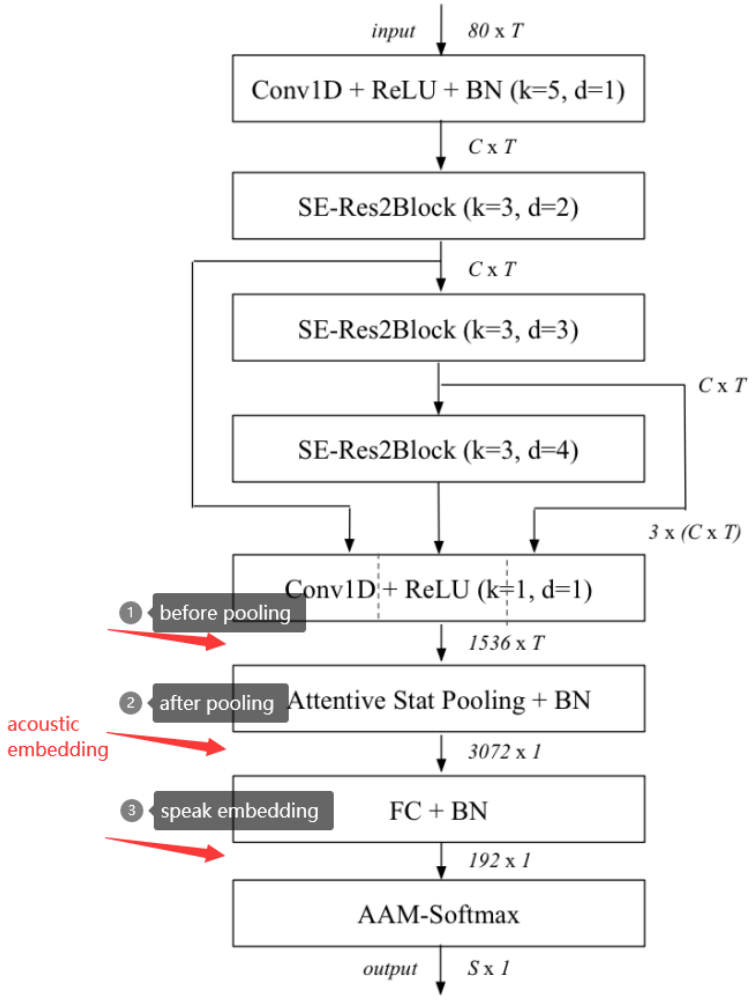
5 条件学习

Aware condition learning

Acoustic-aware training :
Improvement ideas:

Layer	Module	Output
Input	–	80×200×1
Conv2D	3×3×32, Stride 1	80×200×32
ResNetBlock1	$\begin{bmatrix} 3\times3\times32 \\ 3\times3\times32 \\ \text{SE Layer} \end{bmatrix}$ ×3, Stride 1	80×200×32
ResNetBlock2	$\begin{bmatrix} 3\times3\times64 \\ 3\times3\times64 \\ \text{SE Layer} \end{bmatrix}$ ×4, Stride 2	40×100×64
ResNetBlock3	$\begin{bmatrix} 3\times3\times128 \\ 3\times3\times128 \\ \text{SE Layer} \end{bmatrix}$ ×6, Stride 2	20×50×128
ResNetBlock4	$\begin{bmatrix} 3\times3\times256 \\ 3\times3\times256 \\ \text{SE Layer} \end{bmatrix}$ ×3, Stride 2	10×25×256
Pooling	TSP [37]	20×256
Flatten	–	5120
Dense	–	256
Dense	AM-Softmax [38]	5994

1 before pooling
2 after pooling
acoustic embedding
3 speaker embedding



6 对照试验

Experiments

Backbone	Position	Loss function	Pooling	Emb_dim	Backend	EER(%)	Ckpt
ResNet34L	Before Pooling	AM-Softmax	ASP	256	Cosine	15.477	epoch=78_train_loss=0.07.ckpt
	After Pooling					15.517	epoch=83_train_loss=0.13.ckpt
	At Spk Emb					14.892	epoch=98_train_loss=0.21.ckpt

Baseline: 15.736

Backbone	Position	Loss function	Pooling	Emb_dim	Backend	EER(%)	Ckpt
ECAPA-TDNN	Before Pooling	AM-Softmax	ASP	256	Cosine	15.984	epoch=89_train_loss=0.09.ckpt
	After Pooling					16.215	epoch=76_train_loss=0.10.ckpt
	At Spk Emb					16.176	epoch=98_train_loss=0.13.ckpt

Baseline: 16.412

对比实验结果， 可见加入 acoustic embedding 的先验信息，可以一定程度上提升在多场景下的说话人识别表现。

当使用 ECAPA-TDNN 作为主干网络时，当嵌入到 pooling layer 之前，训练 loss 能降低更多，同时 EER 指标能提升 0.428%；当嵌入到 pooling layer 之后和 speaker embedding 之上的性能提升并不明显。

当使用 ResNet34L 作为主干网络时，加在 pooling layer 之前和之后的效果提升并不明显 ，但是当嵌入到 Speaker embedding 之上时，性能的改善极为明显。

6 对照试验

Experiments

Backbone	Position	Loss function	Pooling	Emb_dim	Backend	EER(%)	Ckpt	Test embedding
ResNet34L	At Spk Emb	AM-Softmax	ASP	256	Cosine	14.892	epoch=98_train_loss=0.21.ckpt	Speaker embedding
ResNet34L	At Spk Emb	AM-Softmax	ASP	256	Cosine	14.959	epoch=98_train_loss=0.21.ckpt	Speaker embedding+acoustic embedding
Backbone	**Position**	**Loss function**	**Pooling**	**Emb_dim**	**Backend**	**EER(%)**	**Ckpt**	**Test embedding**
ECAPA-TDNN	At Spk Emb	AM-Softmax	ASP	256	Cosine	16.176	epoch=98_train_loss=0.13.ckpt	Speaker embedding
ECAPA-TDNN	At Spk Emb	AM-Softmax	ASP	256	Cosine	16.328	epoch=98_train_loss=0.13.ckpt	Speaker embedding+acoustic embedding

在测试过程中，当输入的 Test embedding 为 Speaker embedding，识别表现结果要优于输入的 Test embedding 为 Speaker embedding + acoustic embedding。

分析：对于目标说话人测试列表，注册场景和测试场景相同的情况下，加入 acoustic embedding 会正向的提升识别性能；在注册场景和测试场景不同的情况下，加入 acoustic embedding 会反向降低识别性能。对于非目标说话人测试列表，效果是截然相反的。

从实验结果看出，输入的 Test embedding 为 Speaker embedding 时，识别效果更好。这是可以理解的，因为在多场景说话人识别任务中，注册和测试场景不同的情况比两者场景相同的情况要更为普遍。

7 总结

Conclusion

• 语音数据中的声纹信息和场景信息是强耦合的，这种耦合关系会极大的影响识别的准确率。

• 部分场景的声学分布具有很高的相似性，对于这部分场景信息来说，不容易发生注册-测试失配问题。某些场景和其它场景的声学分布差别较大，对于这部分场景，场景特征对于声纹特征的干扰更加明显。

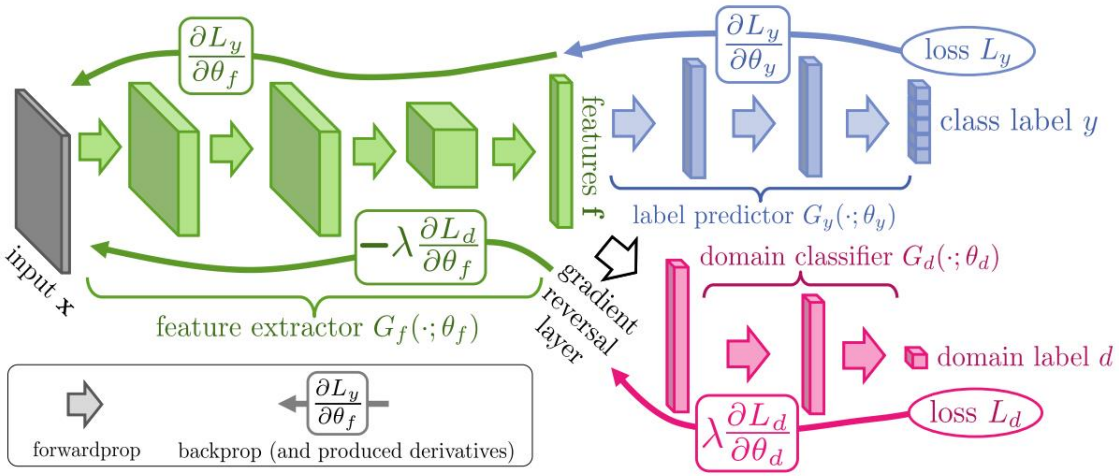
• 不同场景下的说话人分布差异很明显。部分场景下，不同说话人之间的方差较小，在这些场景下的说话人识别任务性能较差。

• 加入 acoustic embedding 的 condition-aware learning 可以在一定程度上滤除说话人特征中的场景信息，提高多场景说话人识别的性能。

8 未来工作

Future Work

- Multi-task training and Genre-adversarial training



- Working on large data sets

将现在的工作迁移到大数据集上进行验证

Thank you for your listening !

Zhenyu Zhou

Instructors: Lantian Li

2022.11.04