用于多通道连续语音分离的多分辨率基于位置的训练

Hassan Taherian[1], 和*DeLiang Wang*[1,2]

[1]俄亥俄州立大学计算机科学与工程系，美国[2]俄亥俄州立大学认知与脑科学中心，美国

taherian.1@osu.edu, dwang@cse.ohio-state.edu

arXiv:2301.06458v1 [eess.AS] 16 Jan 2023

摘要

当多说话人语音重叠时，自动语音识别（ASR）系统的性能会大幅下降。在会议环境中，通常会进行语音分离，以提高自动语音识别系统的鲁棒性。近期，基于位置的训练（LBT）被提出，作为一种用于多通道、与说话人无关的说话人分离的新训练标准。在假设阵列几何结构固定的情况下，对于完全重叠语句和匹配的混响条件，LBT 的表现优于广泛使用的置换不变训练。本文将 LBT 扩展到对话式多通道说话人分离。我们提出多分辨率基于位置的训练，用于估计从低到高时间和频率分辨率下的复数谱图。采用多分辨率基于位置的训练时，卷积核会根据物理空间中的说话人位置进行一致分配。评估结果表明，在录制的 LibriCSS 语料库上，多分辨率基于位置的训练始终优于其他具有竞争力的方法。

索引词— 连续语音分离、复数谱映射、基于位置的训练。

1. 引言

基于深度神经网络（DNN）的多说话人分离方法近年来取得了显著进展 [1]。现有分离方法主要针对短片段中的完全重叠语句。然而，在实际应用中，分离模型需要处理带有偶发语音重叠的长音频录音。近期，研究者提出了连续语音分离（CSS），用于处理对话或会议环境中的重叠语音 [2]。CSS 通常将长音频流划分为短片段。每个片段都由分离模型独立处理，并采用置换不变训练（PIT）评估标准。随后，将分离信号拼接起来，生成各个说话人流。

已有许多研究致力于改进 CSS 框架的不同方面 [3, 4, 5, 6, 7, 8, 9, 10]。我们引入了一种基于片段重叠率的调制因子，以动态调整分离损失 [3]。在 [4], 中，研究者使用循环选择性注意力网络，一次分离一位说话人的语音。 [5] 和 [6] 中的工作提出了新的训练评估标准，将 PIT 扩展为能够捕获长语音上下文的方法。Li等人 [7] 提出了一种双路径分离模型，利用记忆嵌入池中的片段间信息。Wang等人采用多阶段方法，将多通道分离模型与基于掩码的波束形成及后滤波网络相结合，从而取得了出色的分离性能 [8]。其他研究则将无监督多通道源分离用于 CSS。Saijo等人提出了一种空间损失，利用估计的到达方向对解混矩阵施加空间约束 [9]。Boeddeker等人 [10] 提出了一种基于复数角度中心高斯混合模型的波束形成器初始化方案。

在此前的研究中，我们提出了一种新的训练准则，称为基于位置的训练（LBT），根据说话人在物理空间中的位置分配 DNN 输出 [11]。我们证明，在模拟和匹配的混响条件下，对于完全重叠语句，LBT 的表现优于 PIT。本文将 LBT 扩展到会话场景中的多通道说话人分离以及不匹配的混响条件。我们训练了一种多通道输入、单通道输出（MISO）系统，直接从多通道混合信号中估计说话人的实部和虚部频谱图。我们还为卷积神经网络提出了一种多分辨率损失，通过从低到高的时间和频率分辨率估计复数谱图。借助多分辨率损失，卷积核能够根据说话人的位置始终与其对应。我们证明，尽管训练时仅使用模拟的房间脉冲响应（RIR），多分辨率基于位置的训练仍能很好地泛化到真实的会话录音。在 LibriCSS 语料库 [2] 上的实验表明，与其他有竞争力的方法相比，所提出的分离模型具有更优的分离性能。

本研究部分获得了美国国家科学基金会的两项资助（ECCS-1808932 和 ECCS-2125074）、俄亥俄超级计算机中心以及匹兹堡超级计算机中心（NSFACI-1928147）的支持。

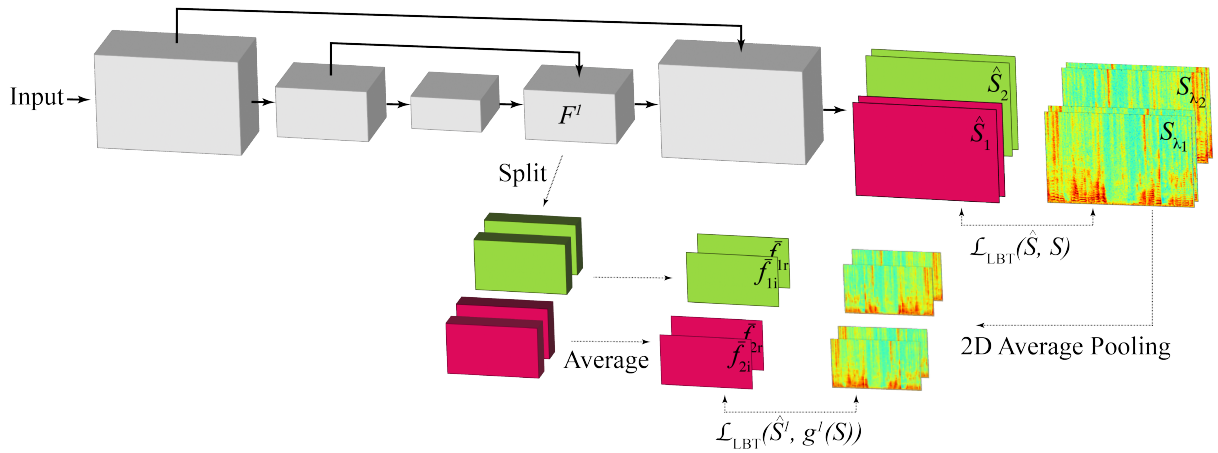

图 1：MC-CSM 网络的多分辨率基于位置的训练示意图，该网络包含 $K_e = 3$ 个编码器层和 $K_d = 2$ 个解码器层。每个密集块之后均使用一个步幅为 2×2 的深度卷积层，分别用于编码器和解码器中的下采样与上采样。每个立方体表示一个输出特征图。跳跃连接将编码器与相应的解码器相连。解码器中间层的特征图分为四组，并根据说话人的位置分配给相应的说话人。

2. 系统描述

2.1. 连续语音分离的基于位置的训练

连续语音分离的目标是从包含偶发语音重叠的多通道信号中生成多个互不重叠的语音流 [2]。只要同一时间段内同时活跃的说话人不超过 N，连续语音分离框架就可以处理任意数量的说话人。在本研究中，我们假设 $N = 2$ 短时间段内同时出现两名以上说话人的情况很少发生。此前大多数连续语音分离研究都采用逐语句 PIT 来训练分离模型 [2, 8, 12]:

$$\mathcal{L}_{\text{PIT}}(\hat{S}, S) = \min_{\phi\in\Phi} \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\phi(n)}), \tag{1}$$

其中 S 和 S 分别是短时傅里叶变换（STFT）域中的估计语音信号和干净语音信号。$\mathcal{L}$ 表示损失函数，符号 Φ 表示所有排列的集合，ϕ 表示其中一个排列。对于没有语音重叠的片段，基于置换不变训练（PIT）的分离模型会将输入信号送至其中一个输出，而另一个输出通道则产生零信号。

近期，有研究提出 LBT，以解决多通道说话人无关的说话人分离中的置换歧义问题 [11]。LBT 利用物理空间中多个说话人的不同空间位置，与 PIT 相比可获得更优的分离性能。此外，LBT 的训练复杂度随说话人数线性增长，在计算上远比复杂度呈阶乘增长的 PIT 高效。本研究将 LBT 评估标准扩展至连续语音分离（CSS）：

$$\mathcal{L}_{\text{LBT}}(\hat{S}, S) = \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\lambda_n}), \tag{2}$$

其中，$\lambda_n \in \{1, 2\}$ 是根据说话人相对于麦克风阵列的方位角或距离按升序排列的说话人索引。与 PIT 不同，式（2）依据说话人方位角或距离对 DNN 输出进行组织。对于非重叠片段，输入会映射到第一个输出，第二个输出则分配一个零信号。

2.2. 多通道复数谱映射

我们的分离模型基于多通道复数频谱映射（MC-CSM），其分离性能优于基于掩蔽的波束形成 [13, 14]。我们采用了 [15]提出的用于 MC-CSM 的 Dense-UNet 架构。Dense-UNet 的输入包括所有麦克风处混合信号短时傅里叶变换（STFT）的实部和虚部，以及第一个麦克风的频谱幅度。Dense-UNet 包含一个编码器和一个解码器，分别由 $K_e = 5$ 层和 $K_d = 4$ 层密集连接卷积块组成。每个密集块包含 5 个卷积层，每层有 76 个通道，卷积核大小为 3×3，步幅为 1×1。在最后一个密集块之后，我们使用一个具有 4 个通道的 1×1 卷积层，生成每个说话者实部和虚部的估计值。MC-CSM 包含约 690 万个参数。关于 Dense-UNet 架构的更多细节，请参阅 [15]。模型使用 [16]提出的损失函数进行训练:

$$\begin{aligned}\mathcal{L}(\hat{S}, S) = &\left\|\hat{S}^{(r)} - S^{(r)}\right\|_1 + \left\|\hat{S}^{(i)} - S^{(i)}\right\|_1 \\ &+ \left\|\left|\hat{S}\right| - |S|\right\|_1,\end{aligned} \tag{3}$$

其中，上标 r 和 i 表示实部和虚部，$|.|$ 用于计算幅度，$\|.\|_1$ 用于计算 ℓ_1 范数。

2.3. 多分辨率基于位置的训练

多分辨率基于位置的训练的核心思想是，通过逐步提高时间和频率分辨率，使用解码器中的专用卷积核来估计复数频谱图。图 1 展示了具有 $K_e = 3$ 个编码器层和 $K_d = 2$ 个解码器层的 Dense-UNet 的多分辨率基于位置的训练。对于解码器中的每个中间层 k，我们将输出特征图 F^k 的通道维度分成四组：

$$f^k_{1r}, f^k_{1i}, f^k_{2r}, f^k_{2i} = \text{Split}(F^k) \tag{4}$$

其中，f^k_{nr} 和 f^k_{ni} 分别是说话人 $n \in \{1, 2\}$ 的实部和虚部特征图。然后，对每组内的特征图沿通道维度求平均。对于每个说话人，在第 k 层生成语音信号在 STFT 域中的低分辨率估计：

$$\hat{S}^k_n = \bar{f}^k_{nr} + i\bar{f}^k_{ni} \tag{5}$$

¯其中，f 是平均后的特征图，i 表示虚数单位。对于每个中间层，计算低分辨率估计与干净信号之间的基于位置的训练损失，并将其加到最终损失中：

$$\mathcal{L}_{\text{LBT}}(\hat{S}, S) + \sum_{k}^{K_d-1} \mathcal{L}_{\text{LBT}}(\hat{S}^k, g^{K_d-k}(S)) \tag{6}$$

其中 $g^x(.)$ 是一个二维平均池化函数，核大小为 2×2，步幅为 2×2，并递归应用 x 次。直观而言，我们分配一组卷积核，以不同的时间和频率分辨率学习每位说话人的实部和虚部频谱图。根据排序后的说话人方位角或距离，每个解码器层中的卷积核分配都保持一致。

请注意，这与多尺度损失 [17] 不同；后者为基于 PIT 的分离模型使用共享解码器，要求在每一层之后生成波形估计。我们强调，卷积核分配保持一致对于多分辨率损失至关重要。我们预计，将多分辨率损失与置换不变训练准则结合使用效果不佳，因为 PIT 允许标签置换，卷积核的顺序可能因混合信号而异。

3. 实验设置

3.1. 数据集

我们使用 Lib- riCSS 语料库 [2]评估所提出的分离模型。该语料库用于一项包含语音重叠的对话语音识别任务。该数据集包含 10 小时从 Librispeech 开发集获取的录音，这些录音经扬声器重放，以捕捉真实房间混响。扬声器围绕桌子摆放，与麦克风阵列的距离为 33 至 409 厘米。录音分为 6 个会话，重叠率各不相同：0S（无重叠，话语间隔 0.1–0.5 秒）、0L（无重叠，话语间隔 2.9–3.0 秒）、10%、20%、30% 和 40% 的重叠。LibriCSS 语料库使用半径为 4.25 厘米的 7 通道圆形麦克风阵列录制。

我们遵循 [8] 中描述的设置，使用镜像法 [18, 19] 模拟训练数据和验证数据。我们使用 LibriSpeech 数据集创建了 192K 个具有不同重叠比例的双说话人混合信号。为模拟 LibriCSS 录音设备，我们将每个混合信号与 7 通道麦克风阵列的 RIR 卷积。为生成 RIR，我们模拟长、宽、高尺寸随机取自 [5×5×3, 10×10×4] 米范围内的矩形房间，并将麦克风阵列放置在房间中央。说话人位置从 360 个候选方位角中均匀采样，范围为 -180° 至 180°，分辨率为 1°。随机选择说话人与阵列之间的距离，使两位说话人至少相距 5 厘米。混响时间（T60）在 0.2 至 0.6 秒之间随机采样。此外，还以 10 至 30 dB 的随机信噪比向混合信号中添加稳态环境噪声。

使用 LibriCSS 语料库提供的默认自动语音识别系统 [2] 对分离模型进行评估。LibriCSS 包含两种评估场景：按话语评估和连续评估。在按话语评估场景中，计算预先分段的混合信号的两个分离信号各自的词错误率（WER），并选取词错误率较低的结果。连续评估场景则评估包含 8–10 个话语的长混合信号片段。将两个流的解码结果合并后进行评估。

在两种场景中，输入混合信号均按 2.4 秒的片段进行处理，片段移位为 1.2 秒。处理后的片段使用 [2] 提出的拼接算法进行拼接。所有信号均以 16 kHz 采样。我们采用帧长为 32 ms、帧移为 8 ms 的短时傅里叶变换（STFT）。对多通道输入混合信号的样本方差进行归一化。此外，还对短时傅里叶变换特征进行全局均值和方差归一化。

3.2. 说话人计数

对于没有说话人重叠的片段，分离模型应在其中一个输出中生成零信号。然而，观察发现，模型有时会为非重叠片段输出可辨识的残余信号，从而导致更多解码错误。随后 [8], 我们使用帧级说话人计数（SC）来找出非重叠帧，并通过合并两个分离后的输出去除残余信号。对于每一帧，SC 都会生成三类的概率：零个、一个或两个说话人。

表 1：LibriCSS 上各对比系统在按话语评估和连续评估中的词错误率（%）。‘$MISO_1$+SC’ 和 ‘MC-CSM+SC’ 分别指采用说话人计数网络的 $MISO_1$ 和 MC-CSM。

			按话语						连续					
	评估标准	多分辨率分辨率	0S	0L	10%	20%	30%	40%	0S	0L	10%	20%	30%	40%
未处理	–	–	11.8	11.7	18.8	27.2	35.6	43.3	15.4	11.5	21.7	27.0	34.3	40.5
双向长短期记忆网络 [2]	PIT	–	8.3	8.4	11.6	16.0	18.4	21.6	11.9	9.7	13.4	15.1	19.7	22.0
Conformer [12]	PIT	–	7.2	7.5	9.6	11.3	13.7	15.1	11.0	8.7	12.6	13.5	17.6	19.6
$MISO_1$ [8]	PIT	–	7.7	7.5	7.9	9.6	11.3	13.0	10.7	10.5	10.9	11.5	13.8	15.3
$MISO_1$+SC [8]	PIT	–	–	–	–	–	–	–	7.9	8.5	8.5	10.5	12.3	14.3
MC-CSM	PIT	–	7.0	7.5	7.9	9.3	11.3	13.0	10.0	9.4	11.1	11.8	14.3	15.2
	方位角	–	8.6	8.8	9.2	10.3	11.3	12.4	13.5	14.5	13.6	13.4	14.7	15.4
	PIT	✓	7.6	7.6	7.8	9.4	11.0	12.8	11.4	12.5	11.3	13.0	15.1	16.3
	方位角	✓	7.3	7.9	7.6	**9.2**	**10.8**	**11.4**	9.2	9.8	9.1	10.5	12.2	**12.8**
	距离	✓	8.7	10	10.5	12.9	14.6	17.8	12.2	12.8	12.4	14.8	17.1	19.0
MC-CSM+SC	PIT	–	**6.5**	**6.9**	7.4	9.3	11.7	13.6	8.0	**7.4**	8.8	10.7	12.6	13.8
	方位角	–	7.4	7.7	8.1	9.5	11.3	12.5	7.9	7.9	8.5	9.4	**11.6**	12.9
	PIT	✓	**6.5**	7.3	**7.3**	9.6	11.3	13.5	8.1	8.1	8.3	10.2	12.4	14.0
	方位角	✓	6.7	7.1	7.5	**9.2**	10.9	12.1	**7.7**	7.7	**8.0**	**9.3**	**11.6**	13.0
	距离	✓	7.6	8.3	9.3	12.1	14.2	17.8	8.3	8.4	9.4	11.7	14.9	16.8

SC 模型基于经过修改的 MC-CSM 架构。具体而言，SC 由 MC-CSM 编码器、两个双向长短期记忆网络（BLSTM）层和一个 softmax 层组成。用于训练的 SC 标签由一个经过预训练、基于深度神经网络（DNN）的语音活动检测器生成 [20]。

4. 评估结果

我们使用置换不变训练（PIT）和基于位置的训练（LBT）准则训练 MC-CSM，并分别采用和不采用多分辨率损失。对于 LBT，输出说话人的分配顺序依据说话人方位角确定。表 1 展示了 MC-CSM 模型按话语评估和连续评估的词错误率（WER）结果。我们发现，在重叠率较高时，采用方位角准则训练的 MC-CSM 与 PIT 的结果相当，但在重叠率较低时表现不如 PIT。不过，在非重叠片段中使用说话人计数（SC）抑制残余信号时，基于方位角的训练取得了明显提升。结合多分辨率损失和基于方位角的训练后，语音识别性能出色，且无论是否使用说话人计数，表现都优于 PIT。例如，在连续评估中，基于多分辨率 LBT 的 MC-CSM 在 40% 重叠条件下，相较于 PIT 将 WER 绝对降低了 2.4%。结果表明，使用模拟房间脉冲响应训练的 LBT 模型能够很好地泛化到真实的麦克风阵列录音。另一方面，采用 PIT 的分离模型在使用多分辨率损失后性能并未提升。这是因为 PIT 准则不考虑顺序，而不同混合信号的特征图在多分辨率损失下无法保持一致的排序。

我们还使用说话人……训练了带有 LBT 的 MC-CSM 使用距离而非方位角。如表中所示，基于距离的训练表现不如基于方位角的训练和 PIT。可能的原因是，在 LibriCSS 数据集中，距离准则的区分能力不足，因为扬声器呈对称排列，且它们与麦克风阵列的距离相近。

最后，我们将 MC-CSM 模型与其他具有竞争力的分离方法进行比较，这些方法均使用 LibriCSS 默认自动语音识别系统进行评估。对于所有方法，我们列出已报告的最佳结果，并将未报告的字段留空。 [2] 和 [12] 中提出的系统分别采用双向长短期记忆网络和 Conformer 架构来估计实值掩码。 [8] 中的系统使用带有时间卷积网络的 U-Net 模型进行复数谱映射。我们还使用了专门的说话人计数网络，以进一步改善词错误率。与其他系统相比，我们使用多分辨率损失训练的 LBT 模型取得了好得多的词错误率。值得一提的是，如 [8]所示，在分离模型之上使用语音增强网络，可以进一步降低词错误率。本研究的重点是分离模型；通过引入语音增强，还有望进一步提升性能。

5. 结语

在本研究中，我们将 LBT 扩展到了对话式多说话人分离。我们提出了一种多分辨率损失，可根据说话人位置一致地分配卷积核。我们证明，在 LibriCSS 语料库上，LBT 的分离和自动语音识别性能显著优于其他有竞争力的方法。

6. 参考文献

[1] D. L. Wang 和 J. Chen，“基于深度学习的监督语音分离：综述”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 26 卷，第 1702–1726 页，2018 年。[2] Z. Chen、T. Yoshioka、L. Lu、T. Zhou、Z. Meng、Y. Luo、J. Wu、X. Xiao 和 J. Li，“连续语音分离：数据集与分析”，载于*ICASSP* 会议论文集，2020 年，第 7284–7288 页。[3] H. Taherian 和 D. L. Wang，“基于重叠率的时域损失调制，用于单声道对话说话人分离”，载于*ICASSP* 会议论文集，2021 年，第 5744–5748 页。[4] Y. Zhang、Z. Chen、J. Wu、T. Yoshioka、P. Wang、Z. Meng 和 J. Li，“采用循环选择性注意力网络的连续语音分离”，载于*ICASSP* 会议论文集，2022 年，第 6017–6021 页。[5] W. Zhang、Z. Chen、N. Kanda、S. Liu、J. Li、S. Emre Eskimez、T. Yoshioka、X. Xiao、Z. Meng、Y. Qian 和 F. Wei，“采用组级置换不变训练分离长篇语音”，载于*Interspeech* 会议论文集，2022 年，第 5383–5387 页。[6] T. von Neumann、K. Kinoshita、C. Boeddeker、M. Delcroix 和 R. Haeb-Umbach，“Graph-PIT：用于任意数量说话人连续分离的广义置换不变训练”，载于*Interspeech* 会议论文集，2021 年，第 3490–3494 页。[7] C. Li、Z. Chen 和 Y. Qian，“采用记忆嵌入模型的双路径建模，用于连续语音分离”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 30 卷，第 1508–1520 页，2022 年。[8] Z.-Q. Wang、P. Wang 和 D. L. Wang，“用于按话语和连续语音分离的多麦克风复数谱映射”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 2001–2014 页，2021 年。[9] K. Saijo 和 R. Scheibler，“用于无监督多通道声源分离的空间损失”，载于*Interspeech* 会议论文集，2022 年，第 241–245 页。[10]C. Boeddeker、T. Cord-Landwehr、T. von Neumann 和 R. Haeb-Umbach，“基于空间混合模型的会议分离初始化方案”，载于*Interspeech* 会议论文集，2022 年，第 271–275 页。[11] H. Taherian、K. Tan 和 D. L. Wang，“通过基于位置的训练实现多通道说话人无关的说话人分离”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 30 卷，第 2791–2800 页，2022 年。[12] S. Chen 、 Y. Wu 、 Z. Chen 、 J. Wu、J. Li、T. Yoshioka、C. Wang、S. Liu 和 M. Zhou，“采用 Conformer 的连续语音分离”，载于*ICASSP* 会议论文集，2021 年，第 5749–5753 页。[13] D. S. Williamson、Y. Wang 和 D. L. Wang，“用于单声道语音分离的复数比率掩蔽”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 24 卷，第 483–492 页，2016 年。[14] K. Tan、Z.-Q. Wang 和 D. L. Wang，“神经谱空域滤波”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 30 卷，第 605–621 页，2022 年。[15] Y. Liu 和 D. L. Wang，“分而治之：一种用于说话人无关的单声道说话人分离的深度 CASA 方法”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 27 卷，第 2092–2102 页，2019 年。

[16] Z.-Q. Wang 和 D. L. Wang，《基于深度学习的语音去混响目标抵消》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，第 941–950 页，2020 年。[17] E. Nachmani、Y. Adi 和 L. Wolf，《未知数量多个说话人的语音分离》，载于国际机器学习会议论文集，2020 年，第 7164–7175 页。[18] J. B. Allen 和 D. A. Berkley，《用于高效仿真小房间声学的镜像法》，美国声学学会杂志，第 65 卷，第 4 期，第 943–950 页，1979 年。[19] R. Scheibler、E. Bezzam 和 I. Dokmani´c，《Pyroomacoustics：用于音频房间仿真和阵列处理算法的 Python 软件包》，载于*ICASSP* 论文集，2018 年，第 351–355 页。[20]V. Manohar。ASpIRE 语音活动检测模型。 [在线]。网址：https://kaldi-asr.org/models/m4