arXiv:2601.02231v1 [eess.AS] 5 Jan 2026

空间特征在基于基础模型的说话人日志中的作用

Marc Deegen[1], *Tobias Gburrek*[1],*Tobias Cord-Landwehr*[1],*Thilo von Neumann*[1], *Jiangyu Han*[2],*Luk´aˇs Burget*[2],*Reinhold Haeb-Umbach*[1]

[1] 帕德博恩大学，通信工程系，德国[2] 布尔诺理工大学，Speech @FIT，捷克{deegen,gburrek,cord,vonneumann,haeb}@nt.upb.de{ihan, burget}@fit.vut.cz

摘要

近期，说话人日志领域借助 WavLM 等大型预训练基础模型，在多个数据集上取得了最先进的性能。DiariZen 等系统利用了丰富的单通道表征，但受限于单通道音频，无法使用多通道录音中可用的空间线索。本文通过评估多种基于多通道空间特征对模型进行条件化的策略，分析了将空间信息融入最先进的单通道说话人日志系统所带来的影响。在会议类数据集上进行的实验表明，空间信息可以提升说话人分段性能，但对于所提出的系统而言，整体提升幅度小于预期。这表明，对所有 WavLM 层的特征进行聚合，已经捕获了准确区分说话人所需的大部分信息，包括在重叠语音区域中的信息。这些发现揭示了利用空间线索增强基于基础模型的说话人日志的潜力与局限。

索引词— 说话人日志，WavLM，空间信息，远场会议数据，多通道音频

1. 引言

说话人日志是许多语音处理系统的基本组成部分，例如会议转录和多说话人自动语音识别（自动语音识别） [1–3]。它回答“谁在何时说话”这一问题，预测输入录音中每位说话人的时间活动。说话人日志信息可用于提升后续下游任务的性能。

说话人日志有多种范式。传统的模块化说话人日志系统依赖于提取并聚类说话人表征，例如 x 向量 [4]。端到端神经日志（EEND）方法直接根据输入音频预测逐帧说话人活动 [5, 6]。结合这两种范式的混合方法是端到端神经日志与向量聚类（EEND-VC）框架 [7], , 该框架在录音的短片段上局部执行 EEND，随后通过聚类跨片段提取的说话人嵌入，拼接片段级预测结果。

WavLM 等大型预训练基础模型的引入，显著提升了说话人分段性能。通过从大量无标注数据中学习，WavLM 能够提供强大的语音表征，有效降低对特定任务训练数据集的依赖。DiariZen 系统采用了这一方法，将 WavLM 提取的特征集成到 EEND-VC 框架中的基于 Conformer 的端到端神经说话人分离模型中，从而实现了最先进的说话人分段性能。

然而，由于大多数基础模型仅使用单通道音频进行预训练，依赖这些表征的系统无法利用多通道录音中包含的空间信息。与此相反，有些方法会显式利用空间信息进行说话人分段，例如通过估计接收语音的到达时间差（TDOA） [12–14] 或到达方向（DOA） [15]。事实证明，空间信息对于重叠语音区域尤其有益。在这些区域，纯频谱系统往往难以应对，而只要说话人位于不同空间位置，空间方法就能更有效地分离并归属同时发声的说话人 [12,13,15,16]。然而，与单通道系统相比，空间系统通常使用规模小得多的数据集进行训练，而单通道系统则可以从大量预训练数据中获益 [8]。

为了让单通道说话人分段系统利用多通道输入，可以采用 DOVER-Lap，将各个通道的输出合并为一个联合说话人分段假设 [17]。 [18], 提出了一种计算开销更低、但效果更好的方法：将通道间通信模块集成到 WavLM 特征提取的早期层中，从而使 WavLM 能够处理多通道输入。

在本文中，我们采用一种替代方法。我们开发了一个辅助网络，用于从多通道输入中提取空间信息，并将其输出与单通道 WavLM特征相结合。通过这种融合，系统不仅能够利用 WavLM特征所捕获的语义和声学表征，还能利用空间信息。

为此，我们分析了将空间辅助网络集成到 DiariZen 说话人分段流程中的多种方案，评估其支持说话人分离性能的适用性。在此，我们评估并分析了两种方法：一是直接引入利用神经网络从空间特征中提取的嵌入，二是与预训练的空间说话人分段模块进行融合；实验在多个会议类数据集上开展。此外，辅助网络的设计不依赖于输入通道数量和麦克风阵列几何结构，从而不会将原系统限制于特定的麦克风阵列。

本文其余部分安排如下：第 2 节概述 DiariZen 框架，并详细介绍如何使用辅助多通道网络整合所提出的空间特征。第 3 节介绍数据集和实验设置，随后从说话人分段错误率（DER）的角度展示并分析说话人分段性能，最后在第 4 节给出结论。

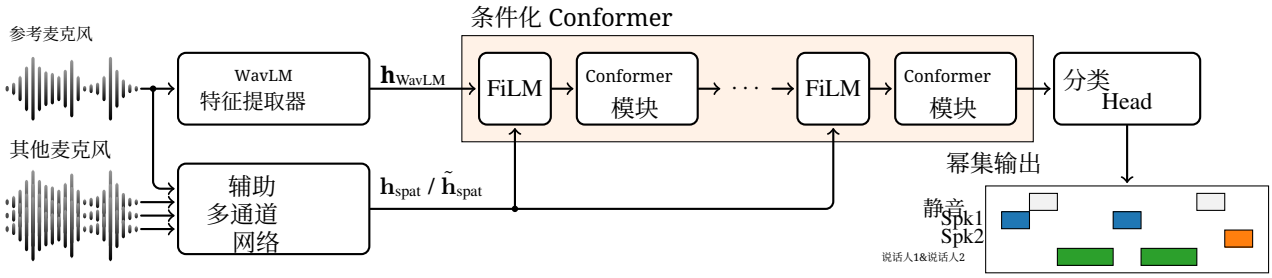


图 1。空间信息增强的 DiariZen 架构概览。首先提取单通道 WavLM 特征，然后使用 FiLM 层将其与空间线索相结合。空间线索由辅助多通道网络提取，该网络以由相位差和幅度构成的空间特征作为输入。

2. 空间信息辅助的 DiariZen

本文的分析基于单通道多说话人说话人分段框架 DiariZen [9]。为研究从多通道信号中提取的空间信息对说话人分段的影响，本文对 DiariZen 进行了扩展，加入了空间特征提取模块，如图 1 所示。这里假设采用紧凑型麦克风阵列配置，可提供通道间相位差（IPDs） [19] 和幅度信息等空间线索，这些信息与基于 TDOA 和到达方向的方法中用于说话人分段的信息密切相关。本文将这些特征与 DiariZen 框架中的 WavLM特征相结合，以分析空间信息能否进一步提升说话人分离性能。

2.1. DiariZen

DiariZen 遵循 EEND-VC [7,20] 框架，其中 EEND 模块采用针对说话人分段场景微调的 WavLM [8] 特征提取器。在 EEND-VC 框架中，输入音频被划分为较短的重叠片段。EEND [5,6] 模型会分别处理每个片段，首先估计帧级说话人活动，并生成局部说话人日志输出。由于各片段是独立处理的，说话人身份在不同片段间并不一致，因此需要额外的对齐与合并阶段来解决说话人标签歧义。此过程通过后续的向量聚类（VC）实现：从非重叠语音区域中提取每个局部检测到的说话人的说话人嵌入，并使用凝聚层次聚类（AHC）或基于 x 向量的变分贝叶斯 HMM 聚类（VBx） [21], 在整段录音范围内进行聚类，同时约束来自同一片段的嵌入不能合并。

本研究聚焦于图 2 所示的 DiariZen 框架中的局部 EEND 模块。DiariZen 通过对所有 WavLM 层的输出进行可学习加权求和，提取 WavLM特征。聚合后的 WavLM特征经线性层投影并进行层归一化，然后输入带有分类头的 Conformer。该分类头采用幂集分类进行训练，以预测说话人分段输出 [22]。在幂集分类中，所有可能的活跃说话人组合（包括静音类别、单个活跃说话人以及重叠说话人）都被表示为不同的目标类别。这种方法

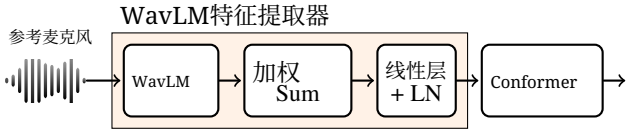


图 2. DiariZen 框架中的局部 EEND 模块示意图（改编自 [9]）。

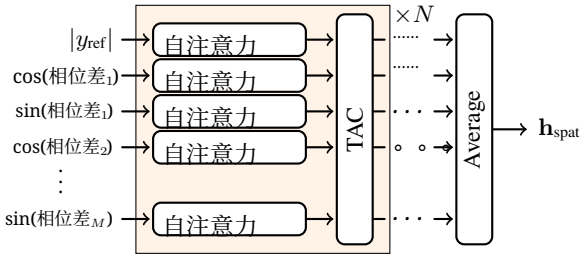


图 3。用于估计空间线索 $\mathbf{h}_{\text{spatial}}$ 的空间编码器架构。所有变换后的输入特征依次通过 N 个编码器层，各层采用共享权重的自注意力，并通过 TAC连接；最后一层之后对输出取平均。

有效地将多标签说话人活动检测问题转化为单标签多类分类问题，并可通过交叉熵损失进行优化。

2.2. 辅助多通道网络

空间信息通过计算所有 M 个非冗余麦克风对中第 m对的相位差特征 IPD_m 来获取。为解决固有的相位不连续性，对相位进行正弦和余弦变换，从而获得相位的连续表示 [23]。这些特征以及第一个麦克风通道的幅度谱图 y_{ref} 共同构成辅助多通道网络的空间输入特征。

2.2.1. 空间编码器

本文测试了辅助网络的两种变体。第一种称为空间编码器，如图 3 所示。在每个编码器层中，对所有空间特征应用具有共享权重的自注意力，并在自注意力之后通过变换、平均和拼接（TAC） [24] 连接促进通道间交互。该架构实现了跨通道信息交换，并将空间特征转换到嵌入空间中，因此在经过最后一个编码器层后，可以跨通道对表示进行平均，而不会丢失关键的空间信息，从而得到单通道空间嵌入 $\mathbf{h}_{\text{spat}}$。因此，所得的空间编码器及其后续模块均不依赖于输入通道数和具体的麦克风阵列配置。

2.2.2. 空间 *Conformer* 和空间说话人分离

或者，也可以通过增加投影层、层归一化和一个额外的 Conformer 来扩展空间编码器，使其结构类似于图 2 所示的单通道 DiariZen 系统。其目标

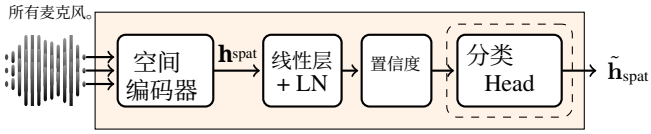


图 4。空间 Conformer 和空间说话人分离网络的示意图。空间编码器提取的特征被用作与 DiariZen 类似的网络结构的输入。

˜是在空间辅助网络中获得更复杂的抽象表示$\mathbf{h}^{\text{spat}}$，供 DiariZen 网络使用。下文将此配置称为“空间 Conformer”，如图 4 所示。进一步级联分类层后，该网络可作为独立的空间说话人分段模块进行训练，称为“空间说话人分段”配置；它还可以作为辅助网络，为 DiariZen 系统提供空间线索$\mathbf{h}^{\text{spat}}$。

2.2.3. DiariZen 的空间条件化

˜辅助网络的输出 $\mathbf{h}^{\text{spat}}$ 或 $\mathbf{h}^{\text{spat}}$被集成到 DiariZen 中，作为特征级线性调制（FiLM） [25] 层的条件输入，如图 1 所示。首先，在 Conformer 之前应用一个 FiLM层，然后在每个 Conformer模块之前再应用一个 FiLM层，以确保整个网络始终能够获得空间信息。

3. 实验

实验采用了文献 [10], 中介绍的 DiariZen 剪枝和微调版本的重新实现。配置中的超参数（如分段长度和跳跃步长）遵循 DiariZen 框架的设置。用于训练系统的幂集分类假设每帧最多有 2 名并发说话人。在所有多通道实验中，无论相应数据集提供的麦克风总数是多少，均使用四个麦克风。麦克风的选择以最大化麦克风之间的间距为原则，从而确保最佳地捕获空间信息。

请注意，本研究关注的是局部 EEND 模块的性能。因此，在以下评估中，片段级端到端神经说话人日记输出以预言机方式拼接。将估计的局部说话人活动与真实活动进行比较，为每个局部说话人关联一个真实说话人标签。分配的预言机说话人标签用于消除不同片段之间的排列歧义。这样，性能评估便可聚焦于局部 EEND 模块的性能，而空间特征的潜在优势也应能在此显现。此外，计算对话错误率时不使用容差区间。

在系统的训练和评估中，使用了多说话人、会议场景和多通道数据集 AMI [26], AliMeet- ing [27], AISHELL-4 [28], 以及 NOTSOFAR-1 [29] 。由于 AISHELL-4 数据集未提供官方开发集，因此采用与 DiariZen 相同的开发集划分。与 [18], 中的实验不同，本分析不包含 CHiME-6 [30] 数据集。其多房间录音设置很可能会使性能提升主要源于不同房间之间录音条件的差异，而非有效利用空间线索，因此不适合公平评估空间信息的影响。训练使用了全部四个数据集的训练集组合。表1显示了活跃说话人的数量、各数据集及其组合数据集的规模。

表1。数据集属性（#Spk = #说话人数，#Hrs =#小时数）。

数据集	训练		Dev		Test	
	#Spk	#Hrs	#Spk	#Hrs	#Spk	#Hrs
AMI	3–5	79.7	4	9.7	3–4	9.1
AISHELL-4	3–7	97.2	3–7	10.3	5–7	12.7
AliMeeting	2–4	111.4	2–4	4.2	2–4	10.8
NOTSOFAR-1	4-8	39.8	4-6	13.4	3-7	16.5
综合	2–8	328.1	2–7	37.6	2–7	49.1

3.1. 参考系统

首先，在表2中评估单通道 DiariZen（编号 1）基线。该基线在四个数据集上的宏观说话人错误率为 12.2%，其中单说话人区域为 8.7%，重叠语音区域为 20.1%。图 4 所示的纯空间系统“空间说话人分离”（编号 4）用作第二个基线。在该系统中，说话人分段仅采用第 2.2.2 节所述的辅助网络。该网络使用与 DiariZen 基线相同的幂集损失和说话人分离目标进行训练，宏观说话人错误率为 14.3%，其中单说话人区域为 10.6%，重叠区域为 23.6%。

尽管整体性能不如 DiariZen 基线，但结果表明，空间说话人分离系统在重叠语音区域也未能带来预期的改进，这说明空间特征在处理重叠方面提供的额外益处有限。由于 WavLM 最初采用掩码预测损失 [8], 进行训练，因此其主要针对单人说话建模进行了优化。我们推测，WavLM特征在重叠区域表现出乎意料地好，可能是因为模型采用了跨所有 WavLM 层的可学习加权和，从而能够整合较早层中的信息；这些层更接近原始波形，可能已经包含有助于区分重叠说话人的线索。对所学习特征和权重的进一步分析留待未来工作。不过，仅凭空间特征便能实现有效的说话人分段，这一事实表明，结合空间特征与频谱信息或可进一步提升整体说话人分离性能。

此外，作为上限基准，我们评估了 DiariZen + 预言机说话人数（编号 9）。该系统纳入了每帧的预言说话人数量信息，宏观说话人错误率为 4.9%（编号 8）。在此设置中，预言说话人数被用作 Conformer 之前 FiLM 层的条件信号，而 Conformer 内部不再应用额外的 FiLM 条件化。

3.2. 空间支持的DiariZen

为评估在 DiariZen 框架中集成空间特征的效果，我们评估了第 2.2 节所述的不同辅助多通道网络。首先，我们采用了图 3 所示、第 2.2.1 节所述的空间编码器辅助网络。训练期间，空间编码器采用随机初始化，并与剪枝后的 DiariZen 框架中的预训练 WavLM 和 Conformer 模块联合训练。然而，该 DiariZen +空间编码器系统（编号 5）的表现与 DiariZen 基线相当，宏观说话人错误率为 12.5%，这表明在此配置下，编码后的空间特征并未带来可测量的收益。

鉴于空间编码器相较于其他组件较为轻量，因此在 DiariZen + 空间 Conformer（ID 6）系统中探索了更大的模型变体。这里评估的是第 2.2.2 节所述、如图 4 所示的辅助网络，去除了分类头且未进行说话人分离预训练。尽管模型容量有所提升，该系统的宏观说话人错误率仅为

表 2。采用预言聚类，对所提出的空间支持系统与单人说话/多通道 DiariZen 系统进行对话错误率（DER）比较，并分别报告重叠语音（OV）区域和单说话人（Single）区域的结果。

ID	系统	AMI (重叠语音 / 单说话人)	AliMeeting (重叠语音 / 单说话人)	AISHELL-4 (重叠语音 / 单人说话)	NOTSOFAR-1 (重叠语音 / 单人说话)	宏平均 (重叠语音 / 单人说话)
1	DiariZen [9]	13.1（21.7 / 9.9）	12.5（22.2 / 7.0）	9.1（16.4 / 8.3）	14.2（19.9 / 9.4）	12.2（20.1/ 8.7）
2	DiariZen-Large Conformer	13.2（21.3 / 10.2）	12.6（22.1 / 7.1）	9.6 (16.0 / 8.9)	14.1（19.7 / 9.5）	12.9 (19.8/ 8.9)
3	多通道 DiariZen [18]	12.8（- / - ）	12.0（- / - ）	**8.9**（- / - ）	14.1（- / - ）	12.0（- / - ）
4	空间说话人分离	14.5（23.9 / 11.1）	14.0 (25.4 / 7.4)	10.0（22.5 / 8.6）	18.5（22.5 / 15.1）	14.3（23.6 / 10.6）
5	DiariZen +空间编码器	13.5（22.4 / 10.2）	12.6（22.1 / 7.1）	9.5 (17.9 / 8.6)	14.3（20.2 / 9.4）	12.5（20.7/ 8.8）
6	DiariZen +空间 Conformer	13.5（22.1 / 10.3）	13.1 (23.0 / 7.3)	9.4（18.4 / 8.5）	14.7（20.4 / 9.9）	12.7（21.0/ 9.0）
7	DiariZen +空间说话人分离	12.5（20.8 / 9.4）	12.1（21.4 / 6.7）	**8.9** (18.5 / 7.8)	13.5（19.0 / 8.8）	11.7（19.9/ 8.2）
8	+ 联合微调	**12.2**(20.5 / 9.2)	**11.8** (21.2 / 6.3)	**8.9** (17.4 / 8.0)	**13.4**(18.8 / 8.8)	**11.6** (19.5/ 8.1)
9	DiariZen + 真实说话人数	3.6（10.1 / 1.1）	6.0 (14.6 / 1.1)	1.6 (4.8 / 1.2)	8.2 (14.7 / 2.8)	4.9（11.1/ 1.6）

表 3。采用 VBx 聚类的部分系统的宏平均说话人错误率表现。

ID	系统	宏观说话人错误率 (OV / 单人说话)
1	DiariZen [9]	14.8（27.1 / 9.6）
4	空间说话人分离	16.8（27.7 / 11.6）
7	DiariZen +空间说话人分离	14.3 (26.9 / 9.2)
8	+联合微调	14.1 (26.1 / 9.1)
9	DiariZen +真实说话人数	9.0（22.4 / 3.1）

12.7%，同样未能优于 DiariZen 基线。

随后，对 DiariZen +空间说话人分离（编号 7）配置进行了评估，该配置将第 3.1 节中的预训练空间说话人分离流程用作辅助多通道网络。此设置在空间 Conformer 上增加了一个分类头，更重要的是，利用针对说话人分离目标的预训练，旨在提供结构更完善、对说话人活动更具判别力的空间线索。在此，空间说话人分离流程保持冻结，以保留其已学习的说话人分段能力，同时对其余模块进行微调，使其适应用于条件控制的空间表示。该系统的宏观说话人错误率为 11.7%，比 DiariZen 基线提高了 0.5 个百分点。值得注意的是，这一提升在所有评估数据集上均保持一致。

为确认观察到的提升并非仅仅源于参数数量增加，我们进行了另一项实验：DiariZen-Large Conformer（编号 2），将 Conformer 的规模扩大了一倍。如表 2 所示，这个更大的模型并未带来任何性能提升，其宏观说话人错误率为 12.9%。这表明，Di- ariZen + 空间说话人分离（编号 7）所取得的提升确实归因于空间线索的集成，而非模型规模的增大。

最后，对经过预训练的 DiariZen + 空间说话人分离配置进行微调（编号 8），同时解除空间说话人分离辅助网络的冻结。这样，空间说话人分段模块便能与 Conformer 及通过 FiLM 层进行的调制共同适配，最终宏观说话人错误率达到 11.6%，比纯频谱 DiariZen 系统提高了 0.6 个百分点。

3.3. 分析与讨论

结果表明，与单通道系统相比，在预言聚类设置下，利用空间信息的 DiariZen 系统能够提升说话人分段性能。表 3 进一步表明，采用 VBx 聚类而非预言聚类时，观察到的性能提升依然存在，正如 {v1} 中所示。与基线相比，取得了类似的相对增益，这表明在局部 EEND 模块层面获得的收益能够有效地延续至完整的说话人分段流程。然而，由于系统的幂集输出最多只能包含两名同时说话的说话人，因此在存在三名或更多活跃说话人的片段中，性能必然受到限制；这些片段约占评估数据总时长的 3.5%。

总体而言，加入空间信息带来的提升并没有最初预期的那么大，尤其是单通道系统在重叠语音区域中的表现已经出乎意料地好。正如第 3.1 节所讨论的，一种可能的解释是，对所有 WavLM 层进行可学习的加权组合，已经使模型能够获取有助于区分重叠说话人的信息。

利用多通道输入的另一种方法是让 WavLM 能够处理多通道数据，正如 [18] 中提出的那样。表 2 中的编号 3 展示了取得的结果（摘自 [18]）。可以看出，该方法取得了与本文提出的方法相近的提升。我们认为，这是让 DiariZen 具备多通道处理能力的两种并行方案：我们的方法引入了明确的空间线索，另一种则针对多通道处理对基础模型本身进行调整。

4. 结论

本研究探讨了空间信息是否以及如何进一步提升一种基于自监督基础模型特征的先进单通道说话人分段系统。为此，我们分析了将多通道录音中的空间线索融入 DiariZen 框架的多种策略。采用未经训练的辅助编码器整合空间线索，并不能提升说话人分离性能；而采用针对说话人分离目标预训练的空间说话人分段网络，则能在所有评估数据集上带来幅度虽小但稳定的提升。结果证实，在真实的会议场景中，空间信息可以补充单通道表征。然而，正如最初预期的那样，这些提升并未集中在重叠语音区域。

5. 致谢

本文所述研究始于 JSALT 2025 期间，并由 JHU 提供支持。BUT 的研究人员则通过 OP JAK 项目 CZ.02.01.010023 0200008518，获得捷克共和国教育、青年和体育部（教育部）的支持。IT4I 超级计算机上的计算工作由教育部通过 e-INFRA CZ（编号：90254）提供支持。

6. 参考文献

[1] T. J. Park、N. Kanda、D. Dimitriadis、K. J. Han、S. Watanabe 和 S. Narayanan，“说话人分段综述：深度学习的最新进展”，《计算机语音与语言》，第 72 卷，页码 101317，2022 年。[2] S. Horiguchi、Y. Fujita、S. Watanabe、Y. Xue 和 P. Garcia，“ 用于端到端神经说话人分段的基于编码器 - 解码器的吸引子”，《*IEEE*音频、语音与语言处理汇刊》，第 30 卷，页码 1493–1507，2022 年。[3] I. Medennikov、M. Korenevsky、T. Prisyach、Y. Khokhlov、M. Korenevskaya、I. Sorokin、T. Timofeeva、A. Mitrofanov 等，“目标说话人语音活动检测：晚宴场景下多说话人分离的一种新方法”，载于*ISCA I nterspeech*论文集，2020 年，页码 274–278。[4] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur，“X 向量：用于说话人识别的鲁棒深度神经网络嵌入”，载于*IEEE*国际声学、语音与信号处理会议论文集，2018 年，页码 5329–5333。[5] Y. Fujita、N. Kanda、S. Horiguchi、K. Nagamatsu 和 S. Watanabe，“采用无排列目标函数的端到端神经说话人分段”，载于*ISCA Interspeech*论文集，2019 年，页码 4300–4304。[6] Y. Fujita、N. Kanda、S. Horiguchi、Y. Xue、K. Nagamatsu 和 S. Watanabe，“采用自注意力的端到端神经说话人分段”，载于*IEEE*自动语音识别与理解研讨会论文集，2019 年，页码 296–303。[7] K. Kinoshita、M. Delcroix 和 N. Tawara，“真实对话语音中端到端神经分段与基于聚类的分段集成技术的进展”，载于*ISCA Interspeech*论文集，2021 年，页码 3565–3569。[8] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda 等，“Wavlm：面向全栈语音处理的大规模自监督预训练”，《*IEEE*信号处理精选主题期刊》，第 16 卷，第 6 期，页码 1505–1518，2022 年。[9] J. Han、F. Landini、J. Rohdin、A. Silnova、M. Diez 和 L. Burget，“利用自监督学习进行说话人分段”，载于*IEEE*国际声学、语音与信号处理会议论文集，2025 年。[10] J. Han、P. P´alka、M. Delcroix、F. Landini、J. Rohdin、J. Cer`nocky 和 L. Burget，“通过对自监督模型进行结构化剪枝，实现高效且可泛化的说话人分段”，*arXiv*，2025 年。[11] A. Gulati、J. Qin、C.-C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang 等，“Conformer：用于语音识别的卷积增强型Transformer”，载于*Proc. ISCA Inter- speech*论文集，2020 年，页码 5036–5040。[12] T. Gburrek、J. Schmalenstroeer 和 R. Haeb-Umbach，“利用自组织声学传感器网络进行会议转录的空间说话人分段”，载于阿西洛马会议：信号、系统与计算机，2023 年，页码 1399–1403。[13] T. Cord-Landwehr、T. Gburrek、M. Deegen 和 R. Haeb-Umbach，“通过结合基于到达时间差的分段与基于说话人嵌入的聚类进行会议的时空频谱说话人分段”，载于*Proc. ISCA Interspeech*论文集，2025 年，页码 5223–5227。[14] S. Horiguchi、Y. Takashima、P. Garc´ıa、S. Watanabe 和 Y. Kawaguchi，“采用分布式麦克风的多通道端到端神经说话人分段”，载于*IEEE*国际声学、语音与信号处理会议论文集，2022 年，页码 7332–7336。

[15] S. Araki、M. Fujimoto、K. Ishizuka、H. Sawada 和 S. Makino，“一种用于真实会议的基于到达方向的说话人分段系统”，载于*HSCMA*论文集，2008年，第29–32页。[16] J. Wang、Y. Liu、B. Wang、Y. Zhi、S. Li、S. Xia、J. Zhang、F. Tong等，“面向多通道多方会议的空间感知说话人分段”，载于*ISCA Interspeech*论文集，2022年。[17] D. Raj、P. Garcia、Z. Huang、S. Watanabe、D. Povey、A. Stolcke 和 S. Khudanpur，“DOVER-Lap：一种融合重叠感知说话人分段输出的方法”，*IEEE SLT*论文集，2021年。[18] J. Han、R. Wang、Y. Masuyama、M. Delcroix、J. Rohdin、J. Du 和 L. Burget，“用于多通道神经说话人分段的空间感知自监督模型”，载于*arXiv*预印本，2025年。[19] H. Song 和 J. W. Shin，“利用频谱掩码和全频段通道间相位差进行多声源定位”，载于*ISCA Interspeech*论文集，2021年，第671–675页。[20] K. Kinoshita、M. Delcroix 和 N. Tawara，“整合端到端神经分段与基于聚类的日志：兼得两者之长”，载于*IEEE*国际声学、语音与信号处理会议论文集，2021年，第7198–7202页。[21] F. N. Landini、J. Profant、M. Diez S´anchez 和 L. Burget，“说话人分段中x-vector序列的贝叶斯隐马尔可夫模型聚类（VBx）：理论、实现及标准任务分析”，《计算机语音与语言》，第71卷，编号101254，第1–16页，2022年。[22] A. Plaquet 和 H. Bredin，“用于神经说话人分段的幂集多类交叉熵损失”，载于*ISCA Interspeech*论文集，2023年，第3222–3226页。[23] Z.-Q. Wang、J. Le Roux 和 J. R. Hershey，“多通道深度聚类：用于说话人无关语音分离的判别式频谱与空间嵌入”，载于*IEEE*国际声学、语音与信号处理会议论文集，2018年，第1–5页。[24] Y. Luo、Z. Chen、N. Mesgarani等，“端到端麦克风排列与数量不变的多通道语音分离”，载于*IEEE*国际声学、语音与信号处理会议论文集，2020年，第6394–6398页。[25] E. Perez、F. Strub、H. de Vries、V. Dumoulin 和 A. Courville，“FiLM：利用通用条件层进行视觉推理”，*AAAI*人工智能大会论文集，第32卷，第1期，2018年4月。[26] J. Carletta、S. Ashby等，“AMI会议语料库：预告”，载于*MLMI*论文集，2005年，第28–39页。[27] F. Yu、S. Zhang、Y. Fu、L. Xie、S. Zheng、Z. Du、W. Huang、P. Guo等，“M2met：2022年IEEE国际声学、语音与信号处理会议多通道多方会议转录挑战赛”，载于*IEEE*国际声学、语音与信号处理会议论文集，2022年，第6167–6171页。[28] Y. Fu、L. Cheng、S. Lv等，“AISHELL-4：用于会议场景下语音增强、语音分离、语音识别和说话人分段的开源数据集”，载于*ISCA Interspeech*论文集，2021年，第3665–3669页。[29] A. Vinnikov、A. Ivry、A. Hurvitz、I. Abramovski、S. Koubi、I. Gurvich、S. Peer、X. Xiao等，“NOTSOFAR-1挑战赛：远距离会议转录的新数据集、基线和任务”，载于*ISCA Interspeech*论文集，2024年，第5003–5007页。[30] S. Watanabe、M. Mandel、J. Barker、E. Vincent、A. Arora、X. Chang、S. Khudanpur、V. Manohar等，“CHiME-6挑战赛：应对未分段录音中的多说话人语音识别”，载于*CHiME*论文集，2020年，第1–7页。