arXiv:2303.01192v1 [eess.AS] 2 Mar 2023

通过为注意力头分配辅助损失，改进基于 Transformer 的端到端说话人分离

Ye-Rin Jeoung、*Joon-Young Yang*、*Jeong-Hwan Choi* 和 *Joon-Hyuk Chang*

韩国首尔汉阳大学电子工程系

摘要

基于 Transformer 的端到端神经说话人分离（EEND）模型利用多头自注意力（SA）机制，实现对重叠语音区域中说话人标签的准确预测。本研究为提高 SA-EEND 模型的训练效果，提出为 Transformer 层中的 SA 头引入辅助损失。具体而言，我们假设，如果某个 SA 层的注意力权重矩阵与单位矩阵的模式相似，则这些矩阵存在冗余。随后，我们明确约束此类矩阵，使其呈现与语音活动检测或重叠语音检测任务相关的特定说话人活动模式。因此，我们期望所提出的辅助损失能够引导 Transformer 层呈现更多样化的注意力权重模式，从而减少 SA 头中假定存在的冗余。在用于双说话人分离任务的模拟数据集和 CALLHOME 数据集上，所提出方法的有效性得到验证；与传统 SA-EEND 模型相比，该方法分别将说话人日志错误率降低了 32.58% 和 17.11%。

索引词— 说话人日志、端到端神经说话人日志、多头注意力、辅助损失

1. 引言

说话人日志旨在确定“谁在何时说话”，可用于分析从电话对话 [1], 、访谈、正式会议 [2, 3], 和电视节目 [4]中获取的多说话人长时音频。传统上，说话人日志技术通过组合多个独立设计的模块来实现，例如语音活动检测（VAD）、重叠语音检测（OSD）、说话人嵌入提取和聚类。然而，这种方法存在局限性：其组成模块并非针对说话人日志任务进行联合优化，而是分别针对相关子任务进行优化。此外，重叠语音区域尤其难以处理，因为说话人嵌入提取器通常经过训练，只能嵌入单说话人信息。

为克服这些局限，研究者提出了端到端神经说话人日志（EEND）方法，能够以端到端的方式直接从一系列声学特征中预测说话人活动标签。具体而言，自注意力（SA）EEND模型 [5] 由多层Transformer层 [6] 堆叠而成，并输出帧级说话人后验概率。SA-EEND模型仅使用说话人标签与通过置换不变训练（PIT）方法从最后一层获得的预测之间的二元交叉熵（BCE）损失进行训练 [5]。然而，以这种方式训练的SA-EEND模型会导致较低层存在冗余（即较低层对目标任务的贡献并不显著），尤其是在Transformer层数增加时 [7]。为解决这一问题， [7] 提出添加残差连接以连接相邻的Transformer层，并对中间Transformer层应用辅助二元交叉熵损失。这种改进后的架构称为RX-EEND [7] ，与原始SA-EEND模型相比，性能得到了显著提升。

在本研究中，我们从不同于 [7]所述的角度出发，旨在提升 SA-EEND模型的性能。具体而言，我们关注 [5] 提出的观察结果：SA-EEND模型的注意力权重矩阵往往呈现出类似单位矩阵的模式，而我们认为这对于说话人日志任务是冗余的。因此，受近期一项研究 [8], 的启发，我们提出对注意力权重矩阵施加显式约束，使其呈现由与说话人日志相关的辅助任务导出的特定模式。在 [8], 中，为了促使一个 SA头在长时非语音区域内准确关注短语音话语，我们约束一个 SA头，使其呈现由真实语音活动检测标签导出的语音活动模式。与之不同的是，本研究聚焦于多说话人场景下的 VAD 和 OSD，二者对于实现良好的说话人日志性能都十分重要 [9–11]。我们使用从重叠语音话语的真实 VAD 或 OSD 标签中导出的模式，以辅助损失函数的形式约束 SA-EEND模型的注意力权重矩阵。由此，我们期望注意力权重矩阵呈现与目标任务相关的模式，而不是类似单位矩阵的模式；后者会为整个输入序列分配相等的权重。我们在双说话人日志任务上使用模拟数据集和真实数据集验证了所提方法的有效性。

2. SA-EEND综述

本节简要回顾作为基线系统的 SA-EEND [5] 模型。SA-EEND 模型根据一个长度为 T 的特征向量序列 $X = [\ _1 \cdots \ _T]$ $_t = [y_{t,1}, \cdots, y_{t,S}]^\top \in \mathbf{x},,\mathbf{x},$ $\{0,1\}^S$ S 预测说话人标签序列 $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_T]$。说话人标签 $\mathbf{y}$表示时间帧索引 t 处说话人的活动情况。经过线性层和层归一化（LN）处理后，[12], $\mathbf{x}_t$ 被转换为嵌入 $\mathbf{e}_t^0$，随后输入堆叠的 Transformer 编码器块。

$$\mathbf{e}_t^0 = \text{LN}(\text{Linear}(\mathbf{x}_t)) \ , \ \mathbf{e}_t^0 \in \mathbb{R}^D, \tag{1}$$

$$E^p = \text{EncoderBlock}_p^D(E^{p-1}), 1 \le p \le P, \tag{2}$$

其中，Linear(·) 表示一个线性层， p 索引编码器块， $E^p = [\mathbf{e}_1^p, \cdots, \mathbf{e}_T^p]$ 表示嵌入序列。每个 Transformer 编码器块由多头自注意力（MHSA）和全连接前馈网络组成。自注意力使用缩放点积 [6] 生成帧级注意力权重，以考虑全局特征关系。

$$Attention(Q, K, V) = \text{softmax}(QK^\top/\sqrt{d})V = AV, \tag{3}$$

其中， Q、 K和 $V \in \mathbb{R}^{T\times d}$ 分别表示查询、键和值 [6] d $A \in$ ；是隐 $\mathbb{R}^{T\times T}$ 藏空间的维度；并表示注意力权重矩阵。自注意力层的输出随后传入一个前馈网络，该网络由两个线性变换和 ReLU 激活函数 [13]组成。最后，经过所有编码器块后，在时间帧 t处， S个说话人的说话人后验概率 $\hat{\mathbf{y}}_t = [\hat{y}_{t,1}, \cdots, \hat{y}_{t,S}]$按如下方式得到：

$$\hat{y}_t = \text{sigmoid}(\text{Linear}(\text{LN}(\mathbf{e}_t^P))). \tag{4}$$

为了训练 SA-EEND 模型，预测值 $\hat{\mathbf{y}}_t$ 与真实标签$\mathbf{y}_t$ 之间的损失函数 $\mathcal{L}_d$ 计算如下：

$$H(y_{t,s}, \hat{y}_{t,s}) = -y_{t,s} \ln \hat{y}_{t,s} - (1 - y_{t,s}) \ln(1 - \hat{y}_{t,s}), \tag{5}$$

$$\mathcal{L}_d = \frac{1}{TS} \min_{\phi_1,\cdots,\phi_S \in \Phi_S} \sum_{t=1}^{T} \sum_{s=1}^{S} H(y_{t,s}^{\phi_s}, \hat{y}_{t,s}), \tag{6}$$

其中， $H(\cdot,\cdot)$ 表示二元交叉熵损失， Φ_S 表示所有可能的说话人排列，而$\mathbf{y}_{\phi_s} = [y_{1,s}^{\phi_s}, \cdots, y_{T,s}^{\phi_s}] \in \{0,1\}^T$是按照说话人排列 ϕ_s排列后的说话人标签序列。

3. 提出的辅助损失

3.1. 说话人级语音活动检测损失

为了帮助自注意力头学习不同说话人的语音活动模式，我们定义了一种新的辅助损失，称为说话人级语音活动检测（SVAD）损失。计算 SVAD 损失时，使用说话人标签序列$\mathbf{y}^{\phi_s}$生成目标掩码 $\mathcal{M}_s = \mathbf{y}_{\phi_s}^{\top}\mathbf{y}^{\phi_s}$ $(1 \le s \le S)$，并用它们约束注意力权重矩阵，如下所示：

$$\mathcal{L}_S = \sum_{s=1}^{S} \mathcal{L}(\mathcal{M}_s, A_s^h) = \sum_{s=1}^{S} \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} H(m_{ij}, a_{ij}), \tag{7}$$

其中，为 $\mathcal{M}_s$ 选取的注意力权重矩阵 A_s^h 由第 h 个自注意力头计算得出， m_{ij} 和 a_{ij} 分别表示 $\mathcal{M}_s$ 和 A_s^h 中第（i，j）个元素。由于说话人日志的目标是为每位说话人生成语音活动检测标签，我们期望目标掩码 $\mathcal{M}_s$ 的模式能够帮助自注意力头学习有效的注意力机制，避免产生类似身份的模式。

3.2. 重叠语音检测损失

为帮助自注意力头学习整体语音活动模式，同时区分重叠语音和单人说话区域，我们定义了一种名为重叠语音检测损失的辅助损失。重叠语音检测标签序列 $\boldsymbol{\psi} = [\psi_1, \cdots, \psi_T]$ 首先定义如下：

$$\psi_t = \begin{cases} 0 & \text{if } t \text{ is non-speech frame} \\ k & \text{if } t \text{ is single-speaker speech frame} \\ 1 & \text{if } t \text{ is overlapped speech frame} \end{cases}. \tag{8}$$

接着，与第 3.1 节所述类似，使用目标掩码 $\mathcal{M}_{OSD} = \boldsymbol{\psi}^{\top}\boldsymbol{\psi}$ 来定义重叠语音检测损失。

$$\mathcal{L}_O = \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} (n_{ij} - a_{ij})^2, \tag{9}$$

其中， n_{ij} 表示 $\mathcal{M}_{OSD}$的第（i，j）项， a_{ij} 是应用重叠语音检测损失的注意力权重矩阵中对应的项。通过这种方式，我们希望注意力权重能够隐式地区分单人说话区域和重叠语音区域，而不是赋予相同的权重，导致出现类似身份的模式。

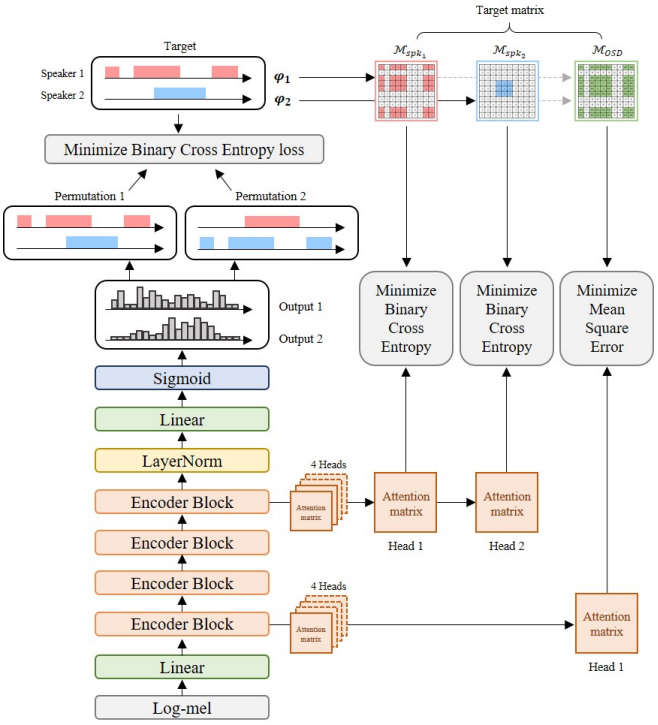

图 1。包含所提出的辅助损失、适用于两位说话人的 SA-EEND 模型

3.3. 使用所提出的辅助损失进行模型训练

由于我们认为单位矩阵是冗余的，因此选择注意力权重矩阵最接近单位矩阵模式的SA头来应用所提出的辅助损失。为实现此方法，我们计算每个注意力头的注意力权重矩阵的迹，并按降序排列这些值，以选出冗余的SA头。

说话人日志损失 $\mathcal{L}_d$ 与辅助损失 $\mathcal{L}_S$ 和 $\mathcal{L}_O$ 共同使用，以帮助每个注意力头不仅区分是否存在语音，还区分每位说话人的语音。最后，我们将所提出系统训练所用的总损失函数定义为

$$\mathcal{L}_{Total} = \mathcal{L}_d + \alpha\mathcal{L}_S + \beta\mathcal{L}_O, \tag{10}$$

其中， α 和 β 是用于表示各辅助损失应用程度的超参数。

4. 实验设置

4.1. 数据集

模拟数据集 Sim2spk 是按照 [5], 中描述的语音混合模拟算法创建的，重叠率为 34.4%。具体而言，Sim2spk 基于 Switchboard-2（第一、二、三阶段）、Switchboard Cellular（第一、二部分） [14], 以及 NIST说话人识别评测（2004年、2005年、2006年和2008年）语料库 [15–18], 生成，所有语料均以 8 kHz 采样。Sim2spk 中的所有话语均由两位不同的说话人组成，每位说话人的语音都是通过随机选取其 10 至 20 条话语生成的。随后 [5], , 每条话语均与随机选取的模拟房间脉冲响应进行卷积，

表 1. 双说话人数据集的统计信息

数据集	训练	适配	评估	
	Sim2spk	CH	Sim2spk	CH
# 混合语音	100,000	155	500 / 500 / 500	148
重叠率 (%)	34.4	14.0	34.4 / 27.3 / 19.6	13.1

卷积概率为 0.5，并加入取自 MUSAN 数据集 [19]的背景噪声样本。对于评估数据集，我们准备了模拟数据集和真实数据集。模拟数据集采用上述相同的方法创建，重叠率分别为 34.4%、27.3% 和 19.6%。对于真实评估数据集，我们使用了 CALLHOME（CH） [20] 数据集中的双人电话会话话语。

4.2. 训练与评估

采用具有四个 Transformer 编码器块的 SA-EEND模型 [5] 作为基线系统，其中每个编码器块均使用四个注意力头进行多头自注意力。输入特征为 23 维对数尺度梅尔滤波器组能量，使用 25 ms 的帧长和 10 ms 的帧移 [5]提取。将单帧特征与左侧 7 帧和右侧 7 帧进行拼接，并将从一个话语中提取的特征序列按 10 [5]的比例进行时间下采样。按照第 3 节所述，将提出的辅助损失应用于 SA-EEND模型。无需在公式（10）中确定 α和 β的最优值，在应用 $\mathcal{L}_S$时将 α设为 1，否则设为 0； β以相同方式设置。对于重叠语音检测损失，我们在公式（8）中设置$k = \sqrt{0.5}$，使 $\mathcal{M}_{OSD}$的对应项变为 0.5。EEND 模型使用 Sim2spk 数据集预训练 200 个周期，随后使用 CH 双说话人数据集适配 200 个周期。在预训练阶段，使用 Adam 优化器 [21]，预热步数为 100,000 [6]；在适配阶段，学习率设为 0.00001。最终的说话人活动预测使用 0.5 的阈值以及窗口大小为 11 帧的中值滤波获得。使用说话人日志错误率（DER） [22] 作为评估指标，并在每个片段的起始和结束处采用 0.25 秒的边界容差。

5. 结果与分析

5.1. 性能比较

表2展示了 SA-EEND [5], RX-EEND [7], 以及所提出的方法在说话人错误率方面的表现。对于 SA-EEND [5]和 RX-EEND [7]模型，表中既列出了文献 [5, 7] 报告的说话人错误率，也列出了我们自行实现后得到的结果，以供比较。为考察应用所提出的辅助损失的效果，我们将 $\mathcal{L}_S$ 或 $\mathcal{L}_O$ 应用于其中一个编码器块。由于训练数据集中有两位说话人，因此根据第3.3节所述的标准，将 $\mathcal{L}_S$ 应用于两个自注意力头，而将 $\mathcal{L}_O$ 应用于一个自注意力头。如表所示，在第4个编码器块应用 SVAD损失（$\mathcal{L}_S$）显著改善了两个数据集上的基线系统表现，但其他情况下的性能变化微乎其微。与 SVAD损失不同，OSD损失在不同编码器块上均改善了基线系统的表现，只有在第2个块上的表现有所下降，且两个数据集上均如此。有趣的是，SVAD损失和 OSD损失应用于第2个编码器块时，表现都最差。

表2. 不同端到端神经说话人日志方法在说话人错误率（%）方面的性能比较。 †: 我们的实现

方法	Sim2spk			Real
	ρ = 34.4	ρ = 27.3	ρ = 19.6	CH
SA-EEND [5]	7.91	8.51	9.51	13.66
SA-EEND†	6.30	6.14	6.17	10.46
RX-EEND [7]	4.18	3.93	4.01	9.17
RX-EEND†	4.28	3.76	3.86	9.04
采用 $\mathcal{L}_S$				
第 1 个编码器	5.81	5.67	5.58	10.33
第 2 个编码器	6.29	6.14	6.05	10.48
第 3 个编码器	5.43	5.27	5.27	10.05
第 4 个编码器	4.37	4.55	4.35	8.75
采用 $\mathcal{L}_O$				
第1个编码器	5.08	4.88	5.09	9.00
第2个编码器	7.10	7.31	7.48	10.51
第3个编码器	4.64	4.36	5.00	9.27
第4个编码器	4.57	4.45	4.54	8.93
$\mathcal{L}_S$ | $\mathcal{L}_O$				
4th Enc. 1st	**4.29**	**4.11**	**4.15**	**8.67**
4th Enc. 2nd	4.60	4.43	4.38	9.09
4th Enc. 3rd	4.48	4.19	4.35	8.81
4th Enc. 4th	4.47	4.2	4.41	8.81

根据模拟数据集的结果，我们在 SA-EEND模型中同时使用了两种辅助损失，如公式（10）所示，并采用了$\alpha = 1$ 和 $\beta = 1$。如表2最后一行所示，同时应用 SVAD损失和 OSD损失的表现通常优于仅应用其中一种辅助损失。两个数据集上的最佳表现均来自将 SVAD损失和 OSD损失分别应用于第4个和第1个编码器块的方案。我们的最佳模型在模拟和真实数据集上的表现也可与近期提出的 RX-EEND [7] 模型相媲美。

5.2. 自注意力头的可视化

为研究所提出的辅助损失对 SA-EEND模型 Transformer编码器块中多头自注意力的影响，图2对四个自注意力头的注意力权重矩阵（即公式（3）中的 A）进行了可视化。在该图中，注意力权重矩阵是在第4个编码器块中计算得到的，因为该块距离生成最终说话人后验概率的输出层最近。比较第一行与第三、第四和第五行可以发现，采用所提出的辅助损失后，与注意力头3和4对应的注意力权重矩阵的非对角元素呈现出更清晰的模式，并具有更大的权重。因此，结合表2所示的说话人错误率改进可以推断，通过应用所提出的辅助损失，使注意力权重矩阵中的类似身份的模式更加多样化，从而实现了对 SA-EEND模型更有效的训练。有趣的是，对于注意力头1和2，图2(c)和2(d)中观察到的模式与图2(a)中的模式不同。在图 2(e)中，该图对应于在第4个编码器块同时施加两种所提出辅助损失的 SA-EEND模型，其注意力权重矩阵模式与图2(c)中的模式相似。这表明，当所提出的重叠语音检测损失和 SVAD损失同时应用于同一个编码器块时，SA-EEND模型受 SVAD损失的影响大于重叠语音检测损失。最后，有趣的是

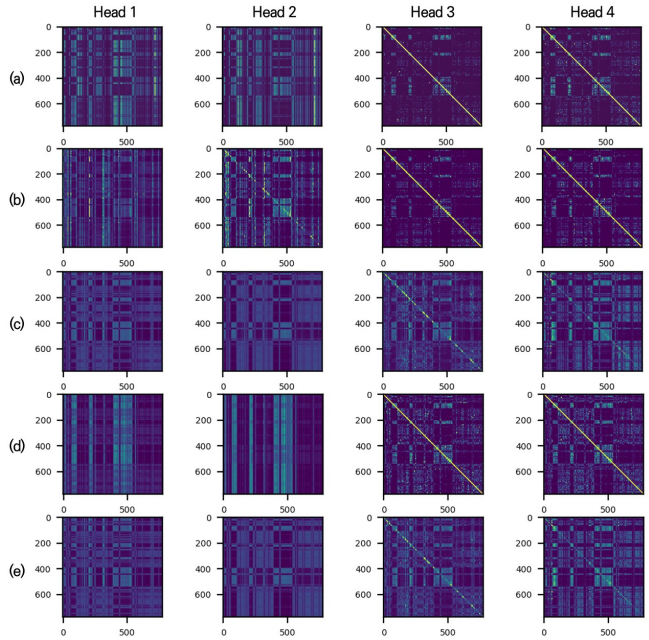

图2。第4个编码器块中注意力权重矩阵的可视化：(a) SA-EEND，(b) RX-EEND，(c) 使用 SVAD 损失的 SA-EEND，(d) 使用重叠语音检测损失的 SA-EEND，以及 (e) 在第4个编码器块上同时应用两种辅助损失的 SA-EEND。（CH 录音编号：iaai）

对于注意力头3和4，RX-EEND得到的模式（如图2(b)所示）与基线SA-EEND在图2(a)中得到的模式相似。这可能是因为RX-EEND的性能提升主要归因于应用于较低层编码器块以及最后一个编码器块的辅助二元交叉熵损失。为验证这一点，图3对各种 EEND模型在第2个编码器块中计算得到的注意力权重矩阵进行了可视化。值得注意的是，与其他模型（图3(a)和3(c)）相比，RX-EEND模型（图3(b)）中观察到的模式具有更大的权重，同时与 SA-EEND模型第4个编码器块中的模式（图2(b)）有些相似。有趣的是，对于图3所示的所有模型，两个自注意力头始终呈现类单位矩阵注意力权重矩阵，即对所有帧大多分配相等的权重。这或许可以解释表2所示的所提出辅助损失导致性能下降的情况：当 $\mathcal{L}_S$ 或 $\mathcal{L}_O$ 应用于 SA-EEND模型的第2个编码器块时，便会出现这种情况。

5.3. 自注意力头选择的消融研究

如第1节和第3.3节所述，我们假设与单位矩阵相似的注意力权重矩阵可能是冗余的，并基于这一假设选择应用所提出辅助损失的自注意力头。为验证上述注意力头选择方法，我们使用所提出的辅助损失训练SA-EEND模型，同时随机选择自注意力头进行比较。对于随机选择注意力头，我们直接将所提出的SVAD损失或OSD损失分别应用于第一个或前两个自注意力头，而不考虑注意力权重矩阵与单位矩阵之间的相似性。表3比较了在CH数据集上采用随机选择注意力头和第3.3节所述方法所得的结果。结果表明，基于单位矩阵选择应用所提出辅助损失的自注意力头是有效的，尤其是在采用

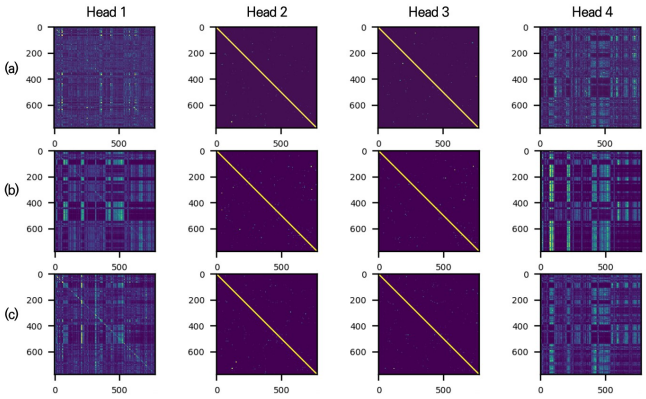

图3。第2个编码器块中注意力权重矩阵的可视化：(a) SA-EEND，(b) RX-EEND，(c) 在第4个编码器块上同时应用两种辅助损失的 SA-EEND。（CH 录音编号：iaai）

表3。不同自注意力头选择方法在 CALLHOME 上的说话人错误率（%）。 $\mathcal{L}_s$ 和 $\mathcal{L}_o$ 应用于第4个编码器块。

Loss		**选择方法**	
$\mathcal{L}_s$	$\mathcal{L}_o$	随机矩阵	单位矩阵
✓		9.30	**8.75**
	✓	9.00	**8.93**
✓	✓	9.22	**8.81**

基于矩阵选择应用所提出的辅助损失的自注意力头是有效的，尤其是在采用 SVAD损失时。

6. 结论

本研究提出使用辅助损失训练 SA-EEND 模型，以便更好地利用其注意力权重矩阵模式类似于单位矩阵的自注意力头。我们利用与语音活动检测（VAD）或重叠语音检测（OSD）相关的语音活动标签定义辅助损失，这两者都可视为说话人日志的重要子任务。为验证这一思路，我们将所提出的辅助损失应用于注意力权重矩阵模式类似于单位矩阵的自注意力头。在双说话人日志任务上的实验结果表明，在模拟和真实数据条件下，所提出的 SVAD 和 OSD损失均能显著提升传统 SA-EEND 模型的性能。对不同 EEND变体的注意力权重矩阵进行可视化和比较，支持了我们的假设与分析。

7. 致谢

本研究获韩国政府（科学技术信息通信部）资助的韩国信息通信技术规划与评估院（IITP）项目支持（编号：2021-0-00456，远程多说话人会议系统超高清语音质量技术开发）

8. 参考文献

[1] P. Kenny、D. Reynolds 和 F. Castaldo，“使用因子分析对电话对话进行说话人分离”，*IEEE*信号处理精选主题期刊，第4卷，第6期，第1059–1070页，2010年。[2] A. Janin、D. Baron、J. Edwards、D. Ellis、D. Gelbart、N. Morgan、B. Peskin、T. Pfau、E. Shriberg、A. Stolcke 等，“ICSI会议语料库”，载于*IEEE*国际声学、语音与信号处理会议论文集，2003年，第364–367页。[3] S. Renals、T. Hain 和 H. Bourlard，“AMI和AMIDA项目中的多方会议解读”，载于*IEEE*免提语音通信与麦克风阵列会议论文集，2008年，第115–118页。[4] F. Vallet、S. Essid 和 J. Carrive，“电视脱口秀说话人分离的多模态方法”，*IEEE*多媒体汇刊，第15卷，第3期，第509–520页，2012年。[5] 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离”，载于*IEEE*自动语音识别与理解研讨会论文集，2019年，第296–303页。[6] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N Gomez、L. Kaiser 和 I. Polosukhin，“Attention is all you need”，载于神经信息处理系统进展论文集，2017年。[7] Y. Yu、D. Park 和 H. K. Kim，“用于端到端说话人分离的残差连接Transformer辅助损失”，载于*IEEE*国际声学、语音与信号处理会议论文集，2022年，第8377–8381页。[8] M. H. Lee、J. H. Chang、S. E. Lee、J. S. Seong、C. Park 和 H. Kwon，“通过惩罚注意力权重对Transformer声学模型进行正则化”，载于*INTERSPEECH*论文集，2022年，第56–60页。[9] H. Bredin、R. Yin、J. M. Coria、G. Gelly、P. Korshunov、M. Lavechin、D. Fustes、H. Titeux、W. Bouaziz 和 M. P. Gill，“Pyannote.audio：用于说话人分离的神经网络构建模块”，载于*IEEE*国际声学、语音与信号处理会议论文集，2020年，第7124–7128页。[10] Y. Takashima、Y. Fujita、S. Watanabe、S. Horiguchi、P. Garc´ıa 和 K. Nagamatsu，“基于语音活动和重叠检测的端到端说话人分离”，载于*IEEE*口语语言技术研讨会论文集，2021年，第849–856页。[11] H. Bredin 和 A. Laurent，“用于重叠感知重分段的端到端说话人分段”，载于*INTERSPEECH*论文集，2021年，第3111–3115页。[12] J. L. Ba、J. R. Kiros 和 G. E Hinton，“层归一化”，*arXiv:1607.06450*，2016年。[13] A. F. Agarap，“使用修正线性单元（ReLU）进行深度学习”，*arXiv:1803.08375*，2018年。[14] J. J Godfrey、E. C Holliman 和 J. McDaniel，“SWITCHBOARD：用于研究与开发的电话语音语料库”，载于*IEEE*国际声学、语音与信号处理会议论文集，1992年，第517–520页。[15] M. P. Alvin 和 A. Martin，“NIST说话人识别评估纪事”，载于*Odyssey*论文集，2004年。

[16] A. Martin、M. Przybocki 和 J. P Campbell，“NIST说话人识别评估项目”，载于生物识别系统，第 241–262 页。2005年。[17] M. A Przybocki、A. F Martin 和 A. N Le，“NIST说话人识别评估纪事——第 2 部分”，载于*Proc. Odyssey*，2006年，第 1–6 页。[18] A. F Martin 和 C. S Greenberg，“NIST 2008年说话人识别评估：电话与室内麦克风通道上的性能”，载于*Proc. INTERSPEECH*，2009年，第 2579–2582 页。[19] D. Snyder、G. Chen 和 D. Povey，“Musan：音乐、语音和噪声语料库”，*arXiv:1510.08484*，2015年。[20] A. F Martin 和 M. A Przybocki，“2000年 NIST 说话人识别评估”，载于费城：语言数据联盟，2001年。[21] D. P Kingma 和 J. Ba，“Adam：一种随机优化方法”，*arXiv:1412.6980*，2014年。[22] J. G Fiscus、J. Ajot 和 J. S Garofolo，“2007年富转录会议识别评估”，载于用于人类感知的多模态技术，第 373–389 页。2007年。