arXiv:2601.02688v1 [cs.SD] 6 Jan 2026

用于语音识别的多通道多说话人Transformer

郭一凡[1], 田尧[1],索红斌[1],万玉龙[1]

[1]数据与人工智能工程系统，OPPO，中国北京{

guoyifan, aaron1, suohongbin, wanyulong}@oppo.com

摘要

随着电话会议和车载语音助手的发展，远场多说话人语音识别已成为一个热门研究课题。近期，有研究提出了多通道Transformer（MCT），展现了Transformer对远场声学环境建模的能力。然而，由于说话人之间的干扰，MCT无法从混合输入音频中为每个说话人编码高维声学特征。基于此，本文提出用于远场多说话人自动语音识别的多通道多说话人Transformer（M2Former）。在SMS-WSJ基准上的实验表明，与神经波束形成器、MCT、采用变换-平均-拼接的双路径循环神经网络，以及基于多通道深度聚类的端到端系统相比，M2Former的相对词错误率分别降低了9.2%、14.3%、24.9%和52.2%。

索引词：多通道自动语音识别、多说话人自动语音识别、Transformer

1. 引言

随着电话会议和车载语音助手日益普及，远场多说话人语音识别已成为热门研究课题。目前主流方法基于置换不变训练（PIT） [1]。具体而言，这些方法首先使用远场语音分离前端[2–8]为每位说话人输出单通道语音特征，然后通过采用PIT的单通道自动语音识别（ASR）后端对这些特征进行解码。与基于串行输出训练 [9] 的方法相比，基于PIT的方法解码速度更快，也更接近实际应用。

最近，研究人员 [10, 11] 提出了一种用于远场语音识别的多通道Transformer（MCT）。实验表明，MCT优于常用的系统 [12–14]；后者使用增强模型作为前端，使用单通道ASR模型作为后端。这主要是因为前端与后端在功能设计上的不一致，限制了后者系统的性能。受此启发，我们尝试绕过“分离前端—识别后端”（Sep-ASR） [2–8] 范式，并提出一种多通道多说话人Transformer（M2Former），其编码器能够直接从混合输入音频中为每位说话人编码高维声学嵌入。

然而，在多说话人场景中使用 MCT 存在一些问题。MCT 的编码器利用通道内连续性信息和跨通道空间信息，对远场环境下的上下文关系进行编码。在利用跨通道信息时，MCT 通过可学习权重融合各通道，这些权重旨在隐式地学习建模空间关系。但是在多说话人情况下，说话人之间的干扰和噪声使编码器难以为每位说话人生成高质量的声学嵌入。

为减轻此类干扰，我们使用二维卷积神经网络（二维卷积神经网络）以及一种改进的多通道注意力机制，称为多通道多说话人注意力（M2A）。二维卷积神经网络广泛用于图像任务和语音增强任务 [15–19]，用作特征解耦模块。这是因为它们能够通过不同的滤波器通道学习多种方面 [15, 17, 20, 21]。例如，研究人员 [15] 发现，卷积神经网络的不同滤波器会关注来自不同方向的信号。基于这一点，我们首先使用二维卷积神经网络对多通道输入进行解耦。由于信号源之间总是存在物理或内容差异，因此卷积神经网络的每个输出通道都应有很高的概率仅包含与一个说话人（或噪声）相对应的信息。随后，我们提出 M2A 模块来编码上下文关系。与 MCT 使用所有通道不同，M2A 仅利用相似度较高的通道之间的跨通道信息。因此，对应不同声源的信号会被分别编码，从而在一定程度上避免上述干扰。

此外，在从输入混合信号中分离说话人特定特征时，我们采用了一种基于谱聚类 [22] 的方法。与基于投影的方法 [4–6, 8]相比，后者使用线性投影层为说话人预测掩码，前者有两项优势。1）聚类后，对应不同声源的通道会被分配到不同的簇。在每个簇内使用 M2A，可以更好地避免声源之间的干扰。2）无需训练额外的线性层，因此更容易应用于说话人数量更多且不固定的情形。此外，聚类所需的相似度矩阵可以通过 M2A 轻松获得。除此之外，我们还采用了一种 [23] 区分噪声和语音的方法。丢弃噪声簇后，背景噪声便会被滤除。

基于上述方法，我们提出了用于远场多说话人自动语音识别的 M2Former。在 SMS-WSJ [24]上的实验表明，与神经波束形成器 [4],MCT、带变换-平均-拼接的双路径循环神经网络 [6]以及基于多通道深度聚类 [3] 的系统相比，M2Former 的相对词错误率降低分别为 9.2%、14.3%、24.9% 和 52.2%。

2. 提出的方法

本节介绍所提出的 M2Former框架，如图 1 所示。

M2Former 的解码器结构和功能与原始语音Transformer [25, 26]中的解码器一致。它将每位说话人的单通道声学嵌入分别解码为文本。

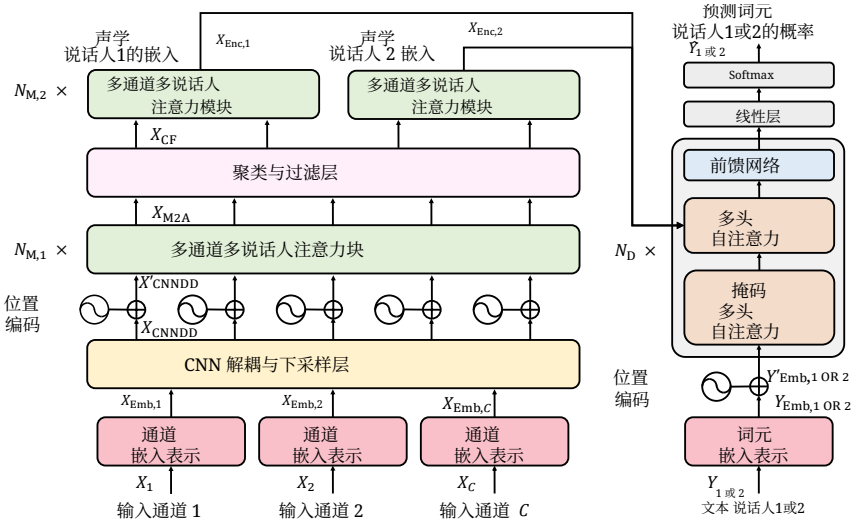

图 1： 多通道多说话人*Transformer*。 C 为输入通道数。 N_M 和 N_D 分别表示 *M2A* 模块和解码器模块的数量。假设输入音频中有两位说话人。

因此，我们将重点介绍编码器的各个组成部分，它是 M2Former 的核心。

2.1. 通道嵌入

为了获得适用于网络学习的输入特征，我们在编码器前添加通道嵌入层，紧接在 [10]之后。具体来说，我们使用短时傅里叶变换（STFT）的幅度 X^{mag} 和相位特征 X^{pha} 作为输入特征 $X \in \mathbb{R}^{C\times T\times 3F}$。然后，我们使用线性投影层获得嵌入$X_{\text{Emb}} \in \mathbb{R}^{C\times T\times D}$ （为简便起见，忽略偏置）：

$$X_{\text{Emb},i} = [X_i^{\text{mag}} W^{\text{mag}}, X_i^{\text{pha}} W^{\text{pha}}] \cdot W^{\text{Emb}} \tag{1}$$

这里， X_i 表示 X 的第 i个通道，$X_{\text{Emb},\ i}$ 表示 X_{Emb}的第i个通道， W表示线性投影层，$[\ldots,\ldots]$ 表示特征拼接。

2.2. 卷积神经网络解耦与下采样层

受 [15–19], 启发，我们采用二维卷积神经网络模块，将输入的多通道特征 X_{Emb} 解耦为 X_{CNNDD}，以获得更多通道和更具判别性的表示。此外，受频谱稀疏分布的启发，我们使用卷积神经网络在频率和时间维度上对输入特征进行下采样，以提高计算效率。因此，我们得到 $X_{\text{CNNDD}} \in \mathbb{R}^{C'\times T'\times D'}$，其中 $C' > C$、 $T' < T$ 和 $D' < D$。

2.3. 多通道多说话人注意力模块

为了利用通道内和跨通道信息，直接编码每个说话人的上下文关系，我们提出了 M2A。遵循 [10, 11], , M2A 由通道内注意力层（图2a）和跨通道注意力层（图2c）组成。

如引言中所述，卷积神经网络的每个输出通道都应有较高概率仅包含与一个说话人（或噪声）对应的信息。因此，我们可以首先像常用的基于 Transformer 的单通道自动语音识别编码器那样，在每个通道中使用自注意力 [10, 11, 26]（图2a）。它可以学习通道内的上下文连续性信息。

然后，我们提出跨通道注意力层，以利用帧间的连续性信息以及通道间的空间信息。如图2b所示，MCT [10,11] 使用可学习权重 P组合通道（为简洁起见，我们不介绍 P的具体设计）。这种方式会导致属于不同说话人的通道相互干扰。这是因为通道与声源之间的对应关系存在许多排列方式，难以用固定参数进行建模。为避免这一问题，我们提出了 M2A 跨通道注意力（图2c）。令 $X_c, X'_c \in \mathbb{R}^{T'\times D'}$ 分别表示跨通道注意力层第 c 个通道的输入和输出特征。然后， X' 的计算方式如下：

$$\begin{aligned}
\hat{X}_c &= \sum_i z_{ci} \cdot X_i \\
Q_c &= X_c W^q + \mathbf{1} \cdot (\mathbf{b}^q)^{\mathrm{T}} \\
K_c &= \hat{X}_c W^k + \mathbf{1} \cdot (\mathbf{b}^k)^{\mathrm{T}} \\
V_c &= \hat{X}_c W^v + \mathbf{1} \cdot (\mathbf{b}^v)^{\mathrm{T}} \\
X'_c &= \text{softmax}\left(Q_c K_c^{\mathrm{T}} / \sqrt{d_k}\right) V_c
\end{aligned} \tag{2}$$

其中 W^*, $\mathbf{b}^*$ 是可学习权重和偏置， d_k 是缩放因子。为简洁起见，此处省略多头机制。 z_{ci} 是通道间相似度矩阵 $Z \in \mathbb{R}^{C'\times C'}$ 中第 c 行、第 i 列的元素，它反映通道 c 与 i之间的相似度：

$$Z = \text{softmax}\left(1/T \sum\nolimits_t X_t X_t^{\mathrm{T}} / \sqrt{d_k}\right) \tag{3}$$

由于不同声源的信号通常互不相关，基于相似度进行组合可以减轻上述干扰。在编码器的最后一个 M2A 模块之后（图 1），我们对对应于同一说话人的通道取平均，从而为每位说话人获得单通道输出。

2.4. 聚类与滤波层

聚类与过滤（CF）层如图 3 所示。首先，我们使用最近的 M2A 模块生成的通道间相似度矩阵 Z 进行谱聚类 [22], 为每个通道分配一个标签。之后，我们计算每个输出通道的帧间相似度差（IFSD）值 [23]

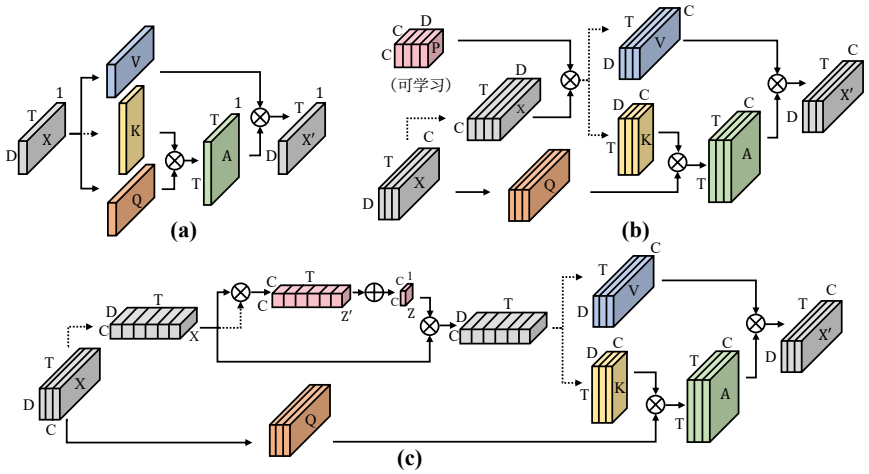

图2：多通道　　注意力块：*(a)*通道内注意力。*(b) MCT* 的跨通道注意力。*(c) M2A* 的跨通道　　*nel*
注意力。 ⊗, ⊕*an*　虚线分别表示矩阵乘法、沿时间维度求平均和矩阵转置。

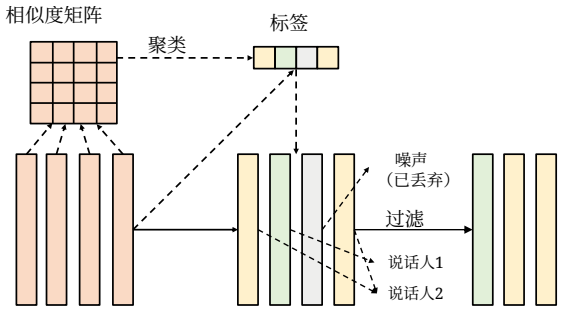

图 3：聚类与滤波层。梯度仅沿实线传播。假设有 *4* 个通道和 *2* 个说话人，灰色标签的 *IFSD* 值最低。

，并对属于同一标签的值求平均：

$$\mathrm{IFSD}\,(X) = \frac{1}{T}\sum_{t}\left[\hat{\mathbf{x}}_t^{\mathrm{T}}\cdot\hat{\mathbf{x}}_{t+1} - \alpha\cdot\hat{\mathbf{x}}_t^{\mathrm{T}}\cdot\hat{\mathbf{x}}_{t+\tau}\right] \qquad (4)$$

其中，$X\in\mathbb{R}^{D\times T}$是输入特征，$\hat{}\mathbf{x}_t\in\mathbb{R}^{D}$是 X中第 t 帧的归一化结果，α 和 τ 是超参数。最后，我们保留 IFSD 值最高的 n 个标签，并丢弃其余标签（n 是说话人数量）。其依据是，IFSD 值表示通道由语音而非噪声主导的概率。因此，保留的标签对应语音源，其余标签则对应噪声。

如图 1 所示，我们还尝试在属于同一聚类的通道之间使用 M2A，结果更好。

3. 实验

3.1. 数据集

为了评估我们提出的模型，我们在 SMS-WSJ [24] 数据集上进行了多项实验。SMS-WSJ 是一个六通道多说话人远场数据集。SMS-WSJ 包含 2、3 和 4 个说话人同时发言的三个版本，分别记为 SMS-2、SMS-3 和 SMS-4。本文主要在 SMS-2 上进行实验，另外两个数据集仅用于验证。SMS-2 的训练集、验证集和测试集分别包含 33561、982 和 1332 条话语，总时长为 93.3 小时。SMS-3 和 SMS-4 共用与 SMS-2 相同的原始数据划分配置。

3.2. 基线系统与所提模型的配置

3.2.1. 基线系统的分离前端

为展示我们方法的性能，我们在基线系统中采用了三种具有代表性的多通道分离系统作为前端。我们使用帧长为 25 ms、帧移为 10 ms 的 STFT 特征作为原始特征。第一个输入通道被用作参考通道。

多通道深度聚类（MC-DPCL） [3]：MC-DPCL 使用具有 300 个隐藏单元的 3 层双向长短期记忆网络作为嵌入提取器。它将由 STFT 特征生成的幅度特征和相位特征作为输入。为了使其能够进行端到端训练，我们采用了 [27] 中的方法使用软 k 均值。

带变换-平均-拼接的双路径循环神经网络（DPRNN-TAC） [6]：该系统使用 6 个双路径循环神经网络模块，每个模块具有 128 个隐藏单元。DPRNN-TAC 的特征维度和分段大小分别为 64 和 24。它直接将时域信号作为输入。

神经波束形成器（NB） [4]：我们使用一个包含 3 层、每层有 300 个隐藏单元的双向长短期记忆网络，并采用两个线性投影层，分别作为两个说话人的掩码估计器。

3.2.2. 基线系统的自动语音识别后端

基线系统的自动语音识别后端是一个统一的、基于 Transformer 的单通道 AED 模型 [26]，该模型使用 PIT 与前端系统进行端到端训练。编码器包含 12 个模块，解码器包含 6 个模块；多头注意力有 4 个头，注意力维度为 256；位置前馈网络有 1024 个单元。

3.2.3. 所提出的模型

CNNDD中的二维卷积神经网络模块包含八个卷积神经网络层，其输出通道数分别为 6、6、10、10、20、20、40、40。所有卷积神经网络的卷积核尺寸均为 3×3，且两个维度上都采用零填充。第一层和第二层的步幅分别为 [2,2]和 [2,1],，其余各层的步幅均为 1。根据验证集上的结果，IFSD 的 α 设为 5.3。为了使 M2Former 的规模相当，我们将 $N_{\mathrm{M},1}$ 与 $N_{\mathrm{M},2}$ 之和设为 6，并将 N_{D} 设为 6；注意力维度设为 256，多头注意力的头数设为 4，前馈单元设为 1024。

表1：所提出模型和基线模型实验的词错误率（*WER%*）。

SMS-2	cv_dev93	test_eval92	MEAN
MC-DPCL [3]	34.7	39.6	37.2
DPRNN-TAC [6]	22.3	25.1	23.7
NB [4]	18.4	20.7	19.6
M2Former	**16.6**	**18.9**	**17.8**

表2：*SMS-2* 上的消融实验（字错误率%）。*[-]* 表示实验中移除了该模块）。

SMS-2	cv_dev93	test_eval92	MEAN
[-] CNNDD	35.5	40.1	37.8
[-] M2A$_1$	24.2	28.7	26.5
[-] M2A$_2$	18.3	21.0	19.7
[-] IFSD	17.9	20.8	19.4
完整模型	**16.6**	**18.9**	**17.8**

3.2.4. M2Former 与基线系统的损失函数

根据 [28],，损失函数 $\mathcal{L}$ 由编码器的 CTC损失 [29] 和解码器的注意力损失组成。CTC损失用于通过 PIT 确定最佳排列。

$$\mathcal{L} = \sum_i [\lambda \mathcal{L}_{\text{ctc}}(X_{\text{Enc},i}, Y_{\hat{\pi}(i)}) + (1-\lambda)\mathcal{L}_{\text{att}}(\hat{Y}_i, Y_{\hat{\pi}(i)})]$$
$$\hat{\pi} = \text{argmin}_{\pi \in \mathcal{P}}(\sum_i \mathcal{L}_{\text{ctc}}(X_{\text{Enc},i}, Y_{\pi(i)})) \quad (5)$$

其中，$i = 1, 2, \ldots\ldots N$表示说话人标识，而 $\mathcal{P}$，Y，$\hat{Y}$,X_{Enc} 分别表示所有可能的排列、真实词元、预测词元和编码器输出。

3.3. 实验结果

*3.3.1. M2Former*与基线系统的比较

我们在SMS-2数据集上进行实验，比较M2Former与基线系统的性能（表1）。我们提出的模型在SMS-2数据集上取得了最佳结果。具体而言，就相对词错误率降低幅度而言，它分别比基于NB、DPRNN-TAC和MC-DPCL的端到端系统高出9.2%、24.9%和52.2%。

3.3.2. 编码器各模块的消融实验

我们进行了若干消融实验，以验证编码器中各个模块的有效性（表2）。在 CNNDD 的实验中，我们将原始卷积神经网络替换为两个3×3 2D 卷积神经网络，其通道维度为 40，步幅分别为 [2,2] 和 [2,1]。对于 M2A$_1$, 实验，我们仅将 $N_{\text{M设为2}}$ 6，并将 $N_{\text{M设为1}}$ 0。M2A$_2$ 实验的设置方法与 M2A$_1$，类似，不同之处在于我们在 CF 层后添加了一个线性层，以平滑 CF 层引入的失真。在 IFSD 实验中，我们将簇数设为 2，并且在 CF 层中不丢弃任何通道。

我们可以发现，CNNDD 对整体性能的贡献最大，因为它能够将输入解耦为对应不同方面的输出。M2A$_1$ 和 M2A$_2$ 也很有用。M2A$_1$ 能够对声学嵌入进行编码，而不受其他说话人的干扰。M2A$_2$ 则能够在编码时消除 CF 导致的失真以及可能存在的残余噪声的影响。IFSD 方法也有助于噪声消除。

表3：*M2A* 和 *MC-SMS-2* 实验的字错误率（*WER，%*）。

SMS-2	cv_dev93	test_eval92	MEAN
s-CNN + MCT	34.8	38.1	36.5
s-CNN + STA [30,31]	31.4	35.6	33.6
s-CNN + M2A$_1$	40.2	47.7	44.0
CNNDD +MCT	21.2	24.7	23.0
CNNDD +STA	19.7	22.8	21.3
卷积神经网络 +M2A$_1$	18.3	21.0	19.7

表4：*M2Former* 在 *SMS-WSJ* 数据集上针对不同说话人数量的性能（字错误率%）。第二列表示推理时是否预先知道说话人数量。

	已知	cv_dev93	test_eval92	MEAN
SMS-2	✓	16.6	18.9	17.8
	✗	16.9	19.3	18.1
SMS-3	✓	28.5	32.7	30.6
	✗	30.8	34.5	32.7
SMS-4	✓	33.4	37.1	35.3
	✗	37.9	41.8	39.9

3.3.3. M2A 与 *MCT* 的实验比较

为了进一步探究 M2A 的性能，我们进行了实验，将 M2A 与 MCT 和时空注意力（STA） [30,31] 进行比较（表3）。这三个模型都在编码器中使用了 6 个注意力块。由于 MCT 只能用于输入通道数固定的情况，因此我们没有在 CF 层之后使用注意力块。我们还考虑了将 CNNDD 替换为 3.3.2 节中提到的单一 CNN（s-CNN）的情况。

比较表3中的第1行与第4行（或第2行与第5行、第3行与第6行），再次证明了 CNNDD 的重要性。比较第1、2、3行可以发现，如果没有 CNNDD 的解耦，M2A 更容易受到各通道内声源间干扰的影响，而 MCT 和 STA 则可以隐式地学习一些说话人相关信息。比较第4、5、6行可以发现，CNNDD-M2A 能更好地避免不同声源对应通道之间的干扰，相较于 CNNDD-MCT，词错误率降低了 14.3%。

3.3.4. 不同说话人数下的性能 *ers*

我们进行了一系列实验，以探索 M2Former 在不同说话人数下的性能（表4）。当说话人数未知时，使用特征间隙启发式方法 [32] 来确定说话人数。

通过比较同一数据集中的每两行，我们可以发现，利用解耦通道能够相对准确地估计说话人的数量。这表明模型已经学会根据声源将特征解耦到不同通道。此外，尽管模型未使用超过两位说话人的数据进行训练，但在 SMS-3 和 SMS-4 上仍能取得不错的效果。这可能表明，该模型可以轻松应用于说话人数量更多的场景。

4. 结论

我们提出了一种用于语音识别的端到端多通道多说话人 Transformer。通过采用CNN解耦与下采样层、多通道多说话人注意力块以及聚类与过滤层，编码器能够直接从混合输入中编码每位说话人的声学特征。实验结果表明，所提出的模型优于分离-识别形式基线和MCT。

5. 参考文献

[1] D. Yu、M. Kolbæk、Z.-H. Tan 和 J. Jensen，“用于说话人无关多说话人语音分离的深度模型置换不变训练”，发表于 *2017 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2017 年，第 241–245 页。[2] L. Drude 和 R. Haeb-Umbach，“利用深度聚类嵌入对盲源分离中的空间特征与频谱特征进行紧密整合”。发表于*Interspeech*，2017 年，第 2650–2654 页。[3] 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”，发表于 *2018 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2018 年，第 1–5 页。[4] X. Chang、W. Zhang、Y. Qian、J. Le Roux 和 S. Watanabe，“MIMO-Speech：端到端多通道多说话人语音识别”，发表于*2019 IEEE* 自动语音识别与理解研讨会 *(ASRU)*。IEEE，2019 年，第 237–244 页。[5] Y. Luo 和 N. Mesgarani，“Conv-TasNet：超越语音分离的理想时频幅度掩蔽”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 27 卷，第 8 期，第 1256–1266 页，2019 年。[6] Y. Luo、Z. Chen、N. Mesgarani 和 T. Yoshioka，“端到端麦克风置换与数量不变的多通道语音分离”，发表于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2020 年，第 6394–6398 页。[7] R. Gu、S.-X. Zhang、L. Chen、Y. Xu、M. Yu、D. Su、Y. Zou 和 D. Yu，“通过空间特征学习增强端到端多通道语音分离”，发表于 *ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2020 年，第 7319–7323 页。[8] C. Subakan、M. Ravanelli、S. Cornell、M. Bronzi 和 J. Zhong，“语音分离只需注意力”，发表于*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2021 年，第 21–25 页。[9] N. Kanda、Y. Gaur、X. Wang、Z. Meng 和 T. Yoshioka，“用于端到端重叠语音识别的序列化输出训练”，*arXiv* 预印本 *arXiv:2003.12687*，2020 年。[10] F.-J. Chang、M. Radfar、A. Mouchtaris、B. King 和 S. Kunzmann，“用于语音识别的端到端多通道Transformer”，发表于*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2021 年，第 5884–5888 页。[11] F.-J. Chang、M. Radfar、A. Mouchtaris 和 M. Omologo，“用于语音识别的多通道Transformer传导器”，*arXiv* 预印本 *arXiv:2108.12953*，2021 年。[12] J. Heymann、L. Drude 和 R. Haeb-Umbach，“用于声学波束形成的基于神经网络的频谱掩蔽估计”，发表于*2016 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2016 年，第 196–200 页。[13] T. Ochiai、S. Watanabe、T. Hori 和 J. R. Hershey，“多通道端到端语音识别”，发表于国际机器学习会议。PMLR，2017 年，第 2632–2641 页。[14] K. Kumatani、W. Minhua、S. Sundaram、N. Str¨om 和 B. Hoffmeister，“用于远场语音识别的多几何空间声学建模”，发表于*ICASSP 2019-2019 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2019 年，第 6635–6639 页。[15] T. N. Sainath、R. J. Weiss、K. W. Wilson、A. Narayanan、M. Bacchiani 和 A. Senior，“基于原始多通道波形的说话人位置与麦克风间距不变声学建模”，发表于*2015 IEEE* 自动语音识别与理解研讨会 *(ASRU)*。IEEE，2015 年，第 30–36 页。[16] Y. Qian、M. Bi、T. Tan 和 K. Yu，“用于噪声鲁棒语音识别的超深卷积神经网络”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 24 卷，第 12 期

，第 2263 页–2276，2016。

[17] S.-W. Fu、T.-W. Wang、Y. Tsao、X. Lu 和 H. Kawai，“基于全卷积神经网络优化直接评估指标的端到端波形语句增强”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 26 卷，第 9 期，第 1570–1584 页，2018 年。[18] Y. Hu、Y. Liu、S. Lv、M. Xing、S. Zhang、Y. Fu、J. Wu、B. Zhang 和 L. Xie，“DCCRN：用于相位感知语音增强的深度复值卷积循环网络”，*arXiv* 预印本 *arXiv:2008.00264*，2020 年。[19] T. Park、K. Kumatani、M. Wu 和 S. Sundaram，“使用频率对齐网络的鲁棒多通道语音识别”，载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020 年，第 6859–6863 页。[20] D. Wei、B. Zhou、A. Torrabla 和 W. Freeman，“理解 CNN 内部的类内知识”，*arXiv* 预印本 *arXiv:1507.02379*，2015 年。[21] A. Mahendran 和 A. Vedaldi，“通过反演理解深度图像表征”，载于*IEEE*计算机视觉与模式识别会议论文集，2015 年，第 5188–5196 页。[22] A. Ng、M. Jordan 和 Y. Weiss，“关于谱聚类：分析与算法”，神经信息处理系统进展，第 14 卷，2001 年。[23] Y. Guo、Y. Chen、G. Cheng、P. Zhang 和 Y. Yan，“基于复值神经网络和帧间相似度差分法的远场语音识别”，载于*2021 IEEE*自动语音识别与理解研讨会（*ASRU*）。IEEE，2021 年，第 1003–1010 页。[24] L. Drude、J. Heitkaemper、C. Boeddeker 和 R. Haeb-Umbach，“SMS-WSJ：用于多通道源分离与识别的数据库、性能度量及基线方案”，*arXiv* 预印本 *arXiv:1910.13934*，2019 年。[25] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，“注意力就是你所需要的一切”，神经信息处理系统进展，第 30 卷，2017 年。[26] L. Dong、S. Xu 和 B. Xu，“Speech-Transformer：一种用于语音识别的无循环序列到序列模型”，载于*2018 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018 年，第 5884–5888 页。[27] Y.-J. Lu、X. Chang、C. Li、W. Zhang、S. Cornell、Z. Ni、Y. Masuyama、B. Yan、R. Scheibler、Z.-Q. Wang等，“Espnet-se++：用于鲁棒语音识别、翻译和理解的语音增强”，*arXiv* 预印本 *arXiv:2207.09514*，2022 年。[28] S. Watanabe、T. Hori、S. Kim、J. R. Hershey 和 T. Hayashi，“用于端到端语音识别的混合 CTC/注意力架构”，*IEEE*信号处理精选主题期刊，第 11 卷，第 8 期，第 1240–1253 页，2017 年。[29] A. Graves、S. Fern´andez、F. Gomez 和 J. Schmidhuber，“连接时序分类：使用循环神经网络标注未分段序列数据”，载于第 *23* 届国际机器学习会议论文集，2006 年，第 369–376 页。[30] D. Wang、T. Yoshioka、Z. Chen、X. Wang、T. Zhou 和 Z. Meng，“使用自组织麦克风阵列进行连续语音分离”，载于*2021*年第*29*届欧洲信号处理会议（*EUSIPCO*）。IEEE，2021 年，第 1100–1104 页。[31] S. Horiguchi、Y. Takashima、P. Garcia、S. Watanabe 和 Y. Kawaguchi，“使用分布式麦克风的多通道端到端神经说话人日志”，载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022 年，第 7332–7336 页。[32] M. Bolla，多重图谱性质与分类性质之间的关系。DIMACS，离散数学与理论计算机科学中心，1991 年。