arXiv:2303.00692v1 [eess.AS] 1 Mar 2023

利用多路音频信号中的冗余信息进行远场语音识别

Feng-Ju Chang、*Anastasios Alexandridis*、*Rupak Vignesh Swaminathan*、*Martin Radfar*、*Harish Mallidi*、*Maurizio Omologo*、*Athanasios Mouchtaris*、*Brian King*、*Roland Maas*

Alexa Speech，亚马逊，美国

摘要

为了实现稳健的远场自动语音识别（ASR），现有技术通常采用声学前端（AFE），并级联神经转换器（NT）自动语音识别模型。然而，AFE 输出可能并不可靠，因为 AFE 中的波束形成输出可能会指向错误方向。解决这一问题的一种有效方法，是利用波束形成阶段之前、声学回声消除之后（回声消除后）的麦克风信号。我们认为，回声消除后信号和 AFE 输出具有互补性，因此可以利用这些信号之间的冗余信息，弥补 AFE 处理可能产生的错误。我们提出了两种融合网络，用以探索这种冗余信息并整合这些多通道（MC）信号：（1）基于频率-LSTM 的融合网络；（2）基于卷积神经网络的融合网络。我们将多通道融合网络扩展至 Conformer 转导器模型，并以端到端方式进行训练。在商用虚拟助手任务上的实验结果表明，与仅使用 AFE 输出的模型相比，使用 AFE 输出和两个回声消除后信号并结合融合网络，词错误率（WER）最高可相对降低 25.9%，参数量仅增加 $\leq 2\%$。

索引词— 声学前端、AFE、端到端自动语音识别、Conformer 转导器、卷积神经网络

1. 引言

基于神经转换器（NT）的模型，如循环神经网络转换器（RNN-T）[1], 、Conformer 转导器（CT）[2,3], 以及 ConvRNN-T [4]，在单通道自动语音识别中取得了当前最先进的成果。在许多情况下，语音助手距离声源较远（例如使用智能音箱时），且处于嘈杂环境中（存在来自各种声源或多个说话者的噪声），这使得声学场景颇具挑战性。对于这类情况，语音助手通常采用基于多麦克风或麦克风阵列的声学前端（AFE），利用波束形成和自适应降噪技术 [5–11]，为自动语音识别提供增强后的信号。

通常，在已知麦克风阵列的情况下，AFE 会根据信噪比（SNR）目标进行优化
几何结构和特定噪声场假设 [7]。因此，增强后的信号对于自动语音识别任务而言可能并非最优，因为优化过程并未考虑自动语音识别目标。此外，干扰源导致的不必要波束切换，以及对噪声场的强假设，可能会降低 AFE 的性能，进而影响下游自动语音识别任务。近年来，基于神经网络构建的波束形成器 [12–14] 放宽了上述限制，其效果优于基于信号处理方法的 AFE。此外，神经波束形成器可以与自动语音识别目标一起直接进行端到端优化 [15–23]。

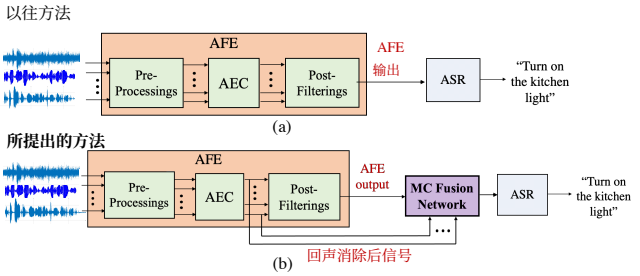


图 1。所提出方法的高级示意图。与以往方法 [15, 20–23] (a) 不同，我们在将信号输入自动语音识别模型之前，通过多通道融合网络 (b) 利用 AFE 输出与回声消除后信号之间的冗余。

本研究旨在充分利用 AFE [24]，生成的信息，其中还包括用于抑制回放信号或其他已知干扰信号的多通道声学回声消除（MCAEC）处理。声学回声消除输出可以与波束形成信号互为补充，尤其是在噪声较大的情况下。基于这一特点，我们设计了融合网络，利用 AFE 输出和回声消除后信号生成隐藏表示，并将其传递给自动语音识别组件，如图 1[1] 所示。这样，自动语音识别模型便可通过利用这种互补性进行端到端训练。[25,26] 中提出了类似的方法。然而，在 [25,26] 中，所有输入通道都直接传递给自动语音识别音频编码器，这使其计算开销更大，也降低了其在流式自动语音识别中的吸引力。

两个融合网络分别基于：(a) 频率-LSTM [27]，由神经波束形成器 [15] 对回声消除后信号进行预处理（NBF+F-LSTM）；以及 (b) 卷积神经网络（CNN）。在 NBF+FLSTM 融合网络中，我们扩展了 [15] 提出的神经波束形成方法（该方法最初是为 DNN-HMM自动语音识别声学建模提出的），并利用 F-LSTM 基于多路信号生成最终的隐藏表示。请注意，在 [15], 中，作者使用的是 AEC前麦克风信号，而非回声消除后信号。另一方面，CNN 融合网络的设计受 [4, 28, 29], 启发，但我们在此使用卷积来建模多通道信号。我们的融合网络用于增强 CT 自动语音识别模型 [2, 3]。请注意，我们的所有模型均采用端到端训练、支持流式处理，而融合网络在模型架构中增加的参数不到 2%。实验表明，所提出的方法优于仅使用 AFE 输出的 CT，词错误率最高可相对降低 25.9%。

[1]本文所示的 AFE 模块是对 [24] 的简化版本。AFE 输出表示 MCAEC、波束形成和降噪算法组合后的结果。

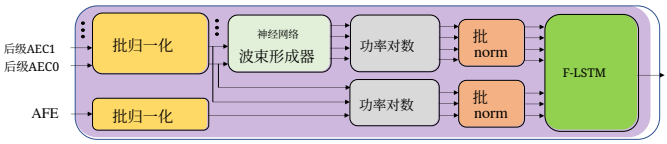


图 2。神经波束形成器+频率-LSTM（NBF+F-LSTM）融合网络。注意，相同颜色的批归一化层表示共享参数。

2. 用于自动语音识别的神经转导器

基线系统（图 1(a)）与我们的方法（图 1(b)）的共享组件是 NT 自动语音识别。它通常由编码器网络、预测网络和联合网络组成。编码器网络将输入时间帧$\mathbf{x_t}$ 编码为高层表示 h_t^{enc}。预测网络利用先前预测的词片$\mathbf{y_{u-1}}$ 并生成中间表示 h_u^{pred}。联合网络先对 h_t^{enc} 和 h_u^{pred}执行联合操作，再将输出传入一系列带激活函数的全连接层，最后经过 softmax 得到词片的概率分布。神经转换器基于 RNN-T损失进行训练，采用前向-后向算法 [1] 计算声学帧与转录词片之间所有可能的对齐。编码器和预测网络的架构可以有所不同，常见选择包括 RNN层 [1]，、Transformer [30] 或 Conformer 块 [2, 3]。在本研究中，我们使用 Conformer 编码器和 LSTM 解码器（CT）。请注意，我们的融合网络可应用于任何 NT，并不限于 CT。

3. 提出的多通道融合网络

我们的目标是让 CT 自动语音识别模型能够利用经不同技术处理的多路信号，并探索其中的冗余，包括：(1) 由基于信号处理的 AFE [24] 生成的增强信号，该 AFE 执行多通道声学回声消除、自适应波束形成和降噪；(2) 声学回声消除后的信号（回声消除后信号）。这些信号经多通道融合网络（图 1）处理后，形成高层表示，并输入 CT 模型的编码器网络。我们提出了两种多通道融合网络。

3.1. 神经波束形成器（窄带融合）+F-LSTM 融合网络

该方法采用一种可训练的神经波束形成器，其灵感来自 [15] 中作者将神经波束形成技术用于 DNN-HMM 自动语音识别声学建模的工作。该方法主要包括两个步骤：首先，使用神经波束形成器对多通道输入进行空间滤波；然后，通过频率 LSTM（F-LSTM）[27] 层融合 AFE 输出和回声消除后信号。融合网络的架构如图 2 所示。首先，对输入信号的频率区间进行批归一化

以稳定训练并加快收敛。

神经波束形成网络。 经过 M 批归一化的 AEC 后信号 $\mathbf{X}(t,\omega)$进入神经波束形成模块，在其中执行如下滤波操作：

$$Y(t,\omega,\mathbf{p}) = \mathbf{w}^H(t,\omega,\mathbf{p})\mathbf{X}(t,\omega) \tag{1}$$

其中， H表示厄米特运算，$\mathbf{w}(t,\ \omega,\ \mathbf{p})$ 是位置 $\mathbf{p}$ 处进行波束形成的复权重向量； t 和 ω 表示

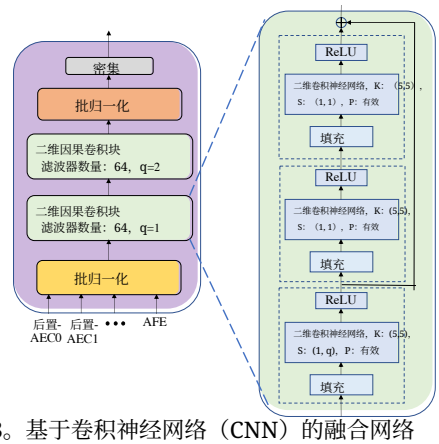


图 3。基于卷积神经网络（CNN）的融合网络

时间帧和角频率。

对于给定频率 ω，公式 (1) 中的复向量乘法可表示为矩阵形式：

$$\begin{bmatrix} Re(Y) \\ Im(Y) \end{bmatrix} = \begin{bmatrix} Re(w_0) & Im(w_0) \\ -Im(w_0) & Re(w_0) \\ \vdots & \vdots \\ Re(w_{M-1}) & Im(w_{M-1}) \\ -Im(w_{M-1}) & Re(w_{M-1}) \end{bmatrix}^T \begin{bmatrix} Re(X_0) \\ Im(X_0) \\ \vdots \\ Re(X_{M-1}) \\ Im(X_{M-1}) \end{bmatrix} \tag{2}$$

为简洁起见，省略了 $(t,\omega,\mathbf{p})$，而 w_i 表示复波束形成权重向量 $\mathbf{w}$的第 i 个元素。根据这一公式，可以通过使用 Ω 个 $2\times 2M$矩阵（每个频率 bin 对应一个矩阵），将波束形成操作纳入神经网络，其中 Ω 为频率区间总数。这些矩阵表示波束形成器权重。请注意，复信号特征被视为两个实值输入；复数神经波束形成器权重的实部和虚部均可学习。

初始权重经过初始化，使波束形成器朝向围绕阵列均匀分布的n_l 个观察方向。在训练过程中，根据 RNN-T 损失函数对 n_l 组权重进行优化。神经波束形成器会生成朝向 n_l 个观察方向的 n_l 个信号。接下来，将所有神经波束形成信号与原始麦克风阵列信号及 AFE 输出一并纳入考虑，总计 $n_l + M + 1$ 个信号。对于每个信号，我们通过对实部和虚部应用平方求和运算符，计算其在每个频率 bin 上的功率。随后，功率经过 ReLU 激活函数和一个对数函数（这两个函数也沿频率轴运算），以估算对数功率谱。最后，我们在每个信号的频率轴上再添加一个批归一化层。

使用频率 LSTM 进行融合。最后，将 $n_l + M + 1$ 信号融合为单一的隐藏表示。为此，我们应用一个 F-LSTM 层 [27]。F-LSTM 是一种常规 LSTM 层，沿频率轴而非时间轴运行。为了准备输入 F-LSTM，我们将所有信号拼接成一个向量，并沿频率轴以特定窗口 w和步幅 s 分块，从而在频域中创建长度为 w的序列。随后将此输入传入 LSTM。LSTM 的输出被投影为维度为 d_{flstm} 的隐藏表示，并作为 CT 音频编码器的输入。

3.2. 卷积神经网络(CNN)融合网络

图3展示了基于卷积神经网络（CNN）的多通道融合网络。其设计借鉴了 ConvRNN-T [4]，并进行了若干调整。首先，输入通道包含 AFE 输出和回声消除后信号。它们拼接后形成一个形状为 $T \times \Omega \times C$的张量，其中 T为时间帧数，Ω 为频率区间数，$C = M + 1$ 为输入通道数。在输入两个卷积块之前，批归一化层会沿频率轴对张量进行均值归一化。初步研究表明，这种批归一化对于稳定训练至关重要。每个卷积块包含三个卷积层，我们使用两个卷积块，因此总计有 6 个卷积层，如图3所示。每个卷积层后接一个 ReLU 激活，并在第一个 ReLU 的输入和输出之间添加额外的残差连接。为确保卷积计算具有因果性（因为我们的模型支持流式处理），我们在时间维度左侧进行零填充，填充数量为卷积核大小（K）减一。请注意，第二个卷积块将频率区间下采样两倍（步幅 $q = 2$），以保持模型规模。经过卷积块后，第二个批归一化层沿频率轴对 CNN输出的均值和尺度进行归一化；我们的研究发现，这有助于更好地收敛。最后，全连接层将归一化后的输出投影到 d_{cnn} 维，并将其馈送至 CT音频编码器。

4. 实验

4.1. 数据集

为评估所提出的方法，我们使用自有的去标识化远场数据集 [2]开展了一系列自动语音识别实验。这些数据集面向英语和虚拟助手任务。设备定向语音数据由配备7个麦克风、孔径为63毫米的智能音箱采集。训练集包含～ $500K$ 小时的单通道（SC）数据和 ～ $60K$ 小时的多通道（MC）数据。单通道数据由AFE 输出 [24]组成；AFE以7个麦克风信号作为输入。多通道数据除AFE 输出外，还包含2个孔径距离的回声消除后信号。在单通道数据（仅包含AFE 输出）上训练/测试多通道模型时，我们只需将这2个回声消除后信号替换为零。

我们使用以下单通道和多通道测试集：(1) *HEAD*：10 小时的单通道数据，代表远场语音助手任务的头部数据分布。(2) *MC Clean*：3 小时的高信噪比多通道数据，用于虚拟助手任务。(3) *MC Noisy*：20 小时的多通道数据，在受控条件下为虚拟助手任务采集。该数据集包含更具挑战性的声学场景，涵盖两种常见的设备位置和三种噪声条件。设备位置包括：(1) **DevLoc1**：设备位于扬声器正前方（相距 5 英尺）。(2) **DevLoc2**：设备位于扬声器前方，但距离较远，处于房间角落（相距 15 英尺）。噪声条件包括：(1) **静音**：无背景噪声；(2) **家用电器**：家电噪声，例如空调；(3) **媒体**：来自电视、收音机等设备的噪声。

4.2. 模型配置

4.2.1. 基线方法

我们使用 Conformer 转导器（CT）作为以下所有基线模型中的自动语音识别组件。（1）**AFE CT**：它是仅使用 AFE 输出、不使用融合网络的基线。（2）**CA CT**：该基线采用 [26] 提出的基于通道注意力（CA）的融合网络。我们将 AFE 输出和后置 AEC 都输入该模型，以评估我们的融合网络。为保持模型规模可比，我们使用一个跨通道注意力层，并根据验证集上的性能，从注意力头数量（1、2、4）和头部维度（512、256、128）中选出最佳配置。所有基线中的 CT 均采用相同配置：两个卷积层后接 14 个 Conformer 块。两个卷积层均包含 128 个卷积核，卷积核大小为 3；沿时间轴的步幅分别为 2 和 1。每个 Conformer 块包含一个具有 1024 个单元的前馈网络（FFN）模块、一个卷积核大小为 15 的卷积模块，以及一个嵌入维度为 512、包含 8 个头的注意力模块。参数量为 8144 万。CT 编码器的输出先映射为 512 维隐藏表示，再输入联合网络。预测网络由两个各含 1024 个单元的 LSTM 层组成。其输出也映射为 512 维，以匹配音频编码器的输出。联合网络由一个包含 512 个单元的前馈神经网络组成。我们使用 Adam 优化器 [31]，并按照 [30, 32] 调整学习率。

4.2.2. 融合网络

神经波束形成器（NBF） + F-LSTM：对于神经波束形成器，我们将观察方向的数量设为 $n_l = 4$。初步分析表明，增加观察方向的数量并不会带来明显的性能提升。对于 F-LSTM 层，我们使用一个包含 48 个频率区间、步幅为 15 的窗口，并采用两个包含 16 个单元的双向 LSTM。F-LSTM 的输出被投影为 $d_{flstm} = 192$ 维向量，以匹配基线模型的特征维度。参数总数为 372K，仅为 Conformer 编码器参数量的 $\sim 0.5\%$。

卷积融合网络（CNN）：该网络包含两个二维因果卷积块，每个块均包含三层卷积层。第一个二维因果卷积块包含 64 个大小为 K=(5,5) 的卷积核，时间轴上的步幅为 1，频率轴上的步幅为 q=1。第二个二维因果卷积块的配置与第一个相同，但频率轴上的步幅为 q=2。最后的全连接层将归一化后的 CNN 块输出投影为 $d_{cnn} = 192$ 维向量。参数总数为 1.4M，仅为 Conformer 编码器参数量的 $\sim 1.8\%$。所有模型的规模相近，~ 82 总参数量为百万级。

4.3. 输入和输出配置

AFE CT模型的输入特征为对数滤波器组能量（LFBEs），这些特征从以16 kHz采样的音频采样中提取，帧率为10 ms，窗口大小为32 ms。每一帧的特征与左侧两帧的特征进行堆叠，随后以3倍因子进行下采样，以实现低帧率，最终得到192维特征。所提出的方法使用短时傅里叶变换（STFT）特征，这些特征通过对采用32 ms窗口、帧率为10 ms的片段进行512点傅里叶变换计算得到。我们将复数值分离为实部和虚部，从而使每一帧得到512维特征。同样，我们使用3帧的左侧上下文。使用子词分词器 [33] 从转录文本中创建输出词元；我们使用4K个词元。

[2]据我们所知，目前没有同时包含 AFE 输出和 AEC 后的麦克风阵列信号的数据集。

表 1。所提出模型和基于通道注意力（CA）的融合网络在 **AFE CT** 上，以及在 *HEAD*、*MC Clean* 和 *MC Noisy* 测试集上的词错误率相对降低（WERRs）（%）。这三个模型均使用 AFE 输出和后置 AEC 作为输入。数值越高表示词错误率越优。

融合网络(↓)/测试集(→)	*HEAD*	*MC Clean*	*MC Noisy*
NBF+F-LSTM	5.66	5.31	**10.75**
CNN	**7.76**	**6.10**	10.38
CA [26]	4.61	3.74	8.96

表 2。基于 WERRs(%) **相对于 AFE CT** 的消融研究：在 *HEAD*、*MC Clean* 和 *MC Noisy* 测试集上，将所提出的模型与移除融合网络或后置 AEC 的模型进行比较。数值越高表示 WER 越低。

模型(↓)/测试集(→)	*HEAD*	*MC Clean*	*MC Noisy*
AFE CT	基线	基线	基线
+后置 AEC+窄带融合+频率长短期记忆网络（我们的方法）	**5.66**	**5.31**	**10.75**
+ 后置 AEC（**无融合网络**）	1.89	3.54	7.92
+窄带融合+频率长短期记忆网络（**无后置 AEC**）	0.84	1.18	6.70
+ 后置 AEC+窄带融合+卷积神经网络（我们的方法）	**7.76**	6.10	**10.38**
+ 后置 AEC（**无融合网络**）	1.89	3.54	7.92
+卷积神经网络（**无后置 AEC**）	5.24	**8.66**	6.89

5. 结果与讨论

我们在整个实验[3]过程中报告所提出方法相对于基线的相对词错误率降低（WERR）。给定模型A的词错误率（WER_A）和基线B的词错误率（WER_B），A相对于B的WERR可通过（$\text{WER}_B - \text{WER}_A$）/$\text{WER}_B$计算；WERR值越高表示性能越好。

表 1 展示了所提出模型和基线模型在三个测试集上的词错误率相对降低。通过利用回声消除后信号以及 AFE 输出，所有基于融合网络的 CT 模型都优于仅使用 AFE 的模型（词错误率相对降低值为正）。此外，所提出的 NBF+F-LSTM 和卷积神经网络融合网络优于基于通道注意力的融合网络，这表明在模型规模相当的情况下，它们能够更好地学习跨通道的时频关系（第 4.2.2 节）。我们还可以看到，与 MC Clean 测试集相比，后置 AEC 和融合网络带来的提升在 MC Noisy 测试集上更为显著。例如，NBF+FLSTM 相对于 AFE CT 在 MC Noisy 上的词错误率相对降低为 10.75%，而在 MC Clean 上为 5.31%。这表明所提出的方法能够适应具有挑战性的声学条件。

融合网络和后置 AEC 的影响：我们通过移除融合网络（**无融合网络**）或后置 AEC（**无后置 AEC**），评估模型中各个子组件的贡献。请注意，**无融合网络** 仅将 AFE 输出与后置 AEC 拼接起来，并通过一个 FFN 层投影到 d_{flstm}/d_{cnn}=192 维。结果如表 2 所示。对于 NBF+F-LSTM
基于融合网络的模型中，我们发现移除后置 AEC 所带来的词错误率改善，远小于移除融合网络所带来的改善。在基于 CNN 融合网络的模型中，我们在*HEAD* 和*MC Noisy*上也观察到了类似趋势。在这种情况下，添加后置 AEC 并未进一步提升*MC Clean* 上的表现（相较于 AFE+CNN CT，结果为 6.10% vs. 8.66%）。

对不同信噪比的鲁棒性：在图 4 中，我们进一步在 *HEAD* 测试集上评估了模型对不同信噪比水平的鲁棒性。

[3]基线系统的绝对词错误率低于 10%。根据机构政策，此处不报告绝对词错误率。

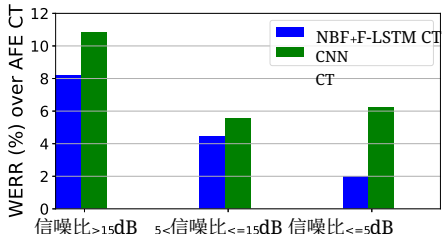


图 4。所提方法相对于不同信噪比下**AFE CT** 的词错误率相对降低（%）。数值越高，词错误率越低。

可以观察到，所提出的两种方法在整个信噪比范围内均带来了提升。与表 1 一致，卷积神经网络融合网络带来的提升最大。

对设备位置和噪声类型的鲁棒性：最后，我们评估了模型在*MC Noisy* 测试集所提供的不同设备位置与噪声类型组合下的表现。结果见表 3。可以看出，在**DevLoc2**条件下，与未使用后置 AEC 的基线系统相比，所提模型表现更佳，相对词错误率改善幅度从 10.52%（静音噪声下的 NBF+FLSTM CT）到 25.91%（家电噪声下的 CNN CT）不等。这表明后置 AEC 有助于提升 AFE 的远场自动语音识别性能。在**DevLoc1**条件下，我们观察到静音噪声和家电噪声下的改善幅度较小，而媒体噪声下的相对性能分别下降了 2.28% 和 6.71%。我们推测，在 DevLoc1 中，噪声源比 DevLoc2 中更靠近目标说话人。后置 AEC 受到媒体噪声的严重污染，因此无法提供有用的补充信息来纠正 AFE 产生的错误。在这种情况下，加入后置 AEC 可能会适得其反。

表 3。所提方法相对于**AFE CT** 在不同设备位置和噪声类型下的词错误率相对降低（%）。数值越大表示词错误率越低。

设备位置	噪声类型	窄带融合+F -LSTM CT	CNN CT
DevLoc1	静音	8.23	**12.05**
～5英尺远	家用电器	**8.21**	6.87
扬声器前方	媒体	-2.28	-6.71
DevLoc2	静音	10.52	**13.86**
距离～15英尺	家用电器	21.47	**25.91**
在扬声器前方	媒体	12.76	**17.92**

6. 结论

我们提出了两种基于 NBF+F-LSTM 和 CNN 的融合网络，以利用冗余信息并整合 AFE 输出与回声消除后信号。实验结果表明，在模型参数仅增加不到 2% 的情况下，我们的融合网络优于基于通道注意力的融合网络。此外，消融研究表明，同时利用 AFE 和回声消除后的麦克风阵列信号可带来最佳改进，例如在*MC Noisy* 测试集上提升 10.75%。此外，我们的方法对不同信噪比、设备位置和噪声类型具有更强的鲁棒性，相较于 AFE CT，词错误率最高可相对降低 25.91%。

7. 参考文献

[1] Alex Graves，《使用循环神经网络进行序列转导》，*ICML* 研讨会，2012年。

[2] Bo Li、Anmol Gulati、Jiahui Yu、Tara N Sainath、Chung-Cheng Chiu、Arun Narayanan、Shuo-Yiin Chang、Ruoming Pang、Yanzhang He、James Qin 等，《一种更好、更快速的流式自动语音识别端到端模型》，载于*ICASSP*，2021年。[3] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Par-mar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zheng-dong Zhang、Yonghui Wu 等，《Conformer：用于语音识别的卷积增强型 Transformer》，*arXiv* 预印本 *arXiv:2005.08100*，2020年。[4] Martin Radfar、Rohit Barnwal、Rupak Vignesh Swaminathan、Feng-Ju Chang、Grant P Strimel、Nathan Susanj 和 Athana-sios Mouchtaris，《ConvRNN-T：用于流式语音识别的卷积增强型循环神经网络转导器》，*Interspeech*，2022年。[5]Reinhold Haeb-Umbach、Shinji Watanabe、Tomohiro Nakatani、Michiel Bacchiani、Bjorn Hoffmeister、Michael L Seltzer、Heiga Zen 和 Mehrez Souden，《数字家庭助理的语音处理：将信号处理与深度学习技术相结合》，*IEEE*《信号处理杂志》，第36卷，第6期，第111–124页，2019年。[6] Maurizio Omologo、Marco Matassoni 和 Piergiorgio Svaizer，《使用麦克风阵列进行语音识别》，载于麦克风阵列，第331–353页。施普林格，2001年。[7] Matthias W¨olfel 和 John McDonough，远场语音识别，约翰·威立父子出版公司，2009年。[8] Kenichi Kumatani、John McDonough 和 Bhiksha Raj，《用于远场语音识别的麦克风阵列处理：从近讲麦克风到远场传感器》，*IEEE*《信号处理杂志》，第29卷，第6期，第127–140页，2012年。[9] Keisuke Kinoshita 等，《REVERB 挑战赛总结：混响语音处理研究的最新进展与遗留挑战》，*EURASIP*《信号处理进展杂志》，第2016卷，第1期，第7页，2016年。[10]Tobias Menne、Jahn Heymann、Anastasios Alexandridis、Kazuki Irie、Albert Zeyer、Markus Kitza、Pavel Golik、Ilia Kulikov、Lukas Drude、Ralf Schl¨uter、Hermann Ney、Rein-hold Haeb-Umbach 和 Athanasios Mouchtaris，《RWTH/UPB/FORTH 系统组合在第四届 CHiME 挑战赛评测中的应用》，载于*CHiME-4* 研讨会，2016年。[11] Reinhold Haeb-Umbach、Jahn Heymann、Lukas Drude、Shinji Watanabe、Marc Delcroix 和 Tomohiro Nakatani，《远场自动语音识别》，《*IEEE* 论文集》，第109卷，第2期，第124–148页，2020年。[12]Tsubasa Ochiai、Marc Delcroix、Tomohiro Nakatani 和 Shoko Araki，《采用基于自注意力的跟踪技术，为移动说话人进行基于掩码的神经波束形成》，*arXiv* 预印本 *arXiv:2205.03568*，2022年。[13] Jahn Heymann、Lukas Drude 和 Reinhold Haeb-Umbach，《用于声学波束形成的基于神经网络的频谱掩码估计》，载于*ICASSP*，2016年。[14] Hakan Erdogan 和 John R 等，Hershey，《使用单通道掩码预测网络改进 MVDR 波束形成》，载于*Interspeech*，2016年。[15] Minhua Wu、Kenichi Kumatani、Shiva Sundaram、Nikko Str¨om 和 Bj¨orn Hoffmeister，《用于远场语音识别的频域多通道声学建模》，载于*ICASSP*，2019年，第6640–6644页。[16] Tsubasa Ochiai、Shinji Watanabe 等人，《多通道端到端语音识别》，*arXiv*预印本 *arXiv:1703 04783*，2017。.[17] Xuankai Chang、Wangyou Zhang、Yanmin Qian、J onathan Le Roux 和 Shinji Watanabe，《MIMO-speech：端到端多通道多说话人语音识别》，发表于*ASRU*，2019年。[18] Xuankai Chang、Wangyou Zhang、Yanmin Qian、J onathan Le Roux 和 Shinji Watanabe，《基于Transformer的端到端多说话人语音识别》，发表于*ICASSP*，2020年。[19] Kenichi Kumatani、Minhua Wu、Shiva Sundaram、Nikko Str¨om 和 Bj¨orn Hoffmeister，《用于远场语音识别的多几何空间声学建模》，发表于*ICASSP*，2019年。[20] Bo Li、Tara N Sainath 等人，《用于鲁棒多通道语音识别的神经网络自适应波束形成》，发表于*Interspeech*，2016年。[21] Xiong Xiao、Shinji Watanabe 等人，《用于多通道语音识别的深度波束成形网络》，发表于*ICASSP*，2016年。[22] Zhong Meng、Shinji Watanabe 等人，《用于鲁棒多通道语音识别的深度长短期记忆自适应波束成形网络》，发表于*ICASSP*，2017年。[23] Yulan Liu、Pengyuan Zhang 和 Thomas Hain，《在基于远场多麦克风的语音识别中使用神经网络前端》，发表于*ICASSP*，2014年。[24] Robert Ayrapetian、Philip Hilmes、Mohamed Mansour、Trausti Kristjansson 和 Carlo Murgia，《无线信道上的异步声学回声消除》，发表于*ICASSP*，2021年。[25] Feng-Ju Chang、Martin Radfar、Athanasios Mouchtaris、Brian King 和 Siegfried Kunzmann，《用于语音识别的端到端多通道Transformer》，发表于*ICASSP*，2021年。[26] Feng-Ju Chang、Martin Radfar、Athanasios Mouchtaris 和 Maurizio Omologo，《用于语音识别的多通道Transformer转导器》，发表于*Interspeech*，2021年。[27] Maarten Van Segbroeck、Harish Mallidih、Brian King、I-Fan Chen、Gurpreet Chadha 和 Roland Maas，《多视角频率LSTM：一种高效的自动语音识别前端》，2020年。[28]Ching-Feng Yeh、J ay Mahadeokar、Kaustubh Kalgaonkar、Yongqiang Wang、Duc Le、Mahaveer Jain、Kjell Schubert、Christian Fuegen 和 Michael L Seltzer，《Transformer转导器：基于自注意力的端到端语音识别》，*arXiv*预印本 *arXiv:1910.12977*，2019年。[29] Abdelrahman Mohamed、Dmytro Okhonko 和 Luke Zettlemoyer，《用于自动语音识别的带卷积上下文的Transformer》，*arXiv*预印本 *arXiv:1904.11660*，2019年。[30] Linhao Dong、Shuang Xu 和 Bo Xu，《Speech-transformer：一种用于语音识别的无循环序列到序列模型》，发表于*ICASSP*，2018年。[31] Diederik P Kingma 和 Jimmy Ba，《Adam：一种随机优化方法》，*arXiv*预印本 *arXiv:1412 6980*，2014。.[32] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin，《Attention Is All You Need》，发表于*NeurNIPS*，2017年。[33] Rico Sennrich、Barry Haddow 和 Alexandra Birch，《使用子词单元进行稀有词神经机器翻译》，发表于*ACL*，2016年。