arXiv:2302.08549v1 [eess.AS] 16 Feb 2023

说话人人变更检测：Transformer Transducer 自动语音识别 R

Jian Wu、*Zhuo Chen*、*Min Hu*、*Xiong Xiao*、*Jinyu Li*

Microsoft，美国华盛顿州雷德蒙德市 One Microsoft Way

摘要

说话人切换检测（SCD）是一项重要特征，它通过在说话人变更点将词序列划分为段落，提高自动语音识别（ASR）系统识别结果的可读性。现有的 SCD 解决方案要么需要额外集成基于时间的判定结果与识别出的词序列，要么将 ASR 与 SCD 紧密集成，从而限制了这两项任务各自达到最佳性能的潜力。为解决这些问题，我们提出了一种针对说话人变更检测任务的新型框架，在现有的 Transformer Transducer 自动语音识别（TT-ASR）网络之上构建一个额外的说话人变更检测模块。我们在该框架中探索了两种说话人变更检测网络变体，它们能够自然地估计每个词的说话人变更概率，同时允许 ASR 和 SCD 采用独立的优化方案，以实现最佳性能。实验表明，与联合说话人变更检测与自动语音识别基线相比，我们的方法在 LibriCSS 和 Microsoft 呼叫中心数据集上显著提升了 F1 分数，且不会降低 ASR 性能。

索引词— 说话人切换检测，Transformer Transducer，端到端自动语音识别，F1 分数

1. 引言

说话人切换检测（SCD）旨在估计音频流中的说话人转换点，该音频流中可能有多名活跃说话人。在现代语音处理应用中，说话人切换检测模块通常提供两项重要功能。首先，准确的说话人变更信号对于改进会话识别系统的显示格式至关重要；当活跃说话人发生变化时，显示的转录文本应另起一行。其次，说话人切换检测是说话人分离应用的重要前端处理 [1, 2], ，在这类应用中，输入语音首先按说话人变更点进行分段，然后再进行聚类等后续处理。

现有的说话人切换检测系统大致可分为两类：基于时间的检测 [3, 4, 5, 6] 和基于词的检测[7, 8]。在基于时间的检测中，说话人切换检测通常依赖输入语音的纯声学特征，并以帧级方式估计判定结果。尽管这种方法具有较高的判定分辨率，但基于时间的判定通常有三项局限。首先，受静音或噪声的影响 [4]，难以准确定位说话人边界。其次，与自动语音识别系统结合时，需要进行后处理，将说话人变更标签插入识别出的词序列中，因为检测到的基于时间的变更点可能不在词边界之间。最后，这种方法仅利用声学信息，而语义信息已被证明有助于说话人切换检测和说话人分离任务 [9, 10, 11]。

与基于时间的检测不同，基于词的说话人变更检测直接估计词与词之间的说话人变更点。这样一来，说话人变更检测的判断自然与自动语音识别的词序列对齐，从而省去了基于时间的系统所需的同步步骤。更重要的是，这种方法使说话人变更检测能够获取自动语音识别模块中的声学和语义信息，从而有望取得更好的性能。另一方面，随着端到端自动语音识别技术的最新进展，过去已有多种任务与自动语音识别相结合，包括端点检测器 [15], 轮替发言 [16], 说话人分离 [17, 18]和多说话人识别 [19], ，并且相较于模块化方法取得了有竞争力的结果。

尽管基于词的说话人变更检测系统已展现出令人鼓舞的结果，现有的联合 SCD-ASR [7] 解决方案仍常受两项局限的影响。首先，在联合模型架构中，很难使两项任务都达到最佳性能。我们的研究发现，在训练过程中加入 SCD 并进行额外调优以提升其性能时，往往会观察到更高的词错误率（WER）。其次，由于说话人变更检测网络需要检测声学上的说话人特征变化，因此需要更长的延迟来确保说话人变更检测模块的稳健性。然而，受自动语音识别应用的流式请求限制，联合 SCD-ASR 模型通常采用低延迟设置，从而限制了说话人变更检测模块的性能。

为解决这些局限，本文提出一种基于现有 Transformer Transducer 自动语音识别（TT-ASR） [20]网络的说话人切换检测方法。我们的模型并未将说话人切换检测直接整合到自动语音识别解码中，而是在 TT-ASR 之上增加了一个用于说话人切换检测的网络。给定部分自动语音识别解码词序列，说话人变更检测网络会预测词级说话人变更概率。具体而言，根据是否能够获取 TT-ASR 识别出的词序列的显式对齐，我们研究了所提框架的两种实现，即 **SCD-ENC**和 **SCD-DEC**。基于所提出的说话人变更检测模型，我们引入了一种分块式流处理管线，可在推理期间灵活调整延迟。我们在两个数据集上评估了这些模型，结果表明，与联合 SCD-ASR 基线相比，所提出的说话人变更检测模型在两个数据集上都取得了显著更高的 F1 分数，同时没有损失自动语音识别准确率。

本文其余部分安排如下：第 2 节和第 3 节分别介绍 TT-ASR 和说话人变更检测方法。第 4 节介绍实验设置，包括数据、模型和训练细节。第 5 节讨论实验结果。第 6 节展望未来工作并总结全文。

2. Transformer Transducer 自动语音识别

我们的 TT-ASR 基于神经转导器结构 [21] ，采用 Transformer [22] 编码器堆栈，并结合流式方案 [20] ，以实现高准确率和低延迟。第i层（$1 \leqslant i \leqslant L$）编码器层的 Z_i 输出采用预归一化方式计算，如下列公式所示

$$\begin{aligned} \bar{Z}_{i-1} &= Z_{i-1} + \texttt{MHSA}_i\left(\texttt{ln}(Z_{i-1})\right), \\ Z_i &= \bar{Z}_{i-1} + \texttt{FFN}_i\left(\texttt{ln}(\bar{Z}_{i-1})\right), \end{aligned} \tag{1}$$

其中，ln(·) 表示层归一化。MHSA$_i$(·) 和 FFN$_i$(·) 分别表示第 i 层的多头自注意力（MHSA）模块和

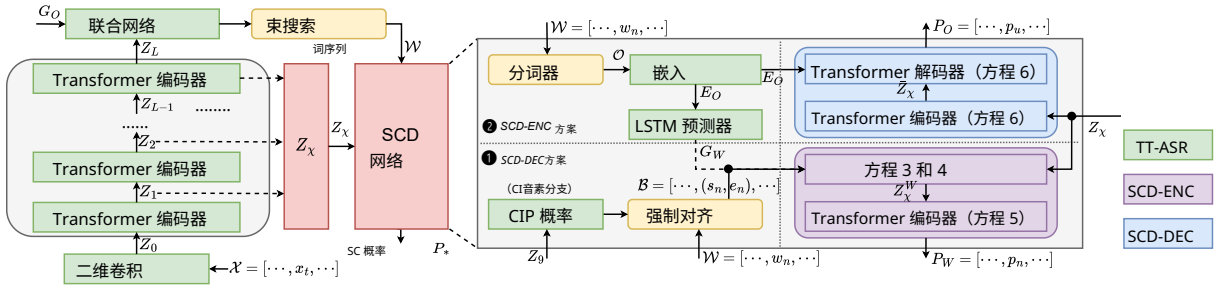

图1。我们基于TT-ASR的说话人变更检测方法示意图。绿色模块表示纯自动语音识别模块，黄色圆角方块表示运行时算法。右侧的蓝色和紫色模块分别展示了SCD-ENC和SCD-DEC的实现。

前馈网络（FFN）。MHSA 的实现采用了学习得到的相对位置编码。给定 T 帧声学特征 $\mathcal{X} = [x_0, \cdots, x_{T-1}]$， Z_0 计算为 $Z_0 = \text{CONV}(\mathcal{X})$，其中 CONV(·) 表示二维卷积下采样层，用于在 Transformer 编码器之前降低帧率。

关于TT-ASR中的解码器，我们使用长短期记忆网络（LSTM）作为预测网络，并使用线性层作为联合网络。给定估计的非空白词元 o_u 和隐藏状态 h_u，预测器在第 $u+1$ 步的输出 g_{u+1} 可写为

$$g_{u+1}, h_{u+1} = \texttt{LSTM}\left(\texttt{Embed}(o_u), h_u\right) \tag{2}$$

其中，Embed(·)表示词元嵌入层。长度为 U 的最佳解码词元（SentencePiece）序列 $\mathcal{O} = [o_0, \cdots, o_{U-1}]$（或$N$词序列 $\mathcal{W} = [w_0, ..., w_{N-1}]$）通过对联合网络输出的概率分布进行束搜索算法获得。

3. 方法

3.1. 特征设计

从自动语音识别编码器中提取的隐藏表示被用作说话人变更检测网络的基本声学特征。除 Z_i 外，还研究了对编码器层输出 $\{Z_1, \cdots, Z_L\}$ 进行加权和 $Z_\omega = \sum_{i=1}^{L} \omega_{i-1} Z_i$。与 [23], 类似，可学习权重 ω 被约束为 $\sum_i \omega_i = 1$。对于语义特征，在不同的实现中采用了来自 TT-ASR 的不同表示，具体内容将在下一节讨论。

3.2. SCD网络实现

以声学特征 $Z_\chi \in \{Z_1, \cdots, Z_L, Z_\omega\}$ 和词序列 $\mathcal{W}$ 作为输入，SCD决策网络旨在显式或隐式地对齐来自两种模态的信息，并为每个词估计最终的说话人变更点，如图1所示。具体而言，我们探索了两种实现方式：SCD-ENC 和 SCD-DEC，具体取决于是否可以通过 TT-ASR模型获得基于对齐的词边界 [24] 。在 SCD-ENC 中，包含一个显式的强制对齐步骤，以从识别出的词序列中获取词边界。随后，对声学特征Z_χ 进行平均，并将其转换为基于词的表示 Z_χ^W，再通过一组 Transformer 编码器进行处理，以生成最终的说话人变更判定。另一方面，在 SCD-DEC 中，我们利用 Transformer 解码器，其中包含词元序列$\mathcal{O}$ 与声学特征 Z_χ之间的源目标注意力机制，从而在不使用显式词边界信息的情况下，直接估计基于词的说话人变更。

SCD-ENC 有了自动语音识别解码得到的词序列 $\mathcal{W}$ $\mathcal{B}$ $[w_0, ..., w_{N-1}]$ 及其词边界 $[\ s_0\ e_0\ \cdots\ s_{N-1}\ e_{N-1}\ ]$ (,), (,), 通过强制对齐，使用一组 Transformer 编码器层来预测每个词的说话人变更概率，如图1中的紫色方块所示。词 w_n 的音频嵌入 $z_\chi^{w_n}$通过对 Z_χ 矩阵中与词边界（s_n， e_n）对应的部分取平均计算得到:

$$z_\chi^{w_n} = \texttt{Mean}\left(Z_\chi^{[s'_n, e'_n]}\right), \tag{3}$$

其中， s'_n 和 e'_n 分别表示卷积层的下采样率和卷积核大小。我们还可以利用 TT-ASR 的预测器输出 $G_O = [g_0, \cdots, g_{U-1}]$，来弥补 $z_\chi^{w_n}$中缺失的语义信息。在本研究中，我们使用词语的最后一个词元表示作为每个词语的文本特征，并将其与原始 $z_\chi^{w_n}$拼接，形成新的 $z_\chi^{w_n}$:

$$z_\chi^{w_n} = \texttt{Concat}\left(\left[z_\chi^{w_n}, g_{w_n}\right]\right), \tag{4}$$

其中， g_{w_n} 是预测器对词语 w_n 输出的最后一个词元表示。最后，. 将音频嵌入序列 $Z_\chi^W = [z_\chi^{w_0}, \cdots, z_\chi^{w_{N-1}}]$输入 Transformer 编码器层，以估计词级说话人变更概率 $P_W = [p_0, \cdots, p_{N-1}]$:

$$P_W = \sigma\left(\texttt{TransformerEnc}(Z_\chi^W)\right), \tag{5}$$

其中， σ（·）是 Sigmoid函数 .

SCD-DEC 由于强制对齐会带来额外的计算开销，我们也可以采用Transformer解码器结构，避免显式对齐声学特征和文本单元；该结构直接根据声学特征和词元嵌入 $E_O = [\text{Embed}(o_0) \cdots, \text{Embed}(o_U)]$（来自ASR）预测词元级说话人变更概率 $[p_0, \cdots, p_{U-1}]$ P_O = Z_χ 。我们将其记作

$$\begin{aligned} \bar{Z}_\chi &= \texttt{TransformerEnc}\left(Z_\chi\right), \\ P_O &= \sigma\left(\texttt{TransformerDec}\left(\bar{Z}_\chi, E_O\right)\right). \end{aligned} \tag{6}$$

在这里，我们添加了若干 Transformer 编码器层，以进一步处理 ¯声学特征 Z_χ，而不是将其直接输入 Transformer 解码器层。由于 P_O 是词元级概率向量，我们通过选取每个词内词元概率的最大值，将其映射为最终的词级结果 P_W。

3.3. 目标函数

根据前文讨论，词级（或词元级）说话人切换检测是一个典型的双分类问题。考虑到众所周知的样本平衡问题，我们采用焦点损失进行网络优化 [25]。给定 $Y_* = \{Y_W, Y_O\}$作为词或词元序列上的说话人变更标签，损失函数写为

$$\mathcal{L}(P_*, Y_*) = -\sum_i \psi_i y_i \log(p_i) + \phi_i (1 - y_i) \log(1 - p_i), \tag{7}$$

其中 $\psi_i = \alpha(1 - p_i)^\gamma$ 和 $\phi_i = (1 - \alpha) p_i$ α^γ,是可调超参数，分别用于调整正负样本以及简单样本和困难样本的权重。

3.4. 分块推理

为处理长音频的说话人变更检测（SCD），我们为模型提出了一种分块处理方案。运行时，每段输入的长音频首先由流式 TT-ASR 进行识别。与 [26], 中的连续语音分离类似，随后将解码后的词序列划分为重叠分块，分块长度为 N_c 个词，回看范围为 N_h 个词，未来上下文为 N_f 个词。之后，使用所提出的说话人变更检测网络对每个分块估计说话人变更点，网络以词分块及其对应的声学特征帧为输入。在 SCD-DEC 中，利用词发射时间确定每个分块的声学特征边界，该时间在束搜索结束后即可获得。在 SCD-ENC 中，则直接对特征 Z_χ^W 进行分段。最后，将分块决策合并，得到整个长音频的决策。这样，通过改变 N_h、N_c N_h 和 的取值，可以灵活调整说话人变更检测网络的延迟。与此同时，无论说话人变更检测的设置如何，自动语音识别仍能保持低延迟。这种灵活性使说话人变更检测和自动语音识别的延迟能够分别调整，有利于提升两项任务的性能。

4. 实验设置

4.1. 数据

使用 Microsoft 内部数据训练自动语音识别和说话人变更检测模型。数据包含 6.5 万小时的匿名化语音，已移除个人身份信息。通过在线数据模拟生成说话人轮次，用于训练说话人变更检测模型。将来自不同说话人的 M 个话语拼接在一起，其中 M 从 2 到 4 均匀采样，用于我们提出的说话人变更检测模型；联合说话人变更检测与自动语音识别基线则从 1 到 4 均匀采样。方程 7 中的训练标签 $\{Y_W, Y_O\}$ 定义如下：将每位说话人话语中最后一个词和词元的位置标记为 1，其余位置标记为 0。

我们构建了一个包含 180 万个词、涵盖多个应用场景的评估集，用于自动语音识别评估。针对说话人变更检测，我们收集了两个测试集：来自 LibriCSS 录音风格（LS-ST）数据的非重叠会话 [26, 27]，以及 Microsoft 供应商采集的呼叫中心（MS-CC）数据，用于评估。LS-ST 包含 114 段 16 kHz 录音，总时长为 13.6 小时。MS-CC 集包含 400 段 8 kHz 音频会话，总时长为 24 小时。LS-ST 和 MS-CC 集中的说话人轮次数分别为 4,834 和 9,409。测试时，先将 8 kHz 音频上采样至 16 kHz。

4.2. 模型结构

本文使用的 TT-ASR 模型包含一个 2 层卷积网络，卷积核大小为 3、步长为 2；一个 18 层 Transformer 编码器；一个 2 层、维度为 1024 的 LSTM 预测器；以及一个 512 维的联合网络。每个编码器块中的多头自注意力层的注意力维度设为 512，包含 8 个头；采用了维度为 2048 的前馈神经网络层，并使用高斯误差线性单元（GELU）。自动语音识别单元使用 4035 个词片段，以及空白符和句末符号 ⟨eos⟩。

[7]本文选用联合说话人变更检测与自动语音识别方法作为基线，并在原始 TT-ASR 输出单元中添加说话人变更符号 ⟨sc⟩ 以指示说话人变更。我们提出的 SCD-ENC 模型由 6 个非因果 Transformer 编码器组成，其多头自注意力和前馈神经网络参数与 TT-ASR 相同；SCD-DEC 则由 1 个包含 2 层的非因果 Transformer 编码器和后续的 4 个 Transformer 解码器层组成。两者的参数总量分别为 19.3M 和 24.0M。

4.3. 训练与评估详情

为 TT-ASR 提取了 80 维对数梅尔滤波器组作为输入特征，窗口为 25 毫秒，帧移为 10 毫秒。对每个话语应用全局均值方差归一化和音量扰动。TT-ASR 使用 16 个 GPU 训练了 300K 步，采用常规线性衰减学习率调度器。预热步数为 25K，峰值学习率设为 $1.6e^{-3}$。在 TT-ASR 中，流式掩码 [20]采用 4 的分块大小，因此延迟为 160 毫秒。用于强制对齐的上下文无关（CI）音素分支 [24] 基于第 9 个编码器层的输出（Z_9）进行训练。在 SCD-ASR 基线中，当采样到多个话语时，会在原始转写文本之间插入说话人变更标记 ⟨sc⟩ 。对于所提出的 SCD 模型，我们在冻结 ASR 参数的同时，使用 8 个 GPU 训练了 175K 步，批量大小为 40K 帧。峰值学习率采用 $1e^{-4}$，预热步数为 25K。公式 7 中的 α 和 γ 分别设为 0.8 和 0.5。本研究中的所有模型均使用 AdamW 优化器进行优化。为评估所提出说话人变更检测网络的流式性能，本研究保留了 $N_h + N_c + N_f = 16$，以说明我们方法的有效性。

4.4. 评估指标

我们采用 F1 分数评估说话人变更模型的质量，并使用了与 [16] 类似的方法。首先，构建标注了说话人变更的假设转写文本和参考转写文本，并在每个说话人变更点处向假设和参考词序列插入一个⟨sc⟩ 标记。随后，对标注后的参考文本和假设文本进行对齐，并在参考词为 ⟨sc⟩的位置计算 F1 分数。对于 TT-ASR 和 SCD-ASR 基线，我们分别直接将输出 ⟨eos⟩ 和 ⟨sc⟩ 符号用作每个样本的说话人变更判定。对于 SCD-ENC 和 SCD-DEC，则将说话人变更概率大于 0.5 阈值的词位置选为变更点。

5. 结果

5.1. 说话人切换检测与自动语音识别之间的平衡

表1展示了所提出模型和各基线模型的说话人切换检测与自动语音识别结果，其中可以观察到一些现象。尽管所有系统都取得了相近的自动语音识别性能，但它们在说话人切换检测任务上的表现却存在显著差异。由于 TT-ASR 仅使用单说话人语句进行训练，基于 ⟨eos⟩ 的说话人切换检测无法捕捉说话人差异的特征，导致其在两个数据集上的 F1 分数都较低。在训练数据中加入说话人轮次后，SCD-ASR 在保持与 TT-ASR 相近词错误率的同时，取得了更高的 F1 分数。然而，该 F1 分数仍远未达到理想水平。一个潜在原因可能是流式自动语音识别中的低延迟设置，即在如此低的延迟下，很难实现高质量的说话人切换检测。通过更新顶部 10 个编码器层的参数，可以进一步提升 SCD-ASR 的 F1 分数。然而，在这一设置下，词错误率从 8.64% 上升到了 8.75%。这一观察结果表明，要通过自动语音识别与说话人切换检测的紧密集成，使两项任务都获得最佳性能，仍然具有挑战性。

另一方面，所提出的两种方法在说话人变更检测结果上均显著优于 TT-ASR 和 SCD-ASR 基线。具体而言，SCD-ENC 的 F1 分数分别达到 91.3% 和 81.7%，比最佳 SCD-ASR 系统相对高出 22.6% 和 14.3%，表明所提出系统的有效性。在所提出的框架中，由于提供了基于对齐的词语边界，SCD-ENC 的表现略优于 SCD-DEC。

表1。我们 180 万测试集的词错误率（%），以及 LS-ST 和 MS-CC 数据集在 TT-ASR 和 SCD-ASR 上的 F1 分数（%）。同时列出了我们提出的 SCD-ENC 和 SCD-DEC 方法的最佳性能，以供比较。

方法	Seed	更新 ASR 编码器	词错误率 (%)	F1 分数（%）↑	
			1.8M	LS-ST	MS-CC
TT-ASR	×	✓	8.65	28.9	20.0
SCD-ASR	TT-ASR	×	8.80	54.2	38.8
		✓	**8.64**	60.7	66.7
		前十	8.75	74.5	71.5
SCD-ENC	TT-ASR	×	8.65	**91.3**	**81.7**
SCD-DEC	TT-ASR	×	8.65	86.4	**81.7**

表 2。采用我们提出的方法，在 LS-ST 和 MS-CC 数据集上的 F1 分数（%）。本文探讨了不同的声学特征和词边界线索。 N_c， N_h， N_f 分别设为 8、4、4

方法	词边界	特征		F1（%）↑	
		声学	Text	LS-ST	MS-CC
SCD-ENC	对齐	Z_0	×	84.7	66.0
		Z_4	×	**91.3**	**81.7**
		Z_9	×	88.2	76.9
		Z_L	×	83.8	71.3
		Z_*	×	85.6	74.2
		Z_ω	×	**90.5**	**80.3**
		Z_4	G_W	91.2	80.4
	发射	Z_4	×	72.2	75.4
		Z_4	G_W	76.0	76.4
SCD-DEC	发射	Z_0	E_O	71.8	75.3
		Z_4	E_O	79.7	**81.7**
		Z_9	E_O	84.2	79.6
		Z_L	E_O	82.8	73.0
		Z_*	E_O	84.6	78.6
		Z_ω	E_O	**86.4**	**81.7**
	对齐	Z_ω	E_O	**90.4**	**82.3**

不过，SCD-DEC 的实现更简单，推理成本也更低，因为 SCD-ENC 中的强制对齐步骤通常计算开销较大。关于这两种实现方式的更多讨论将在以下章节中展开。

5.2. 特征探究

我们比较了声学特征候选项$Z_\chi \in \{Z_0, Z_4, Z_9, Z_L, Z_\omega, Z_*\}$和替代语义特征 G_W 在 SCD-ENC 方法中的有效性，结果见表 2。 Z_* 表示我们使用固定权重 $\omega_i = \sum_l Z_l/L$。

在 SCD-ENC 中，浅层输出 Z_4 优于其他候选输出，这验证了浅层能够捕捉更多说话人特征。从 Z_4 到 Z_L，我们观察到，随着层级变高，输出的 F1 分数呈下降趋势，从 91.3% 和 81.7% 降至 83.8% 和 71.3%，这意味着，最适合自动语音识别的声学表征未必是说话人变更检测任务的最佳特征。加权和特征 Z_ω 的表现同样具有竞争力，但略逊于 Z_4；通过分析每一层的学习权重，我们发现浅层，尤其是前四层，被赋予了更高的权重。使用固定权重会使 F1 分数显著下降，表明不同层的特征对于说话人变更检测任务而言并非同等重要。将文本嵌入特征 G_W 融入基于 Z_4 的声学特征并未带来额外提升。在提供高质量的基于对齐的词边界后，纯声学特征 Z_4 已足以支持 SCD-ENC 网络实现良好表现。

关于 SCD-DEC， Z_ω 在 LS-ST 数据集上的表现优于其他方法，并在 MS-CC 数据上取得了与 Z_4相同的 F1 分数。 Z_ω 的权重分布与 SCD-ENC 学习到的数值分布相似，这说明在我们的基于词的说话人变更检测方法中，浅层的输出贡献更大。特征 Z_0 和 Z_L 仍然表现较差，分别从 86.4% 和 81.7% 降至 71.8% 和 75.3%，以及从 82.8% 和 73.0% 降至相应数值。与 SCD-ENC 相比，表现最佳的 SCD-DEC 模型在 LS-ST 数据集上的 F1 分数相对低了 5.4%。这可能是由于词发射时间导致音频切分不准确，我们在下一节中对此进行了分析。

表 3。采用不同历史上下文和未来上下文大小时，LS-ST 和 MS-CC 数据集上的 F1 分数（%）。我们选择表 2（第 2 行和第 15 行）中的最佳模型进行评估，并设置约束 $N_h+N_f+N_c = 16$。

N_h & N_f	LS-ST		MS-CC	
	SCD-ENC	SCD-DEC	SCD-ENC	SCD-DEC
0	80.0	79.5	73.6	76.6
1	88.6	83.1	78.4	79.9
2	89.8	84.4	80.0	80.8
3	90.5	85.4	81.1	81.5
4	**91.3**	**86.5**	**81.7**	**81.7**

5.3. 词边界对比

在 SCD-ENC 中，我们采用强制对齐得到的精确词边界，为每个词构建音频嵌入特征，因此性能高度依赖于所提供边界信息的质量。在表 2 中，将 SCD-ENC 中基于对齐的词边界替换为束搜索得到的词发射时间后，性能显著下降，尤其是在 LS-ST 数据集上，从 91.2% 降至 72.2%，这是因为实际词结束时间与词发射时间之间存在延迟 [28]。在这种情况下，添加词嵌入特征 G_W 在一定程度上提高了 F1 分数，但效果仍然有限。相反，为了找出导致 SCD-ENC 与 SCD-DEC 性能差距的原因，我们尝试在 SCD-DEC 中使用高质量、基于对齐的词时间信息进行音频分块，并在 LS-ST 和 MS-CC 数据集上分别将性能从 86.4% 和 81.7% 进一步提升至 90.4% 和 82.3%。因此，如果我们能够进一步降低词发射时间的延迟，就有望通过 SCD-DEC 获得与 SCD-ENC 相近的性能。

5.4. 推理上下文大小

我们分别考察了两种实现中推理分块大小的影响，结果如表 3 所示，其中我们保持 $N_h + N_f + N_c = 16$ 不变，并为 N_h 和 N_f选择 $0 \sim 4$。显然，与第一行零上下文的结果相比，即使使用较小的上下文，例如 $N_h = N_f = 1$，也能显著提升性能。总体趋势是，上下文越大，F1 分数越高；但与此同时，由于分块之间的重叠更大，我们需要承担更高的计算成本。未来我们将尝试减少模型参数量和推理计算成本。

6. 结论

本文提出了一种说话人切换检测方法，包含两种实现方式；它们能够与我们此前研究的流式 TT-ASR 模型异步协同工作，且不会损失自动语音识别的准确率。研究表明，自动语音识别编码器的隐藏层输出表示是说话人切换检测任务的有效声学特征。与先前的联合说话人变更检测与自动语音识别模型相比，我们的方法在两个数据集上显著提升了 F1 分数。未来的工作将探索潜在的系统融合、模型大小压缩、计算成本降低，以及将应用场景扩展至多说话人场景，以支持说话人日志任务。

7. 参考文献

[1] Xavier Anguera、Simon Bozonnet、Nicholas Evans、Corinne Fredouille、Gerald Friedland 和 Oriol Vinyals，“说话人分离：近期研究综述”，*IEEE* 音频、语音与语言处理汇刊，第 20 卷，第 2 期，第 356–370 页，2012 年。[2] Tae Jin Park 、Naoyuki Kanda、Dimitrios Dimitriadis、Kyu J Han、Shinji Watanabe 和 Shrikanth Narayanan，“说话人分离综述：深度学习的最新进展”，《计算机语音与语言》，第 72 卷，第 101317 页，2022年。[3] Herv´e Bredin、Claude Barras 等，“利用双向长短期记忆网络检测广播电视中的说话人变更”，载于 *Interspeech* 会议论文集。ISCA，2017年。[4] Marek Hr´uz 和 Zbyn˘ek Zaj´ıc，“用于电话说话人日志系统的卷积神经网络说话人变更检测”，载于*ICASSP* 会议论文集。IEEE，2017年，第 4945–4949 页。[5] Zhenhao Ge、Ananth N Iyer、Srinath Cheluvaraja 和 Ar- avind Ganapathiraju，“利用神经网络说话人分类器的特征进行说话人变更检测”，载于*IntelliSys*。IEEE，2017年，第 1111–1116 页。[6] Leda Sarı、Samuel Thomas、Mark Hasegawa-Johnson 和 Michael Picheny，“用于广播新闻中低延迟说话人变更检测的说话人嵌入预训练”，载于*ICASSP* 会议论文集。IEEE，2019年，第 6286–6290 页。[7] Wei Xia、Han Lu、Quan Wang、Anshuman Tripathi、Yiling Huang、Ignacio Lopez Moreno 和 Hasim Sak，“Turn-to-diarize：受 Transformer Transducer 说话人轮次检测约束的在线说话人分离”，载于*ICASSP* 会议论文集。IEEE，2022年，第 8077–8081 页。[8] Zhiyun Fan、Zhenlin Liang、Linhao Dong、Yi Liu、Shiyu Zhou、Meng Cai、Jun Zhang、Zejun Ma 和 Bo Xu，“通过连续积分发火，利用说话人差异和语音内容进行词元级说话人变化检测”，*Interspeech* 会议论文集，第 3749–3753 页，2022年。[9] Tae Jin Park 和 Panayiotis Georgiou，“通过序列到序列神经网络，利用词汇和声学线索进行多模态说话人分段与分离”，*arXiv* 预印本 *arXiv:1805.10731*，2018年。[10] Miquel 、Angel India Massana、Jos´e Adri´an Rodr´ıguez Fonol- losa 和 Francisco Javier Hernando Peric´as，“基于 LSTM 神经网络的说话人分段：声学与语言建模”，载于*Interspeech* 会议论文集，2017年，第 2834–2838 页。[11] Nikolaos Flemotomos、Panayiotis Georgiou 和 Shrikanth Narayanan，“利用说话人角色信息进行语言学辅助说话人分离”，*arXiv* 预印本 *arXiv:1911 07994*，2019年。。[12] Jinyu Li、Yu Wu、Yashesh Gaur、Chengyi Wang、Rui Zhao 和 Shujie Liu，“大规模语音识别中主流端到端模型的比较”，*arXiv* 预印本 *arXiv:2005.14327*，2020年。[13] Yanzhang He、Tara N Sainath、Rohit Prabhavalkar、Ian Mc- Graw、Raziel Alvarez、Ding Zhao、David Rybach、Anjuli Kan- nan、Yonghui Wu、Ruoming Pang 等，“面向移动设备的流式端到端语音识别”，载于*ICASSP* 会议论文集。IEEE，2019年，第 6381–6385 页。[14] Jinyu Li，“端到端自动语音识别的最新进展”，*APSIPA* 信号与信息处理汇刊，第 11 卷，第 1 期，2022年。

[15] Bo Li、Shuo-yiin Chang、Tara N Sainath、Ruoming Pang、Yanzhang He、Trevor Strohman 和 Yonghui Wu，《迈向快速准确的流式端到端自动语音识别》，载于*Proc. I CASSP*。IEEE，2020年，第6069–6073页。[16] Shuo-yiin Chang、Bo Li、Tara N Sainath、Chao Zhang、Trevor Strohman、Qiao Liang 和 Yanzhang He，《自然会话语音的轮流发言预测》，*arXiv*预印本 *arXiv:2208.13321*，2022年。[17] Laurent El Shafey、Hagen Soltau 和 Izhak Shafran，《通过序列转导联合进行语音识别与说话人分离》，*arXiv*预印本 *arXiv:1907.05337*，2019年。[18] Huanru Henry Mao、Shuyang Li、Julian McAuley 和 Garrison Cottrell，《长对话的语音识别与多说话人分离》，*arXiv*预印本 *arXiv:2005 08072*，2020年。.[19] Naoyuki Kanda、Jian Wu、Yu Wu、Xiong Xiao、Zhong Meng、Xiaofei Wang、Yashesh Gaur、Zhuo Chen、Jinyu Li 和 Takuya Yoshioka，《采用令牌级序列化输出训练的流式多说话人自动语音识别》，*arXiv*预印本 *arXiv:2202.00842*，2022年。[20] Xie Chen、Yu Wu、Zhenghao Wang、Shujie Liu 和 Jinyu Li，《在大规模数据集上开发用于语音识别的实时流式变换器转导器》，载于*Proc. I CASSP*。IEEE，2021年，第5904–5908页。[21] Alex Graves，《使用循环神经网络进行序列转导》，*arXiv*预印本 *arXiv:1211.3711*，2012年。[22] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin，《注意力机制即所需的一切》，神经信息处理系统进展，第30卷，2017年。[23] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 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，2022年。[24] Rui Zhao、Jian Xue、Jinyu Li、Wenning Wei、Lei He 和 Yifan Gong，《关于解决循环神经网络转导器的实际挑战》，载于*ASRU*。IEEE，2021年，第526–533页。[25] Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaiming He 和 Piotr Doll´ar，《用于密集目标检测的焦点损失》，载于*IEEE*计算机视觉国际会议论文集，2017年，第2980–2988页。[26] Z huo Chen、Takuya Yoshioka、Liang Lu、Tianyan Zhou、Zhong Meng、Yi Luo、Jian Wu、Xiong Xiao 和 Jinyu Li，《连续语音分离：数据集与分析》，载于*Proc. ICASSP*。IEEE，2020年，第7284–7288页。[27] Dongmei Wang、Takuya Yoshioka、Zhuo Chen、Xiaofei Wang、Tianyan Zhou 和 Zhong Meng，《使用临时麦克风阵列进行连续语音分离》，载于*EUSIPCO*。IEEE，2021年，第1100–1104页。[28] Jiahui Yu、Chung-Cheng Chiu、Bo Li、Shuo-yiin Chang、Tara N Sainath、Yanzhang He、Arun Narayanan、Wei Han、Anmol Gulati、Yonghui Wu 等，《FastEmit：通过序列级发射正则化实现低延迟流式自动语音识别》，载于*Proc. ICASSP*。IEEE，2021年，第6004–6008页。