arXiv:2302.07521v1 [eess.AS] 15 Feb 2023

基于置信度分数的 Conformer语音识别系统说话人自适应

Jiajun Deng、Xurong Xie、Tianzi Wang、Mingyu Cui、Boyang Xue、Zengrui Jin、Guinan Li、Shujie Hu、Xunying Liu，*IEEE* 会员

摘要——**说话人自适应技术为针对个体用户定制自动语音识别（自动语音识别）系统提供了强有力的解决方案。将无监督的、基于模型的说话人自适应技术应用于数据密集型端到端语音识别系统，会受到说话人级数据匮乏以及系统性能对转录错误敏感的限制。为解决这些问题，本文采用一组紧凑且数据高效的说话人相关（SD）参数表示，以支持最先进的 Conformer 自动语音识别系统的说话人自适应训练和测试时无监督说话人自适应。通过基于置信度分数选择说话人级自适应数据中错误较少的子集，降低了系统对监督质量的敏感性。本文提出了两个轻量级置信度分数估计模块，以生成更可靠的置信度分数。数据选择会加剧数据稀疏性问题；为此，本文使用贝叶斯学习对 SD 参数不确定性进行建模。在基准 300小时 Switchboard 语料库和 233小时 AMI 数据集上的实验表明，所提出的基于置信度分数的自适应方案始终优于基线说话人无关（SI）Conformer模型，以及未进行说话人数据选择的传统非贝叶斯、基于点估计的自适应。在采用外部 Transformer 和长短期记忆网络语言模型重打分后，仍然保持了类似的性能提升。具体而言，在 300小时 Switchboard 语料库上，相较于基线 SI Conformer，在 NIST Hub5'00、RT02 和 RT03 评估集上，词错误率分别绝对降低了 1.0%、1.3% 和 1.4%（相对降低 9.5%、10.9% 和 11.3%），且具有统计显著性。在 AMI 开发集和评估集上，也分别实现了绝对降低 2.7% 和 3.3%（相对降低 8.9% 和 10.2%）的类似词错误率降低。**

索引词—**语音识别、说话人自适应、置信度分数估计、贝叶斯学习、Conformer**

一、引言

THE 过去几十年来，随着深度学习技术的广泛应用，自动语音识别（ASR）系统的性能显著提升[1]–[15]。近年来，语音技术领域的一大趋势是从混合式自动语音识别系统[1]–[3] 转向端到端（E2E）建模 [4]–[8], [10], [11]，即利用单个神经网络将声学特征向量直接映射为表层词序列或词元序列。以连接时序分类（CTC） [16]，、听、关注与拼写（LAS） [4]或基于注意力的编码器-解码器（AED） [17]，、循环神经网络转导器（RNN-T）模型 [18], [19]，以及 Transformer [7], [20]或卷积增强 Transformer（Conformer） [11]，为代表的端到端自动语音识别系统，采用与语音识别错误率一致的单一目标函数进行优化。在这些端到端模型中，Conformer 编码器结构结合了自注意力与卷积模块，旨在同时学习较长距离的全局上下文和局部声学特征。这些优势也促使其在 Conformer-Transformer AED 和 Conformer 变换器中得到广泛应用 [11]–[13], [21]–[23]。

自动语音识别系统（包括基于 Conformer 的系统）面临的一项关键挑战，是对语音中系统性且潜在的变化进行建模。这类变化的一个主要来源是说话人层面特征所代表的因素，例如口音和个体特异性，或年龄、性别等因素所体现的生理差异。说话人之间的差异往往会导致训练数据与评估数据的分布显著不匹配，从而降低识别性能。为此，旨在减少说话人间声学差异所造成不匹配的说话人自适应技术，在当前语音识别系统中发挥着至关重要的作用。第 I-A 节和第 I-B 节分别介绍传统隐马尔可夫模型、混合式自动语音识别系统和端到端自动语音识别系统中的说话人自适应相关综述。

A. 隐马尔可夫模型和混合式自动语音识别系统的说话人自适应

针对隐马尔可夫模型和混合式自动语音识别系统设计的说话人自适应技术可分为三大类：**1)** 辅助的基于说话人嵌入的方法[24]–[27]。**2)** 基于特征变换的方法 [28]–[33]。**3)**基于模型的自适应方法 [34]–[59]。

在辅助的基于说话人嵌入的方法中，会将为每位说话人计算出的紧凑型说话人相关（SD）嵌入向量添加到声学特征中，以便在自动语音识别系统训练和评估期间进行说话人自适应。辅助说话人嵌入向量既可以独立于后端识别系统进行估计，例如从基于高斯混合模型（GMM）的通用背景模型中学习得到的 i-vector [25], [26]；也可以与声学模型的参数联合估计，例如通过在辅助说话人嵌入向量与后端声学模型参数之间交替更新来估计说话人编码 [24]。

在基于特征变换的方法中，会对声学前端应用特征空间变换，以生成说话人不变的输入特征，随后将其馈送至后端

Jiajun Deng、Tianzi Wang、Mingyu Cui、Boyang Xue、ZengruiJin、Guinan Li 和 Shujie Hu 就职于中国香港中文大学（电子邮箱：{jjdeng, twang, mycui, byxue, zrjin, gnli, sjhu}@se.cuhk.edu.hk）。Xurong Xie 就职于中国科学院软件研究所（中国北京，电子邮箱：xurong@iscas.ac.cn）。Xunying Liu 就职于中国香港中文大学（电子邮箱：xyliu@se.cuhk.edu.hk）。通讯作者：Xurong Xie 和 Xunying Liu。

识别系统。估计特征空间变换的一种常用方法是特征空间最大似然线性回归（f-MLLR） [29]。f-MLLR 变换基于 GMM-HMM 系统[28], [29] 在说话人级进行估计，然后应用于声学前端，以规范说话人层面变异 [33]。为补偿说话人之间的声道长度差异，还可以使用基于生理学原理的声道长度归一化（VTLN）特征变换，生成类似说话人不变的特征 [30]–[32]。通过应用分段线性频率扭曲因子 [31], [32]对谱特征进行归一化，即可获得经过说话人级 VTLN 归一化的特征。

在基于模型的自适应方法中，说话人相关建模粒度从更新全部声学模型参数 [34], 到仅更新一小部分参数不等，这在很大程度上取决于可用目标说话人自适应数据的具体数量。为确保良好的泛化能力，并防止在实际应用中过拟合有限的自适应数据，通常会基于结构化变换，单独设计紧凑形式的说话人相关参数，并将其应用于各种神经网络层。这些方法包括使用线性输入网络（LIN） [46], [48], 、线性隐藏网络（LHN） [47],、线性输出网络（LON） [47], 、隐藏单元贡献学习（LHUC） [41]、参数化激活函数（PAct）[45], 、因式分解线性变换 [49], 、基于说话人编码的偏置向量 [37],，以及对说话人无关（SI）隐藏层输出或参数基进行自适应线性插值 [38]–[40]。此外，还可以进一步将说话人自适应训练（SAT） [50], [51]与上述基于结构化变换的自适应方法结合。SAT在训练阶段联合优化说话人相关参数和说话人无关参数，从而得到一个规范模型，使其在测试时自适应中能够更好地适应未见说话人。为进一步确保说话人相关参数估计的稳健性，研究者已在混合自动语音识别系统中探索对说话人自适应目标函数进行正则化的方法，包括使用L2 [52]或Kullback–Leibler（KL）散度正则化 [34]、最大后验自适应 [58], [59], ，以及基于变分推断的贝叶斯自适应（对说话人相关参数的不确定性进行建模） [43], [44]。此外，研究者还探索了其他说话人自适应目标函数，包括使用对抗学习 [53], [54] 和多任务学习[57]。

B. 端到端自动语音识别系统的说话人自适应

相比之下，针对端到端自动语音识别系统的说话人自适应研究还较为有限。其中，基于 i-vector [21],[23], x-vector [23], [60]和 f-MLLR [60], 的辅助说话人自适应特征，或从序列摘要网络 [61], 、基于注意力的说话人记忆 [62] 和说话人感知模块 [63], [64] 中提取的特征，被整合到基于注意力的编码器-解码器模型中，或整合到常规的、未采用卷积增强的 Transformer 模型中。若干针对端到端模型的基于模型的说话人自适应方法采用说话人自适应神经网络的内部组件，例如多通道端到端系统中的说话人相关神经波束形成、编码器、注意力或解码器模块 [65], ，同时在自适应过程中固定其他组件的参数，使其保持说话人无关。类似方法也被应用于 RNN-T [66], [67] 和 Conformer 转导模型 [22]。为降低自适应过程中因说话人级数据有限而导致过拟合的风险，研究者提出了 KL散度正则化 [68]–[70]以及通过文本转语音合成进行目标说话人数据增强 [67], [71] 的方法。多任务学习 [68], [69]和对抗学习 [69]则已应用于基于 CTC 或基于注意力的端到端模型。

为 Conformer 模型开发基于模型的说话人自适应技术面临着若干挑战。首先，说话人特定数据通常数量有限，因此需要采用高度紧凑的 SD参数表示，以降低自适应过程中发生过拟合的风险。其次，当以无监督模式执行 LHUC [41], [42] 等基于模型的自适应方法时，SD参数估计对底层监督错误率 [72], 十分敏感。在这种情况下，由于缺少自适应数据的参考转写，通常需要采用多遍解码框架。首次识别产生初始语音转录，作为后续 SD参数估计以及第二遍语音重新解码的监督信息。监督假设的错误率所导致的敏感性，对于端到端系统尤其具有挑战性，因为这类系统直接学习自适应监督中的表层词或词元序列标签 [73], ，而混合自动语音识别系统自适应则使用潜在语音目标。

为解决这些问题，本文采用 LHN [47], LHUC [41], [42],隐藏单元偏置向量（HUB） [37]和 PAct [45] 形式的紧凑型 SD 参数表示，以便对 ESPnet Conformer 模型进行说话人自适应训练和测试时无监督说话人自适应。通过基于置信度分数的选择，获取更“可靠”的说话人专属自适应数据子集，从而降低对监督质量的敏感性。在 Conformer 模型中，基于注意力的编码器-解码器模型架构和自回归解码器组件都会利用完整的历史上下文。因此，使用传统的基于格的识别假设表示来估计词后验概率和置信度分数并非易事 [74],[75]。为此，本文提出两种轻量级置信度分数估计模块（CEM），在将过度自信的 Conformer 解码器输出概率用作置信度分数之前对其进行平滑处理。基于置信度分数的说话人级数据选择会加剧数据稀疏性问题；为解决这一问题，本文以潜变量后验分布取代 SD 参数，并通过贝叶斯学习对参数不确定性进行建模 [44], [76], [77]。

实验在两项任务上进行：a) 300小时 Switchboard 电话语料库 [78]；以及 b) 233小时 AMI 会议语音转录数据集 [79]。实验结果表明，所提出的基于置信度分数的选择方法，从说话人级别选取数量不等的子集数据，其 LHUC-SAT 自适应性能与基于预言机真实词错误率（WER）进行选择的结果相当。所提出的基于置信度分数的自适应方案始终优于基线说话人无关模型，以及使用 LHUC、HUB、PAct 进行的传统非贝叶斯、基于点估计的自适应，

或 PAct。经过外部 Transformer 和长短期记忆网络语言模型重打分后，仍保持了类似的性能提升。特别是，在 300小时 Switchboard 语料库上，与基线 SI Conformer模型相比，在 NIST Hub5'00、RT02 和 RT03 评估集上，词错误率分别绝对降低了 1.0%、1.3% 和 1.4%（相对降低 9.5%、10.9% 和 11.3%），且具有统计显著性[1] 。在使用个人头戴式麦克风（ihm）、单远场麦克风（sdm）和多远场麦克风（mdm）条件录制的 AMI 开发集和评估集上，词错误率也分别绝对降低了 2.7% 和 3.3%（相对降低 8.9% 和 10.2%）。

本文的主要贡献总结如下：1）据我们所知，本文是首篇针对端到端Conformer自动语音识别系统研究数据高效、紧凑的基于模型的说话人自适应方法的工作。相比之下，先前关于基于结构化变换的说话人自适应的研究仅限于传统的混合DNN-HMM自动语音识别系统 [36]–[49]。另一方面，现有针对端到端自动语音识别系统设计的说话人自适应方法，主要侧重于使用辅助说话人自适应特征 [21], [23], [60]–[64], ，或直接自适应说话人无关模型的全部或部分组件 [22], [65]–[67]。这些数据密集型基于模型的自适应方法所面临的说话人级数据稀疏性问题，基本上仍未得到解决。先前的研究还考察了在端到端系统中使用辅助的基于说话人嵌入的方法[21], [23], [60], [61]。然而，结果表明，这些方法在说话人自适应方面的表现并不一致，例如，在 Switchboard 任务上报告的 IBM Conformer 系统就是如此 [21]。

2）本文首次研究了基于置信度分数的无监督Conformer模型说话人自适应。研究设计了两种分别在话语级或词元级运行的置信度估计模块（CEM），用于平滑过度自信的 Conformer 解码器输出概率，之后将其作为置信度分数，以筛选出说话人级自适应数据中更"可信"的子集。本文还通过一系列消融研究，探讨了构建 CEM 时模型配置以及输入特征选择等具体实现问题。相比之下，类似端到端自动语音识别系统的解码器输出概率平滑方法仅被用于置信度分数估计 [81]–[84], ，而未用于本文所研究的 Conformer自动语音识别系统说话人自适应。

3）据我们所知，本文所 प्रस्तुत的研究还首次探讨了利用完整贝叶斯学习实现端到端Conformer自动语音识别系统的说话人自适应，以解决说话人级数据稀疏性问题。研究使用变分推断对 SD参数不确定性进行显式建模并加以考虑。相比之下，先前的研究主要采用 KL散度正则化 [68]–[70]以及目标说话人数据增强 [67], [71] （通过文本转语音合成），以解决端到端系统说话人自适应和个性化过程中的类似数据稀疏问题。

4) 所提出的基于置信度分数的 Conformer 自适应方法的优势始终得到证实——这一优势。与文献中报道的、在 300 小时 Switchboard 任务上采用最新混合系统和端到端自动语音识别系统所取得的最先进性能相比，这些方法的有效性和竞争力也得到了进一步证明。此外，本文所述方法具有通用性，且附带的实现细节使其能够进一步应用于其他流行的端到端自动语音识别架构，例如 Conformer 变换器 [22],；同时，也为语音处理任务之外更普遍的参数高效和数据高效 Transformer 模型适应提供了进一步见解，例如，在自然语言处理中微调预训练语言模型时使用任务领域相关模块 [85]。

本文其余部分安排如下。第 II 节回顾 Conformer 自动语音识别系统。第 III 节介绍基于结构化隐藏层变换的说话人适应的各种形式及其在 Conformer 模型中的应用。第 IV 节提出基于置信度得分的说话人自适应方案和贝叶斯估计。第 V 节讨论一系列关键实现细节。第 VI 节展示实验结果并进行分析。最后，第 VII 节总结全文并讨论未来研究工作。

II. CONFORMER 端到端自动语音识别系统

Conformer [11], [12] 由 Conformer 编码器和 Transformer 解码器组成。Conformer 编码器包含一个卷积子采样模块，以及一个带有丢弃操作的线性层，其后连接多个堆叠的编码器块。每个 Conformer 编码器块的内部组件包括：一个逐位置前馈网络（FFN）模块、一个多头自注意力（MHSA）模块、一个卷积（CONV）模块，以及末尾的一个逐位置 FFN 模块。其中，CONV 模块还包含以下各层：一个一维逐点卷积层、一个门控线性单元（GLU）激活函数、第二个一维逐点卷积层、一个一维深度卷积层、一个 Swish 激活函数，以及最后一个一维逐点卷积层。所有编码器块还都经过层归一化，并采用残差连接。图 1 展示了一个 Conformer 端到端自动语音识别系统示例：中间的灰色方框展示了整体模型架构，右上方的绿色方框展示了编码器块的组成。左上方展示了可供比较的 Transformer 编码器块。

在 Transformer 和 Conformer 模型的训练中，采用以下 CTC 与注意力错误代价之间的多任务准则插值 [5] 。

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{att} + \lambda\mathcal{L}_{ctc}, \tag{1}$$

其中 $\lambda \in [0,1]$ 是可调超参数；本文通过经验将其设为训练时的 0.2 和识别时的 0.3。

III. Conformer说话人自适应

本节介绍四种基于结构化变换的说话人自适应方法，分别基于 LHN、LHUC、HUB 和 PAct，并介绍它们与说话人自适应训练（SAT）的进一步结合，应用于 ConformerASR系统。

[1]采用匹配句段词错误（MAPSSWE）进行统计显著性检验， [80] 显著性水平为 α=0.05。

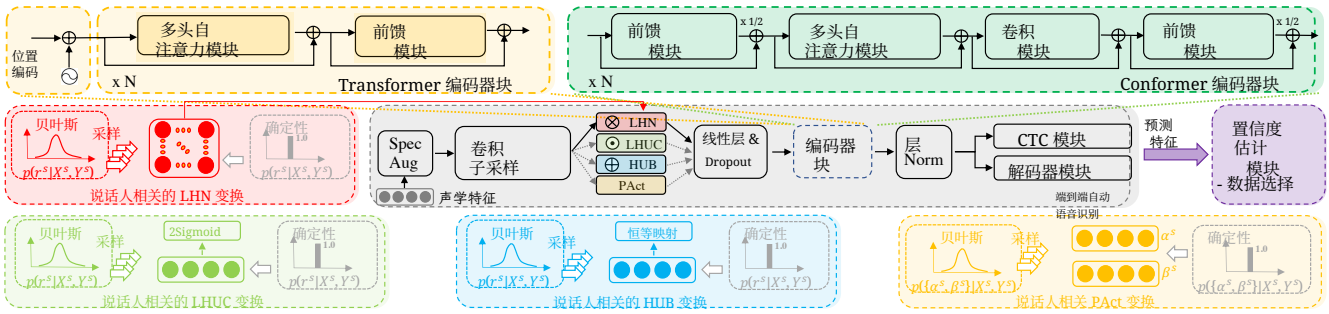


图1。Transformer 和 Conformer 端到端模型的示例显示在灰色框中（中间中央）。相应的 Transformer 和 Conformer 编码器显示在黄色框（左上）和绿色框（右上）中。LHN、LHUC、HUB 和 PAct 的说话人相关变换分别显示在红色框（中间左侧）、浅绿色框（左下角）、蓝色框（底部中央）和浅黄色框（右下角）中，其中贝叶斯估计或固定值、确定性估计分别显示在每个框内的左侧和右侧。

A. 线性隐藏网络

在线性隐藏网络（LHN）基础上的 Conformer说话人自适应中，SD 仿射变换被插入 Conformer编码器模块的隐藏层，其方式类似于采用相似方法对混合DNN自动语音识别系统进行自适应[47]。说话人 s 在第 l 个隐藏层中的 LHN 变换由一个线性矩阵 $\mathbf{A}^{l,s}$ 和一个偏置 $\boldsymbol{b}^{l,s}$ 参数化。第 l 个自适应隐藏层的输出可表示为

$$\boldsymbol{h}^{l,s} = \mathbf{A}^{l,s}\boldsymbol{h}^{l} + \boldsymbol{b}^{l,s}, \tag{2}$$

其中，$\boldsymbol{h}^{l}$ 是第 l层经过非线性激活函数后的隐藏向量。执行自适应时，线性矩阵 $\mathbf{A}^{l,s}$ 和偏置 $\boldsymbol{b}^{l,s}$初始化为单位矩阵和零向量。图1展示了一个基于 LHN 的 Conformer 说话人自适应示例（红色框，中间左侧）。然而，在有限的自适应数据条件下，直接使用包含数百万参数的大型完整矩阵说话人相关线性变换不仅容易过拟合，还会导致计算效率问题。为此，以下小节介绍三种更紧凑的说话人相关变换。

B. 隐藏单元贡献学习

隐单元贡献学习（LHUC）自适应的核心思想是将线性 SD 变换$\mathbf{A}$限制为对角矩阵，以减少 SD 参数的数量；这相当于使用缩放向量来调整激活输出的幅度。设 $\boldsymbol{r}^{l,s}$ 为说话人 s 在第 l 个隐藏层中的 SD 参数。经 LHUC 自适应后的隐藏层输出可表示为

$$\boldsymbol{h}^{l,s} = diag(\mathbf{A}^{l,s})\boldsymbol{h}^{l} = \xi(\boldsymbol{r}^{l,s}) \odot \boldsymbol{h}^{l}, \tag{3}$$

其中，$\odot$ 表示 Hadamard 乘积运算，$\xi(\boldsymbol{r}^{l,s})$ 是由 $\boldsymbol{r}^{l,s}$参数化的缩放向量。本文中，LHUC 自适应采用 $\xi(\cdot)$，即值域为 (0,2) 的逐元素 2Sigmoid(·) 函数。说话人自适应期间，SD 参数 $\boldsymbol{r}^{l,s}$ 初始化为零向量。图 1 中展示了一个基于 LHUC 的 Conformer 说话人自适应示例（左下角的绿色方框）。

C. 隐单元偏置

或者，将SD参数 $\boldsymbol{r}^{l,s}$ 限制为一个加到隐藏层输出上的偏置向量，就会得到隐单元偏置（HUB）适配，类似于说话人编码[37]。令 $\xi(\boldsymbol{r}^{l,s})$ 表示由SD参数 $\boldsymbol{r}^{l,s}$ 参数化的偏置向量，该向量对应说话人 s 在第 l个隐藏层中的偏置。经过HUB适配的隐藏层输出为

$$\boldsymbol{h}^{l,s} = \xi(\boldsymbol{r}^{l,s}) + \boldsymbol{h}^{l}. \tag{4}$$

本文采用恒等激活作为 HUB 自适应的偏置激活函数 $\xi(\cdot)$。SD参数$\boldsymbol{r}^{l,s}$ 初始化为零向量。图 1 中展示了一个基于 HUB 的 Conformer 说话人自适应示例（底部中央的蓝色方框）。

D. 参数化激活函数

本文最后一种说话人自适应方法修改的是隐藏层激活函数的输入，而不是其输出向量。考虑到 [45] 为混合 DNN 系统提出的参数化 ReLU（PReLU）激活函数，本文将说话人自适应后的第 l 个 Conformer 隐藏层 PReLU 激活输出计算如下：

$$h_n^{l,s} = \begin{cases} \alpha_n^{l,s} z_n^{l}, & z_n^{l} \geq 0; \\ \beta_n^{l,s} z_n^{l}, & z_n^{l} < 0, \end{cases} \tag{5}$$

其中，z_n^{l} 是激活输入 $\boldsymbol{z}^{l}$ 的第 n 个元素；$\alpha^{l,s}_n$ 和 $\beta^{l,s}_n$ 分别是 SD 参数 $\boldsymbol{r}^{l,s}=\{\boldsymbol{\alpha}^{l,s},\boldsymbol{\beta}^{l,s}\}$ 中的第 n 个元素，分别对应正值和负值 z_n^{l} 的斜率。SD 参数在自适应过程中初始化为 $\boldsymbol{\alpha}^{l,s}=\mathbf{1}$ 和 $\boldsymbol{\beta}^{l,s}=\mathbf{0}$，这可视为标准 ReLU 激活的情形。基于 PAct 的 Conformer 说话人自适应示例如图 1 所示（右下角黄色方框）。

E. SD 参数估计

令 $\mathcal{D}^{s}=\{\boldsymbol{X}^{s},\boldsymbol{Y}^{s}\}$ 表示说话人 s的适配数据集，其中 $\boldsymbol{X}^{s}$ 和 $\boldsymbol{Y}^{s}$ 分别表示声学特征及其对应的监督令牌序列。为不失一般性并简化表述，在测试时自适应的语境下省略 Conformer 说话人无关参数。通过最小化公式（1）中的损失，可以估计 SD参数 $\boldsymbol{r}^{s}$，其表达式为

$$\hat{\boldsymbol{r}}^{s} = \arg\min_{\boldsymbol{r}^{s}} \{-\lambda_1 \log p_a(\boldsymbol{Y}^{s}|\boldsymbol{X}^{s},\boldsymbol{r}^{s}) - \lambda \log p_c(\boldsymbol{Y}^{s}|\boldsymbol{X}^{s},\boldsymbol{r}^{s})\}, \tag{6}$$

其中， $\lambda_1 = 1-\lambda$、 p_a和 p_c分别为基于注意力和 CTC 的序列概率。在无监督测试时自适应过程中，需要先使用基线说话人无关模型对相应的说话人特定数据进行初始解码，以生成未见测试说话人s 的监督 $\boldsymbol{Y}^s$，随后将其作为自适应阶段中的目标词元标签。因此，SD 参数的估计对底层监督错误率较为敏感。

在说话人自适应训练过程中，与训练说话人相关的一组说话人相关参数会与其余说话人无关参数一同进行优化 $\boldsymbol{\theta}_{SAT}^{SI}$，优化依据为式（1）中的损失函数。该过程可改写为

$$\{\hat{\boldsymbol{\theta}}^{SD}, \hat{\boldsymbol{\theta}}_{SAT}^{SI}\} = \underset{\boldsymbol{\theta}^{SD}, \boldsymbol{\theta}_{SAT}^{SI}}{\arg\min} \sum_{s\in\mathcal{S}} \mathcal{L}(\mathcal{D}^s; \boldsymbol{\theta}^{SD}, \boldsymbol{\theta}_{SAT}^{SI}), \quad (7)$$

其中， $\mathcal{S}$ 为训练说话人集合， $\boldsymbol{\theta}^{SD} = \{\boldsymbol{r}^s\}_{s\in\mathcal{S}}$为说话人相关参数集。这种联合训练形式能够生成说话人归一化程度更高的规范化说话人自适应训练模型，从而在测试时自适应过程中更好地适配未见说话人。

IV. 基于置信度分数的说话人自适应

本节介绍基于置信度分数的Conformer自适应。本文还提出了基于贝叶斯学习的说话人自适应方法，以应对基于置信度分数的数据选择所导致的数据稀疏性增加问题。

A. 基于置信度分数的数据选择

为了降低 Conformer 说话人自适应的性能对监督错误率的敏感性，本研究探索的一种解决方案是基于置信度分数进行选择，从说话人特定的自适应数据中筛选出一个更“可信”、准确且错误更少的子集。置信度分数可用于衡量语音识别系统输出的可靠性。它们已广泛应用于传统的基于 HMM 的自动语音识别系统中的说话人自适应 [86]–[89]，无监督训练 [90]–[92],和半监督学习 [93] 。

基于置信度分数的数据选择的有效性在很大程度上取决于其与词错误率的相关性。研究人员已考察用于语音识别的多种置信度度量形式，这些度量分别在帧 [94],音素[95], 词元或词 [96] 以及话语 [97] 级别进行计算。传统的基于 HMM 的自动语音识别系统通常使用基于格的识别输出，来估计词或词元的后验概率和置信度分数 [74], [98], [99]。研究人员也考察了更先进的置信度分数估计方法，这些方法通过专门设计的置信度分数估计器（例如条件随机场和循环神经网络）整合更丰富的、从格中提取的特征 [100]–[102]。

然而，在许多采用注意力编码器-解码器模型架构和自回归解码器设计的当前端到端自动语音识别系统中（包括本文研究的 Conformer 模型），推断时通常需要完整的历史上下文。这给在基于格的解码中实现有效的束搜索带来了挑战，除非采用基于截断历史等近似方法 [103]。另一种解决方案是直接将 Conformer解码器输出概率，或将其与 CTC 概率结合后，解读为置信度分数。然而，事实证明，这些输出概率过度自信 [82], [83] ，不能直接用作识别准确率的可靠指标。为解决这一问题，本文设计了两种基于神经网络的置信度评估模型，分别生成话语级或词元级置信度分数，以平滑过度自信的 Conformer解码器输出概率；随后在测试时无监督自适应过程中，利用这些分数从说话人级数据中选出更准确、错误更少的子集。图 2 展示了一个置信度评估模型示例，其中包含一个正确的识别假设和一个错误的识别假设。解码器原本为正确的识别假设（右上方的红色矩形）和错误的识别假设（右上方的深绿色矩形）分配的输出概率分别为 0.53（“置信不足”）和 0.93（“过度自信”）。使用置信度评估模型平滑这些输出概率后，正确识别假设（右下方的红色矩形）的概率提高至 0.81，而错误识别假设（右下方的深绿色矩形）的概率则降至 0.62。与原始解码器输出概率相比，这些概率可视为更可靠的置信度度量。

1) 话语级 *CEM*：如图 2 所示，话语级 CEM 是一种轻量级二元分类模型，包含一个简单而有效的三层残差前馈网络，隐藏维度为 64，并采用 Sigmoid输出层。每个隐藏层依次应用批归一化、ReLU 激活和 Dropout 神经操作。第一层和第二层隐藏层的输出之间还采用了跳跃连接。随后，该 CEM 与自动语音识别系统（例如 SI Conformer模型）相连。在识别过程中，从 Conformer模型的不同内部组件提取的一组预测特征被输入 CEM，后者随后为当前话语 k预测一个平滑置信度分数 c_k 。值得注意的是，预测特征在鲁棒置信度评估模型的设计中发挥着重要作用，预测特征选择的更多细节将在第 V-A 节中介绍。为生成用于 CEM训练的话语级置信度分数目标标签，使用未经适配的说话人无关 Conformer对训练数据进行束搜索解码，从而生成识别假设。如果话语 k 的识别假设与相应的真实参考文本完全匹配，则为其二元目标标签 l_k 赋值 1，否则赋值 0。使用置信度分数 $\boldsymbol{c}$与目标标签 $\boldsymbol{l}$ 之间计算得到的交叉熵误差代价来优化 CEM。

$$\mathcal{L}_c = \sum_{k\in\mathcal{B}} (\eta l_k \log(c_k) + (1-\eta)(1-l_k)\log(1-c_k)), \quad (8)$$

其中， $\mathcal{B}$ 是训练样本集， $\eta \in [0,1]$ 是加权因子，在本文中根据经验设为 0。3。

2）词元级 *CEM*：词元级 CEM 与句子级 CEM 采用相同的总体架构，而 Sigmoid输出层则基于更细粒度的词元级特征，生成一个置信度分数，用于表示当前词元正确的概率。由于说话人级别的数据选择是在话语级进行的，因此，句子级

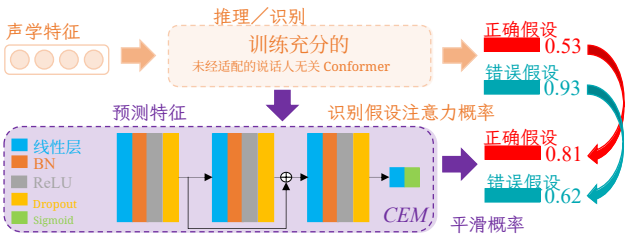


图2。紫色框中展示了一个 CEM架构示例。预测特征从 Conformer模型（浅橙色框）的不同内部位置提取，如第 V-A 节所述。

置信度分数可以通过对每个话语内的词元级置信度度量取平均值来获得。

词元级 CEM训练的目标标签确定方法如下。首先使用编辑距离生成识别假设与真实参考文本之间的对齐。根据该对齐，将正确词元标记为 1，将替换或插入的词元标记为 0。例如，如果真实参考文本的词元序列为“C U H K”，对应的识别假设为“C U L K”，那么标签序列为 $\boldsymbol{l} = [1, 1, 0, 1]$。随后对生成的二元标注CEM训练数据进行下采样，以平衡置信度分数为 1 和 0 的词元数量[2]。这里同样使用与公式（8）类似的二元交叉熵损失，但损失是在词元级别计算的。

B. 贝叶斯适应学习方法

尽管基于置信度分数的数据选择能够减轻监督错误对自适应造成的负面影响，但它会进一步减少本已有限的说话人级自适应可用数据量，例如每位说话人仅有几分钟的 Switchboard 会话语音，并加剧数据稀疏问题。这会导致标准的基于点估计的模型适应过程中出现SD参数不确定性，即便是在第III节中更紧凑的LHUC、HUB和PAct自适应方法中也是如此。为了解决这一问题，本文采用贝叶斯学习[44], [76], [77]，通过后验分布对参数不确定性进行建模。测试话语 $\boldsymbol{X}^s$ 的贝叶斯预测分布通过对SD参数取值进行积分得到，如下所示：

$$p(\tilde{\boldsymbol{Y}}^s|\tilde{\boldsymbol{X}}^s, \mathcal{D}^s) = \int p(\tilde{\boldsymbol{Y}}^s|\tilde{\boldsymbol{X}}^s, \boldsymbol{r}^s) p(\boldsymbol{r}^s|\mathcal{D}^s) d\boldsymbol{r}^s, \quad (9)$$

其中，$\tilde{\boldsymbol{Y}}^s$ 是预测的词元序列，$p(\boldsymbol{r}^s|\mathcal{D}^s)$是根据说话人 s的自适应数据学习得到的SD参数后验分布。通过对自适应数据 $\mathcal{D}^s$ 进行边缘化处理，优化混合注意力与CTC损失，以推断SD潜变量 $\boldsymbol{r}^s$的后验分布。

$$\begin{aligned}\mathcal{L}(\mathcal{D}^s) &= -\lambda_1 \log p_a(\boldsymbol{Y}^s|\boldsymbol{X}^s) - \lambda \log p_c(\boldsymbol{Y}^s|\boldsymbol{X}^s) \\ &= \int \{-\lambda_1 \log p_a(\boldsymbol{Y}^s, \boldsymbol{r}^s|\boldsymbol{X}^s) - \lambda \log p_c(\boldsymbol{Y}^s, \boldsymbol{r}^s|\boldsymbol{X}^s)\} d\boldsymbol{r}^s.\end{aligned} \quad (10)$$

[2]例如，300 小时的 Switchboard 训练集共包含 1,267,693 个词元，其中 1,163,601 个和 104,092 个分别标记为 one 和 zero。CEM 训练中使用的 one 和 zero 标记词元数量分别为 416,368 和 104,092。

常用的变分推断用于学习上述积分中的 SD 参数后验分布。通过最小化由詹森不等式推导出的式（10）中积分的以下上界，引入变分分布 $q(\boldsymbol{r}^s)$ 来近似后验分布 $p(\boldsymbol{r}^s|\mathcal{D}^s)$，其形式如下：

$$\begin{aligned}\mathcal{L}(\mathcal{D}^s) \leq &\int q(\boldsymbol{r}^s)\{-\lambda_1 \log p_a(\boldsymbol{Y}^s|\boldsymbol{X}^s, \boldsymbol{r}^s) - \lambda \log p_c(\boldsymbol{Y}^s|\boldsymbol{X}^s, \boldsymbol{r}^s)\} d\boldsymbol{r}^s \\ &+ \mathrm{KL}(q(\boldsymbol{r}^s)||p(\boldsymbol{r}^s)) \triangleq \mathcal{L}_1 + \mathcal{L}_2,\end{aligned} \quad (11)$$

其中，$p(\boldsymbol{r}^s)$表示SD参数的先验分布，KL(·)表示KL散度。为确保积分$\mathcal{L}_1$ 可微，采用带重参数化技巧的蒙特卡洛采样方法来近似项 $\mathcal{L}_1$，

$$\mathcal{L}_1 \approx \frac{1}{N}\sum_{n=1}^{N} -\lambda_1 \log p_a(\boldsymbol{Y}^s|\boldsymbol{X}^s, \boldsymbol{r}_n^s) - \lambda \log p_c(\boldsymbol{Y}^s|\boldsymbol{X}^s, \boldsymbol{r}_n^s), \quad (12)$$

其中，$\boldsymbol{r}_n^s = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}_n$ 和 $\boldsymbol{\epsilon}_n$ 是从标准正态分布中抽取的第 n个蒙特卡洛采样向量。为简单起见，假设 $q(\boldsymbol{r}^s)$ 和 $= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$以及$p(\boldsymbol{r}^s)$和 $= \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\sigma}}^2)$均服从正态分布。KL散度 $\mathcal{L}_2$ 可显式表示为闭式形式，

$$\mathcal{L}_2 = \frac{1}{2}\sum_k \left(\frac{\sigma_k^2 + (\mu_k - \bar{\mu}_k)^2}{\bar{\sigma}_k^2} + 2\log\frac{\bar{\sigma}_k}{\sigma_k} - 1\right), \quad (13)$$

其中，$\{\bar{\mu}_k, \bar{\sigma}_k\}$ 和 $\{\mu_k, \sigma_k\}$ 分别是对应向量 $\{\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\sigma}}\}$ 和 $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$中的第 k个元素。

V. 实现细节

本节讨论了影响所提出的基于置信度分数的 Conformer说话人自适应方法的性能和效率的一系列重要实现问题。

A. CEM 的特征选择

设计鲁棒置信度评估模型的一个关键问题，是选择信息量最大的预测特征，以便更好地区分正确的识别假设与错误假设。受此前端到端自动语音识别系统置信度分数估计研究的启发，本文研究了从 Conformer 自动语音识别系统的不同组件中提取的各种神经特征，用于 CEM 训练。这些特征包括：**1）内部嵌入特征**，从编码器和解码器模块的最后几层提取；**2）1-best输出分数**，在束搜索过程中由解码器、CTC 模块单独或二者组合生成，针对当前考察的 1-best词元或话语级假设；**3）N-best输出分数**，作为第 2）项的扩展，可利用关于竞争假设或词元之间混淆建模的额外信息；以及**4）特征融合**，融合上述第 1）至第 3）项中的特征。第 VI-A3 节将详细介绍针对 CEM 训练和置信度分数估计的特征选择消融研究。

B. 说话人相关变换的位置

SD参数的基本任务，是应对 Conformer 隐藏层输出中呈现的说话人级变异。与许多其他深度神经网络系统类似，靠近编码器输出端的较高层 Conformer 编码器层，往往比位于底部的层生成去噪效果更好且对扰动不变的特征。因此，较低层 Conformer 隐藏层输出中体现的说话人级变异预计更大，需要通过 SD参数加以应对；相比之下，较高层中的变异则较小。第 VI-A2 节表 II 所示的实验通过在 Conformer 编码器的不同内部位置引入 LHUC 变换，开展了一组消融研究，证实了这一点。实际观察发现，在编码器内部的不同位置中，对卷积下采样模块应用 SD LHUC 变换可获得最佳的说话人自适应性能配置。

C. 先验分布的选择

在对SD参数进行贝叶斯估计时，需根据不同自适应方法中相应激活函数的形式设置合适的参数先验分布。先验分布可以设为特定的简单正态分布。本文中，LHUC、HUB和PAct自适应的先验分布分别设为 $p(\boldsymbol{r}_s) = \mathcal{N}(\mathbf{0}, \mathbf{1})$、 $p(\boldsymbol{r}_s) = \mathcal{N}(\mathbf{0}, \mathbf{0.001})$和$\{p(\boldsymbol{\alpha}_s)\ p(\boldsymbol{\beta}_s)\} = \{\mathcal{N}(\mathbf{1}, \mathbf{1})\ \mathcal{N}(\mathbf{0}, \mathbf{1})\}$。或者，也可以使用根据训练数据估计得到的经验分布作为贝叶斯估计的先验分布。该先验分布 $\mathcal{N}(\bar{\boldsymbol{\mu}},\ \bar{\boldsymbol{\sigma}}^2)$ 可在所有训练说话人的说话人相关变换上计算得到，其形式如下

$$\bar{\boldsymbol{\mu}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \boldsymbol{r}_s,\ \bar{\boldsymbol{\sigma}}^2 = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \{(\boldsymbol{r}_s - \bar{\boldsymbol{\mu}}) \odot (\boldsymbol{r}_s - \bar{\boldsymbol{\mu}})\}, \quad (14)$$

其中， $|\mathcal{S}|$ 是训练说话人集合 $\mathcal{S}$中的实体数量。

D. 参数采样

贝叶斯估计的另一个重要问题是推断的计算效率。贝叶斯自适应中使用的变分推断需要借助蒙特卡洛采样向量，在自适应过程中近似式 (12) 中上界的第一项，并在评估过程中近似式 (9) 中的积分。由于针对每个参数值独立抽取样本，因此自适应过程中的推断成本会随样本数线性增长。为确保贝叶斯自适应的推断成本与传统的非贝叶斯、基于点估计的自适应相当，在自适应过程中仅在式 (12) 中抽取一个样本来更新 SD参数。第 VI-A2 节将通过表 II 中对参数样本数量的消融研究对此进行验证。在评估过程中，贝叶斯自适应所需的式 (9) 中的积分可通过变分分布的均值向量高效近似。

$$\int p(\tilde{\boldsymbol{Y}}^s|\tilde{\boldsymbol{X}}^s, \boldsymbol{r}^s) p(\boldsymbol{r}^s|\mathcal{D}^s) d\boldsymbol{r}^s \approx p(\tilde{\boldsymbol{Y}}^s|\tilde{\boldsymbol{X}}^s, \boldsymbol{\mu}). \quad (15)$$

VI. 实验

本节将研究所提出的基于置信度分数的自适应方案，这些方案采用 LHN、LHUC、HUB 和 PAct 形式的各种 SD参数表示，用于两项任务中说话人无关和说话人自适应训练 Conformer 系统的测试时无监督自适应：广泛使用的 300 小时 Switchboard 会话电话语料库 [78]，以及经过速度扰动的 233 小时 AMI 会议语音转写数据集 [79]。对于所有基于说话人自适应训练的 Conformer 系统，均采用对训练说话人数据进行有监督自适应且不进行数据选择的方法。

本节安排如下。第 VI-A 节和第 VI-B 节分别展示 Switchboard 和 AMI 任务的具体实验设置与结果。对于 300 小时 Switchboard 任务，第 VI-A1 节介绍 Switchboard 语音数据集，以及基线 Transformer 和 Conformer 自动语音识别系统的详细配置。第 VI-A2 节展示了在不同设置下采用各种 LHN、LHUC、HUB 和 PAct 自适应方法的性能。这些初步实验结果用于确认第五节讨论的实现细节和设置，并在后续实验中采用这些设置。在第 VI-A3 节中，通过使用不同置信度度量，将所提出的基于置信度分数的数据选择方法与多种选择方法进行广泛比较，以评估其有效性。最后，在第 VI-A4 节中，通过在外部语言模型重评分前后执行各种形式的结构化变换，并加入基于速度扰动的训练数据增强，展示所提出的基于置信度分数的贝叶斯自适应的优势。此外，还将该自适应方法与最新文献中采用混合和端到端自动语音识别系统在 300 小时 Switchboard 任务上取得的最先进性能进行比较，以进一步展现其竞争力。对于 233 小时 AMI 任务，第 VI-B1 节介绍了在 IHM、SDM 和 MDM 条件下录制的 AMI 会议语音转写数据集，以及基线 Conformer 自动语音识别系统。第 VI-B2 节通过开展一组可比实验，进一步展示了所提出的基于置信度分数的自适应方案在 AMI 数据集上的性能。所有模型均在单块 NVIDIA Ampere A40 GPU 上使用一个线程训练。

A. Switchboard 数据实验

1）实验设置：训练使用 Switchboard-1 会话电话语料库（LDC97S62） [78]，其中包含约 300 小时的音频数据，录自 4804 位说话人。为评估性能，使用分别采集自 80、120 和 144 位说话人的 NIST Hub5'00（LDC2002S09、LDC2002T43）、RT02（LDC2004S11）和 RT03（LDC2007S10）测试集。时长为 3.8 小时的 Hub5'00 语音数据包含两个子集：有 40 位说话人的 CallHome（CHE）和有 40 位说话人的 Switchboard（SWBD）。时长为 6.4 小时的 RT02 数据集包含三个子集：Switchboard 1（SWBD1）、Switchboard 2（SWBD2）和 Switchboard Cellular（SWBD3）。每个子集均包含 40 位说话人。时长为 6.2 小时的 RT03 数据包含两个子集：Fisher（FSH）和 Switchboard（SWBD）。每个子集均包含 72 位说话人。

表I 使用不同适应方法的 Transformer 和 Conformer 系统在 Hub5'00测试集上的性能（词错误率（WER））。「O.V.」表示「总体」。「†」表示与相应的基线说话人无关 Transformer（系统1）或 Conformer（系统10）系统相比，差异具有统计显著性。

ID	模型	自适应方法	# SD参数 每位说话人	Hub5'00 CHE	SWBD	O.V.
1	Transformer	-	-	17.3	9.1	13.2
2		LHN	26，214，400（100%）	17.9	9.5	13.7
3		LHUC	5,120（∼ 0.02%）	16.8†	9.1	13.0
4		HUB	5,120（∼ 0.02%）	16.9	9.1	13.0
5		PAct	10,240 (∼ 0.04%)	17.0	9.1	13.1
6		LHN-SAT	26，214，400（100%）	18.0	9.6	13.8
7		LHUC-SAT	5,120 (∼ 0.02%)	**16.1**†	**8.5**†	**12.3**†
8		HUB-SAT	5,120 (∼ 0.02%)	16.2†	**8.5**†	**12.3**†
9		PAct-SAT	10,240 (∼ 0.04%)	16.5†	8.9	12.8†
10	Conformer	-	-	15.0	7.3	11.1
11		LHN	26，214，400（100%）	15.3	7.4	11.3
12		LHUC	5，120（∼ 0.02%）	14.9	7.3	11.1
13		HUB	5，120（∼ 0.02%）	14.9	7.3	11.1
14		PAct	10，240（∼ 0.04%）	15.0	7.2	11.1
15		LHN-SAT	26，214，400（100%）	15.2	7.5	11.4
16		LHUC-SAT	5，120（∼ 0.02%）	**13.8**†	**7.1**	**10.5**†
17		HUB-SAT	5，120 (∼ 0.02%)	14.2†	**7.1**	10.7†
18		PAct-SAT	10，240 (∼ 0.04%)	14.7	7.2	11.0†

基线 Transformer 和 Conformer 模型均由 12 个编码器块和 6 个解码器块组成，并遵循 ESPnet 配置[3]。每个编码器块或解码器块均配置了 4 头注意力机制，维度为 256，并设有 2048 个前馈隐藏单元。卷积子采样模块包含两个步幅为 2 的二维卷积层，每层之后均接一个 ReLU 激活。输入特征和解码器输出分别采用 80 维梅尔滤波器组、3 维基音参数和大小为 2000 的字节对编码（BPE）词元。SI 和 SAT 模型训练均采用 SpecAugment 技术 [104] 。Noam 优化器的初始学习率为 5.0。丢弃率设为 0.1。训练轮数为 100，评估所用的最终识别模型通过对最后十轮的模型取平均得到。进一步使用在 Switchboard 和 Fisher 转录文本（LDC2004T19、LDC2005T19）上以跨话语上下文 [105] 训练的对数线性插值外部 Transformer 和双向长短期记忆网络（BiLSTM）语言模型（LM），对 Conformer 自动语音识别系统生成的 30 个最佳假设进行重打分。

2）基线自适应的性能：表 I 展示了在 Hub5'00 评估集上，对采用或未采用说话人自适应训练的 Transformer 和 Conformer 系统进行评估时，使用 LHN、LHUC、HUB 和 PAct 的非贝叶斯、基于点估计的自适应方法的性能。自适应过程中使用了全部说话人数据。说话人相关变换应用于卷积下采样模块。对于 Transformer 和 Conformer 系统，可以观察到以下几种趋势：**a**）与基线 SI 系统相比，使用 LHUC、HUB 或 PAct 对非 SAT 系统进行测试时自适应并未带来显著的性能提升（系统 3-5 对比系统 1，以及系统 12-14 对比系统 10）。进一步引入说话人自适应训练后，所得的 LHUC-SAT、HUB-SAT 和 PAct-SAT 系统一致地实现了具有统计显著性的总体词错误率降低，相较于基线 SI 系统，绝对值最高可降低 0.9%（系统 7-9 对比系统 1，以及系统 16-18 对比系统 10）。这证实，在测试时自适应期间，相较于 SI模型，SAT 模型能够更好地适应未见说话人。**b**）该

[3]github.com/espnet/espnet/blob/master/egs/swbd/asr1/run .sh

表 II 不同配置下 LHUC-SAT Conformer 系统在 Hub5'00 测试集上的性能（词错误率（WER）），配置包括：在说话人自适应训练和测试时自适应期间应用 LHUC 变换的位置（第 3-4 列）；贝叶斯自适应中的说话人相关参数先验分布（第 5 列）；以及变分推断中的参数样本数量（第 6 列）。“二维卷积”、“编码器（X）”和“编码器（X-Y）”分别表示卷积下采样模块、“第 X 个”编码器块层，以及从“第 X 个”到“第 Y 个”的编码器块层。

ID	适配参数	位置 SAT	位置 测试时	先验	#Sa-样本	Hub5'00 CHE	SWBD	O.V.
1	-	-	-	-	-	15.0	7.3	11.1
2	确定确定性	二维卷积	二维卷积	-	-	**13.8**	7.1	**10.5**
3			编码器（1）			14.9	7.4	11.2
4			编码器（6）			15.5	7.5	11.5
5			编码器（12）			16.5	7.8	12.2
6		编码器（1）	二维卷积	-	-	15.1	7.4	11.3
7			编码器（1）			14.3	7.1	10.7
8			编码器（6）			16.1	7.7	12.0
9			编码器（12）			16.6	8.1	12.4
10		编码器（6）	二维卷积	-	-	16.7	7.6	12.2
11			编码器（1）			15.1	7.2	11.2
12			编码器（6）			14.5	7.2	10.9
13			编码器（12）			16.7	8.2	12.5
14		编码器（12）	二维卷积	-	-	16.8	7.9	12.4
15			编码器（1）			15.6	7.8	11.7
16			编码器（6）			15.4	7.6	11.5
17			编码器（12）			14.7	7.3	11.0
18		二维卷积与编码器（1）		-	-	13.9	**7.0**	**10.5**
19		二维卷积与编码器（1-6）				14.1	7.1	10.6
20		二维卷积与编码器（1-12）				14.1	7.2	10.7
21	贝叶斯	二维卷积		经验。	1	13.8	7.0	10.4
22				$\mathcal{N}(\mathbf{0},\ \mathbf{1})$	1	**13.7**	**6.9**	**10.3**
23					2	13.8	6.9	10.3
24					3	13.8	7.0	10.4
25					4	**13.7**	7.0	**10.3**
26					5	13.8	6.9	10.3

无论是否采用说话人自适应训练，经过 LHN 自适应的系统表现都不如 SI 系统（系统 2、6 与系统 1 相比；系统 11、15 与系统 10 相比）。这是因为 LHN 系统的说话人相关参数过多，在自适应数据有限时容易过拟合，如表 I（第 4 列）所示。**c)**经过自适应的 LHUC-SAT Conformer 系统（系统 16）取得了最低的词错误率，因此被选用于后续消融研究，探讨说话人相关变换位置和贝叶斯自适应的实现问题，相关内容见第 V-B 节至第 V-D 节。

表 II 展示了按照第 V-B 节至第 V-D 节的实现细节所述，采用不同配置的自适应 LHUC-SAT Conformer 系统的性能。可以得出以下几点发现。**a)** 在说话人自适应训练和测试时自适应过程中，始终在同一层应用说话人相关变换（系统 2、7、12、17），比其他做法取得了更好的性能。此外，只需在靠近原始输入特征的卷积下采样模块应用说话人相关变换（系统 2）；在后续任意更高层（系统 18–20）进一步应用说话人相关变换并未带来额外提升。这证实了如第 V-B 节所述，较低层 Conformer 隐藏层输出中的说话人差异大于较高层中的差异。**b)** 基于零均值和单位方差正态先验分布的贝叶斯估计（系统 22），比基于经验先验分布的估计（系统 21）取得了更好的性能。**c)** 如第 V-D 节先前所述，在式（12）的计算中仅抽取一个说话人相关参数样本是合理的。在自适应过程中抽取更多蒙特卡罗样本（系统 22–26）时，未观察到具有统计显著性的词错误率差异。本文后续实验采用了性能最佳的自适应配置（系统 22）。

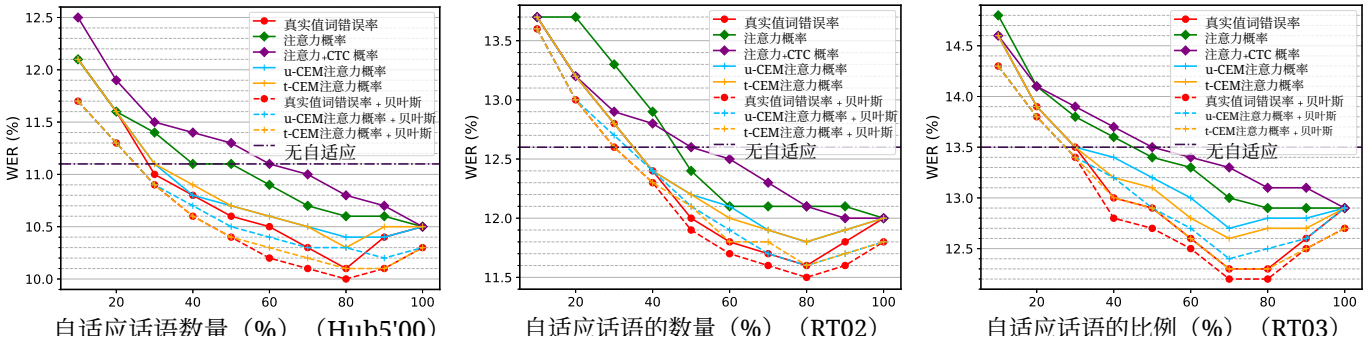


图 3. 使用不同置信度度量选取不同数量的自适应数据时，自适应后的 LHUC-SAT Conformer 系统的性能（词错误率（WER））（图例从上到下：真实值词错误率“真实值词错误率”；原始 Conformer 解码器注意力 Softmax 输出对数概率“注意力概率”，或将其与 CTC 对数概率按 7:3 的比例组合得到的结果“注意力+CTC 概率”；经句子级 CEM 平滑的注意力概率“u-CEM注意力概率”；经词元级 CEM 平滑的注意力概率“t-CEM注意力概率”；“+ 贝叶斯”表示贝叶斯测试时自适应。横坐标表示按置信度度量排序并用于自适应的不同规模的说话人级话语选定子集。

表 III

在 HUB5'00 数据上评估的、使用不同预测特征的语句级和词元级置信度估计模块的性能。“编码器嵌入”和“解码器嵌入”分别指从编码器和解码器最后一个块层提取的隐藏向量。“平均值”表示对帧或输出词元上的特征取平均。“10-best 注意力分数”、“10-best CTC 分数”和“10-best 注意力+CTC（7:3）分数”分别是从解码器模块、CTC 模块及其按 7:3 加权组合得到的前 10 个识别假设的负对数概率。“前 10 名分数逻辑值”表示整个词汇表中与前 10 个解码器 Softmax 概率对应的线性输出。

ID	CEM	预测特征		AUC	EER
1	话语级	内部嵌入	编码器嵌入（平均值）	0.7648	0.3232
2			解码器嵌入（平均值）	**0.7672**	**0.3194**
3		分数	1-best 注意力分数	0.7902	0.2888
4			5-best 注意力分数	0.8524	0.2302
5			10-best 注意力分数	**0.8538**	**0.2279**
6			20-best 注意力分数	0.8501	0.2291
7			10-best CTC 分数	0.8227	0.2634
8			10-best 注意力+CTC (7:3) 分数	0.8491	0.2375
9		融合	编号 (1) +编号 (5)	0.8603	0.2233
10			编号 (2) +编号 (5)	**0.8627**	**0.2204**
11			编号 (1) +编号 (2) +编号 (5)	0.8518	0.2352
12	词元级别	内部嵌入	编码器嵌入（平均值）	0.6956	0.3369
13			解码器嵌入	**0.7991**	**0.2621**
14		分数	Top-1 分数逻辑值	0.7780	0.2791
15			Top-5 分数逻辑值	0.8524	0.2214
16			前 10 名分数逻辑值	**0.8576**	**0.2195**
17			前 20 名分数逻辑值	0.8517	0.2226
18			所有分数逻辑值	0.8541	0.2234
19		融合	编号 (12) +编号 (16)	0.8380	0.2401
20			编号 (13) +编号 (16)	**0.8787**	**0.2076**
21			编号 (12) +编号 (13) + 编号 (16)	0.8251	0.2531

3) 用于自适应的置信度分数估计：表 III 展示了使用不同预测特征时，话语级和词元级置信度估计模块的置信度分数估计性能。等错误率（EER）和接收者操作特征（ROC）曲线下面积（AUC）被用作评估指标[4]。可以观察到几个重要趋势。**a)** 对于话语级和词元级置信度估计模块，基于解码器嵌入的预测特征（编号 2、13）始终比使用编码器嵌入特征（编号 1、12）取得更好的 AUC 和 EER 分数。**b)** 对于这两种置信度估计模块，与 1-best 分数相比，使用 n-best 分数也能持续提升 AUC 和 EER（编号 4–6 与编号 3，或编号 15–18 与编号 14 的比较）。正如第 V-A 节所述，来自竞争假设或词元的额外信息有利于 CEM 预测。**c)** 将解码器嵌入与 10-best 注意力分数或前 10 名分数逻辑值进行融合（编号 10 或编号 20）后，取得了最佳性能，因此在后续实验中分别选作话语级或词元级 CEM 训练的预测特征。这些设置是根据表 III 中与使用其他各种特征（例如内部 Conformer 嵌入）的性能比较结果选定的[5]。

如图 3 进一步所示，我们通过在 Hub5'00、RT02 和 RT03测试集上，使用不同置信度度量，将所提出的基于置信度分数的说话人级数据选择方法与多种选择方法进行广泛比较，评估该方法的有效性。图中展示了经过自适应的 LHUC-SAT Conformer 系统在选取不同数量说话人数据时的词错误率性能。可以观察到几种趋势。**a)** 在三个测试集上，随着自适应数据量的变化，采用所提出的话语级和词元级 CEM 平滑后的 Conformer 注意力概率（蓝线和橙线）所取得的自适应性能始终优于采用可比基线选择方法的结果；后者使用未经映射的原始注意力概率（绿线），以及将其与 CTC 概率进一步结合的结果（紫线）。**b)** 为避免过拟合风险而进行贝叶斯学习时，使用词元级 CEM 平滑置信度度量的贝叶斯自适应性能（橙色虚线）与之相当

[4]等错误率是 ROC 曲线上误接受率与误拒绝率相等时对应的工作点。曲线下面积衡量整个工作阈值范围内的平均分类性能。等错误率越低或曲线下面积越大，性能就越好。

[5]在表 III 中，Conformer 编码器和解码器嵌入都被视为 CEM 预测特征。首先，通过对每个语音话语中获得的帧级编码器嵌入向量取平均值，计算话语级内部 Conformer 编码器嵌入。因此，与单个输出词元级别的编码器嵌入相比，这类话语级编码器嵌入与话语级标签及其相应的话语级置信度分数具有更好的同步性。这解释了为何使用话语级编码器嵌入来预测话语级置信度分数（编号 1，表 III）的性能优于在词元级进行预测（编号 12，表 III）。然而，使用解码器嵌入与 10-best 注意力分数或前 10 名分数逻辑值进行融合，效果明显优于上述两种方法（编号 10 或编号 20，表 III）。因此，在本文表 IV 至表 VII 所列的主要实验中，分别选用这些特征作为话语级或词元级 CEM 的训练预测特征。其次，Conformer 解码器嵌入天然与输出词元一一对应，因此无需在两者之间进行额外对齐（例如，编号 13，表 III）。话语级解码器嵌入则通过对每个话语中的词元级解码器嵌入取平均值获得（例如，编号 2，表 III）。

表 IV：在 HUB5'00、RT02 和 RT03 测试集上，使用 LHUC、HUB 和 PAct，并结合或不结合说话人自适应训练（SAT）与贝叶斯估计的 Conformer 系统，在基于置信度分数的数据选择以及外部 Transformer 加长短期记忆网络语言模型重打分前后的说话人自适应性能（词错误率（WER））。"O.V."表示"总体"。"†"和"∗"分别表示：相较于基线 SI 系统（系统 1、26），以及相应的不采用贝叶斯方法、基于点估计的自适应基线且未进行说话人数据选择的系统（系统（2-4）、（14-16）、（27-29）、（39-41）），具有统计显著性的词错误率差异。

ID	系统	自适应参数	Data 选择	语言模型	Hub5'00			RT02				RT03		
					CHE	SWBD	O.V.	SWBD1	SWBD2	SWBD3	O.V.	FSH	SWBD	O.V.
1	SI模型	-	-		15.0	7.3	11.1	8.7	12.9	15.6	12.6	10.4	16.4	13.5
2	LHUC	确定性	✗	✗	14.9	7.3	11.1	8.7	12.8	15.5	12.6	10.4	16.4	13.5
3	HUB				14.9	7.3	11.1	8.8	12.8	15.4	12.6	10.4	16.4	13.5
4	PAct				15.0	7.2	11.1	8.7	12.7	15.8	12.7	10.4	16.3	13.5
5	LHUC	贝叶斯			14.8	7.2	11.0†∗	8.7	12.8	15.3†	12.5†∗	10.2†∗	16.3	13.3†∗
6	HUB				14.6†	7.3	11.0†∗	8.9	12.7†	15.3†	12.5†	10.3	16.1†∗	13.3†∗
7	PAct				14.6†	7.2	11.0†∗	8.7	12.8	15.6	12.6	10.4	16.3	13.5
8	LHUC	确定性	✓		14.5†∗	7.2	10.8†∗	8.6	12.7†	15.2†∗	12.4†∗	10.2†∗	16.1†∗	13.2†∗
9	HUB				14.5†∗	7.1	10.8†∗	8.6	12.7†	15.2†	12.4†∗	**10.0**†∗	16.4	13.3†∗
10	PAct				14.5†∗	7.1	10.8†∗	8.7	12.5†	15.1†∗	**12.3**†∗	10.2†∗	15.9†∗	13.1†∗
11	LHUC	贝叶斯			**14.2**†∗	**7.0**†∗	**10.6**†∗	8.6	12.6†∗	**15.0**†∗	**12.3**†∗	**10.0**†∗	**15.8**†∗	**13.0**†∗
12	HUB				14.3†∗	**7.0**†∗	10.7†∗	8.6	12.6†∗	15.1†∗	**12.3**†∗	**10.0**†∗	16.0†∗	13.1†∗
13	PAct				**14.2**†∗	**7.0**†∗	**10.6**†∗	**8.5**	**12.3**†∗	15.2†∗	**12.3**†∗	**10.0**†∗	15.9†∗	**13.0**†∗
14	LHUC-SAT	确定性	✗	✗	13.8†	7.1	10.5†	8.6	12.3†	14.7†	12.0†	10.0†	15.7†	12.9†
15	HUB-SAT				14.2†	7.1	10.7†	8.6	12.5†	14.8†	12.2†	10.1†	15.7†	13.0†
16	PAct-SAT				14.7	7.2	11.0†	8.6	12.8	15.6	12.6	10.3	16.4	13.5
17	LHUC-SAT	贝叶斯			13.7†	6.9†	10.3†∗	8.4†	12.1†∗	14.3†∗	11.8†∗	9.8†∗	15.4†∗	12.7†∗
18	HUB-SAT				13.7†∗	7.0†	10.4†∗	8.5	12.0†∗	14.2†∗	11.8†∗	9.7†∗	15.2†∗	12.5†∗
19	PAct-SAT				14.1†∗	7.0†	10.6†∗	8.6	12.5†∗	14.6†∗	12.1†∗	9.9†∗	15.4†∗	12.8†∗
20	LHUC-SAT	确定性	✓		13.7†	6.9†	10.3†∗	**8.3**†∗	12.0†∗	14.5†	11.8†∗	9.8†∗	15.4†∗	12.7†∗
21	HUB-SAT				13.9†	6.9†	10.4†∗	8.4†	12.1†∗	14.4†∗	11.8†∗	9.4†∗	15.4†∗	12.5†∗
22	PAct-SAT				14.1†∗	7.2	10.7†∗	8.6	12.7†	15.2†∗	12.3†∗	10.0†∗	15.7†∗	13.0†∗
23	LHUC-SAT	贝叶斯			**13.4**†∗	**6.8**†∗	**10.1**†∗	**8.3**†∗	**11.9**†∗	14.1†∗	**11.6**†∗	9.5†∗	**14.8**†∗	**12.3**†∗
24	HUB-SAT				13.5†∗	6.9†	10.2†∗	8.4†	12.0†∗	**13.8**†∗	**11.6**†∗	**9.4**†∗	14.9†∗	12.3†∗
25	PAct-SAT				13.6†∗	7.2	10.4†∗	8.4†	12.3†∗	14.6†∗	11.9†∗	9.8†∗	15.2†∗	12.6†∗
26	SI模型	-	-	Transformer + 双向长短期记忆网络	14.2	6.8	10.5	8.4	12.5	14.3	11.9	9.4	15.3	12.4
27	LHUC	确定性	✗		14.2	6.8	10.5	8.1†	12.1†	14.2	11.7†	9.4	15.2	12.4
28	HUB				14.0	6.8	10.4†	8.2	12.1†	14.1	11.7†	9.3	15.3	12.4
29	PAct				14.2	6.8	10.5	8.2	12.0†	14.3	11.7†	9.3	15.3	12.4
30	LHUC	贝叶斯			13.9	6.7	10.3†∗	8.0†	11.7†∗	13.7†∗	11.4†∗	9.1†∗	15.0†	12.1†∗
31	HUB				13.7†	6.7	10.2†∗	8.1†	11.8†∗	14.0†	11.5†∗	9.2†	15.0†	12.2†∗
32	PAct				13.8†∗	6.7	10.3†∗	8.3	11.8†∗	13.9†∗	11.5†∗	9.3	14.8†∗	12.1†∗
33	LHUC	确定性	✓		13.8†∗	6.6	10.2†∗	**7.9**†	11.5†∗	13.8†∗	11.3†∗	9.2	14.8†∗	12.1†∗
34	HUB				13.3†∗	6.7	**10.0**†∗	8.2	11.7†∗	13.9†	11.4†∗	9.1†∗	14.9†∗	12.1†∗
35	PAct				13.6†∗	6.8	10.2†∗	8.0†	**11.3**†∗	13.7†∗	11.3†∗	9.1†∗	14.6†∗	12.0†∗
36	LHUC	贝叶斯			13.4†∗	**6.5**†∗	**10.0**†∗	**7.9**†	11.5†∗	**13.5**†∗	**11.2**†∗	**9.0**†∗	**14.4**†∗	**11.8**†∗
37	HUB				**13.2**†∗	6.7	**10.0**†∗	8.1†	11.5†∗	13.7†∗	11.3†∗	**9.0**†∗	14.7†∗	11.9†∗
38	PAct				13.5†∗	6.7	10.1†∗	8.0†	11.5†∗	**13.5**†∗	**11.2**†∗	**9.0**†∗	**14.4**†∗	**11.8**†∗
39	LHUC-SAT	确定性	✗		13.4†	6.8	10.1†	8.2	11.8†	13.9†	11.5†	9.0†	14.6†	12.0†
40	HUB-SAT				13.5†	6.9	10.2†	8.1†	11.7†	13.8†	11.4†	9.0†	14.8†	12.1†
41	PAct-SAT				14.0	7.0	10.5	8.5	12.1†	14.4	11.9	9.4	15.0	12.3
42	LHUC-SAT	贝叶斯			13.0†∗	6.7	9.8†∗	8.1†	11.6†∗	13.1†∗	11.1†∗	8.5†∗	14.0†∗	11.4†∗
43	HUB-SAT				13.0†∗	6.6∗	9.8†∗	7.7†∗	11.3†∗	13.3†∗	11.0†∗	8.6†∗	14.2†∗	11.5†∗
44	PAct-SAT				13.4†∗	7.0	10.2†∗	8.5	11.9†∗	14.0†∗	11.6†∗	9.0†∗	14.6†∗	11.9†∗
45	LHUC-SAT	确定性	✓		12.8†∗	6.4†∗	9.6†∗	7.9†∗	11.2†∗	13.0†∗	10.9†∗	8.5†∗	13.8†∗	11.3†∗
46	HUB-SAT				12.9†∗	6.5†∗	9.7†∗	7.9†	**11.0**†∗	12.9†∗	10.8†∗	8.4†∗	14.0†∗	11.3†∗
47	PAct-SAT				13.5†∗	6.7∗	10.1†∗	8.3	11.7†∗	13.9†∗	11.5†∗	9.2†∗	14.5†∗	11.9†∗
48	LHUC-SAT	贝叶斯			12.8†∗	**6.2**†∗	**9.5**†∗	7.7†∗	**11.0**†∗	**12.6**†∗	**10.6**†∗	**8.3**†∗	**13.6**†∗	**11.0**†∗
49	HUB-SAT				**12.6**†∗	6.3†∗	**9.5**†∗	**7.6**†∗	11.1†∗	**12.6**†∗	**10.6**†	8.4†∗	**13.6**†∗	11.1†∗
50	PAct-SAT				12.9†∗	6.7∗	9.8†∗	8.0†∗	11.4†∗	13.6†∗	11.2†∗	8.7†∗	14.0†∗	11.5†∗

与使用预言机真实标注词错误率选择方法的结果相比（所有三幅图底部的红色虚线）。此外，词元级 CEM（橙色虚线）的贝叶斯自适应性能略优于话语级 CEM（蓝色虚线）。这可能是因为，细粒度的词元级预测特征比通过对词元级特征取平均值得到的话语级特征更有助于区分正确假设与错误假设。对于性能最佳的基于词元级 CEM 平滑置信度分数的数据选择方法（橙色虚线），最佳的数据选择百分位点约为 80%；后续第 VI-A4 节的自适应实验均采用这一百分位点。

4）基于置信度分数的贝叶斯自适应性能：表 IV 和表 V 进一步展示了在三个测试集上评估的采用 LHUC、HUB 和 PAct 的 Conformer 系统的自适应性能，主要趋势如下可以观察到。

a）在自适应的数据选择中使用经 CEM 平滑的置信度分数后，得到的 LHUC-SAT、HUB-SAT 和 PAct-SAT 说话人自适应 Conformer 系统（系统 20–22）在全部三个测试集上的表现始终优于基线说话人无关 Conformer 系统（系统 1）以及未使用数据选择的相应基线自适应方法（系统 14–16）。尤其是，基于置信度分数的 HUB-SAT Conformer 系统（系统 21）在 RT03测试集的 FSH子集上，相较于基线说话人无关系统（系统 1）和未使用数据选择的相应自适应系统（系统 15），分别实现了具有统计显著性的词错误率降低：绝对降低**1.0%**（相对降低**9.6%**）和绝对降低**0.7%**（相对降低**6.9%**）。

b) 为解决数据选择导致的数据稀疏性增加问题，进一步将贝叶斯学习应用于基于置信度分数的自适应后，贝叶斯自适应

表 V：经过数据增强、使用 900 小时数据训练的 LHUC-SAT Conformer 系统在三个测试集上的性能（词错误率（WER）），比较采用或不采用贝叶斯说话人自适应的情况，并展示基于置信度分数的数据选择以及外部 Transformer 加长短期记忆网络语言模型重打分前后的结果。“†”和“*”分别表示与对应的基线 SI 系统（系统 1、5）以及未进行数据选择的对应非贝叶斯、基于点估计的自适应（系统 2、6）相比，词错误率差异具有统计显著性。

ID	系统	自适应参数	Data 数据选择	语言模型	Hub5'00			RT02				RT03		
					CHE	SWBD	O.V.	SWBD1	SWBD2	SWBD3	O.V.	FSH	SWBD	O.V.
1	SI模型	-	-	✗	15.1	7.2	11.1	8.5	12.6	14.7	12.1	10.3	16.1	13.3
2	LHUC-SAT	确定性	✗		13.9†	7.0	10.5†	8.2†	12.1†	14.1†	11.7†	9.9†	15.7†	12.9†
3		确定性	✓		13.5†*	**6.9**†	10.2†*	8.0†	11.9†*	13.4†*	11.2†*	9.4†*	15.0†*	12.3†*
4		贝叶斯	✓		**13.0**†*	**6.9**†	**10.0**†*	**7.9**†*	**11.8**†*	**13.2**†*	**11.1**†*	**9.3**†*	**14.7**†*	**12.1**†*
5	SI模型	-	-	Transformer + 双向长短期记忆网络	14.0	6.8	10.4	8.0	12.0	13.4	11.3	9.3	15.1	12.3
6	LHUC-SAT	确定性	✗		13.5†	6.7	10.1†	7.8	11.4†	12.6†	10.8†	8.9†	14.7†	12.0†
7		确定性	✓		12.5†*	6.6	9.6†*	7.6†	**10.7**†*	12.0†*	10.3†*	8.3†*	13.8†*	11.2†*
8		贝叶斯	✓		**12.4**†*	**6.5**†	**9.4**†*	**7.4**†*	**10.7**†*	**11.8**†*	**10.1**†*	**8.2**†*	**13.6**†*	**11.0**†*

LHUC-SAT、HUB-SAT 和 PAct-SAT Conformer 系统（系统 23–25）在全部三个测试集上的性能均持续优于相应的非贝叶斯确定性估计自适应方法（系统 20–22）。例如，在 RT02 数据的 SWBD3 子集上，与非贝叶斯自适应的 HUB-SAT Conformer 系统（系统 21）相比，贝叶斯自适应的 HUB-SAT Conformer 系统（系统 24）的词错误率具有统计显著性地绝对降低了**0.6%**（相对降低了**4.2%**）。

c) 仅进行测试时自适应时，也获得了类似的性能提升。采用基于置信度分数的数据选择和贝叶斯学习的 LHUC、HUB 或 PAct 自适应 Conformer 系统（系统 11–13），其性能始终优于相应的未进行数据选择的标准自适应系统（系统 2–4），以及基于置信度分数的非贝叶斯自适应 Conformer 系统（系统 8–10）。

d) 进一步使用外部 Transformer 和双向 LSTM 语言模型对 n-best 输出进行重打分时，也观察到类似且稳定的性能提升。所提出的基于置信度分数的贝叶斯自适应 LHUC-SAT Conformer 系统（系统 48）取得了最佳自适应性能。总体而言，相较于基线说话人无关 Conformer 系统（系统 26），在 Hub5'00、RT02 和 RT03测试集上分别实现了具有统计显著性的绝对词错误率降低**1.0%**、**1.3%** 和**1.4%**（相对降低**9.5%**、**10.9%**和**11.3%**）。

e) 如表 V 所示，在系统训练中进一步加入基于速度扰动的数据增强后，表 IV 中呈现的趋势依然保持不变。经过外部语言模型重评分后，在全部三个测试集上，与基线说话人无关系统及相应的非贝叶斯、基于点估计的自适应系统（系统 8 对比系统 5、6）相比，词错误率分别绝对降低了 **1.0%-1.3%** 和 **0.7%-1.0%**。表 IV 和表 V 所示的实验结果表明，所提出的基于置信度分数的自适应、速度扰动和外部语言模型重评分相辅相成，其改进效果大体上可以叠加。

表 VI 进一步将所提出的基于置信度分数的贝叶斯自适应 LHUC-SAT Conformer 系统（表 V 中的系统 8）的最佳性能，与最新文献中报道的、采用混合和端到端自动语音识别系统在 300 小时 Switchboard 任务上取得的最先进性能进行对比，以展示其竞争力。

B. AMI 数据上的实验

1）实验设置：使用从 1641 名说话人处采集的 AMI 语音数据进行训练；这些数据经过速度扰动处理，音频总时长近 233 小时数据采集自会议环境，使用了头戴式麦克风和远场麦克风阵列。评估使用了 8.9 小时的开发集（DEV），其中包含 72 名说话人的数据；以及 8.6 小时的评估集（EVAL），其中包含 63 名说话人的数据。录音条件包括个人头戴式麦克风（ihm）、单远场麦克风（sdm）和多远场麦克风（mdm）。ESPnet Conformer 自动语音识别系统的配置与第 VI-A1 节所述 Switchboard 任务的配置相同，但使用了大小为 100 的 BPE 词元。为进一步重打分，使用了基于 AMI 转录文本训练的外部 Transformer 语言模型。

2) 基于置信度分数的 *LHUC* 自适应性能：表 IV 中此前展示的实验结果表明，在 Switchboard 任务上，与其他自适应方法相比，LHUC 自适应取得了最佳性能。因此，我们开展了一系列可比较的 LHUC
我们在 AMI 数据上进行了自适应实验，以进一步证明所提出的基于置信度分数的自适应方法的有效性和优势。

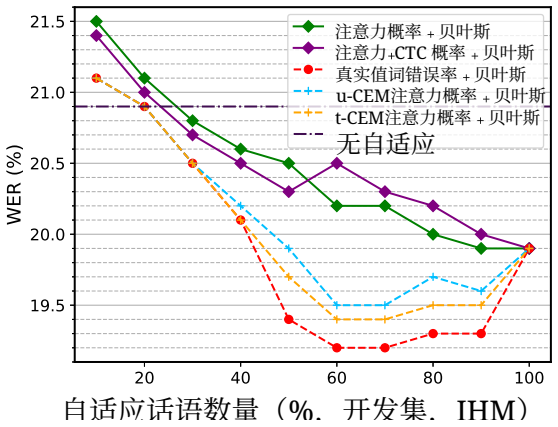


图 4. 使用不同置信度度量选择不同数量自适应数据后，自适应 LHUC-SAT Conformer 系统的性能（词错误率（WER））。所有命名约定均与图 3 一致。

图 4 展示了在 IHM 开发数据上评估的 LHUC-SAT Conformer 系统采用不同说话人数据选择方法进行自适应后的性能。在不同的所选子集中，可以观察到与图 3 中类似的总体趋势。与使用原始 Conformer 输出注意力，或使用注意力加 CTC 概率进行基线选择（绿色和紫色线）的可比方法相比，所提出的话语级 CEM（蓝色点线）和词元级 CEM（橙色点线）始终能带来更好的自适应性能。

表 VI 基于置信度分数的贝叶斯 LHUC-SAT Conformer 系统与其他最先进系统在 HUB5'00 和 RT03 数据上的性能（词错误率（WER））对比。

ID	系统	#段落。	Hub5'00			RT03		
			CHE	SW.	O.V.	FSH	SW.	O.V.
1	亚琛工业大学-2019 混合 [106]	-	13.5	6.7	10.2	-	-	-
2	谷歌-2019 LAS [104]	-	14.1	6.8	(10.5)	-	-	-
3	IBM-2020 AED [107]	29M	14.6	7.4	(11.0)	-	-	-
4		75M	13.4	6.8	(10.1)	-	-	-
5		280M	12.5	6.4	9.5	8.4	14.8	(11.7)
6	Salesforce-2020 Transformer [108]	-	13.3	6.3	(9.8)	-	-	11.4
7	香港中文大学-2021 BLHUC-混合型 [44]	15.2M	12.7	6.7	9.7	7.9	13.4	10.7
8	IBM-2021 CFM-AED [21]	68M	11.2	5.5	(8.4)	7.0	12.6	(9.9)
9	IBM-2021 RNN-T [14]	57M	13.1	6.3	(9.7)	9.5	15.4	(12.6)
10	RWTH-2022 CFM-Hybrid [109]	-	14.1	6.3	10.3	-	-	-
11	RWTH-2022 适配版 CFM-Hybrid [23]	-	12.1	6.3	9.2	-	-	-
12	**SI Conformer{表 V-系统 5}（本文方法）**	45M	14.0	6.8	10.4	9.3	15.1	12.3
13	**+ 自适应{表 V-系统 8}（本文方法）**	+ 30k	12.4	6.5	9.4	8.2	13.6	11.0

表 VII

在 IHM、SDM 和 MDM 条件下的 AMI 开发集和评估集上，采用或不采用贝叶斯估计和基于置信度分数的数据选择时，LHUC 和 LHUC-SAT Conformer 系统在外部 Transformer 重打分前后的性能（词错误率（WER））。† 和 ∗ 分别表示，与相应的基线 SI 系统（系统 1、10）以及相应的未进行数据选择的基于点估计的自适应（系统 2、6、11、15）相比，具有统计显著性的词错误率差异。

ID	系统	自适应参数	Data 选择	语言模型	DEV				EVAL			
					IHM	SDM	MDM	O.V.	IHM	SDM	MDM	O.V.
1	SI模型	-	-	✗	20.9	38.0	34.9	31.3	20.7	42.1	37.8	33.6
2	LHUC	确定性 贝叶斯	✗		20.7†	38.0	34.9	31.3	20.5†	41.6†	37.8	33.3†
3					20.5†*	37.7†*	34.7	30.9†*	20.3†*	41.3†	37.5†*	33.0†*
4		确定性 贝叶斯	✓		20.3†*	37.4†*	34.4†*	30.7†*	19.9†*	41.0†*	37.1†*	32.7†*
5					**20.1**†*	**37.2**†*	**34.2**†*	**30.5**†*	**19.7**†*	**40.9**†*	**36.7**†*	**32.4**†*
6	LHUC-SAT	确定性 贝叶斯	✗		20.4†	36.4†	33.8†	30.2†	20.1†	41.2†	35.9†	32.4†
7					19.9†*	36.0†*	33.0†*	29.6†*	19.7†*	39.6†*	35.6†*	31.6†*
8		确定性 贝叶斯	✓		19.7†*	35.1†*	32.8†*	29.2†*	19.1†*	38.6†*	34.8†*	30.8†*
9					**19.4**†*	**34.6**†*	**32.2**†*	**28.7**†*	**18.8**†*	**37.9**†*	**34.0**†*	**30.2**†*
10	SI模型	-	-	Transformer	20.2	36.9	33.9	30.4	19.8	40.7	36.6	32.4
11	LHUC	确定性 贝叶斯	✗		20.1	36.9	33.8	30.3	19.8	40.6	36.6	32.4
12					19.9†*	36.7	33.5†*	30.0†*	19.7	40.2†*	36.0†*	32.0†*
13		确定性 贝叶斯	✓		19.5†*	36.2†*	33.3†*	29.7†*	19.1†*	39.7†*	35.7†*	31.5†*
14					**19.3**†*	**35.9**†*	**33.0**†*	**29.4**†*	**19.0**†*	**39.5**†*	**35.4**†*	**31.3**†*
15	LHUC-SAT	确定性 贝叶斯	✗		19.7†	35.3†	32.7†	29.3†	19.1†	38.8†	34.9†	30.9†
16					19.4†*	34.5†*	31.8†*	28.6†*	18.7†*	38.1†*	34.0†*	30.3†*
17		确定性 贝叶斯	✓		19.1†*	34.0†*	31.3†*	28.2†*	18.3†*	37.3†*	33.4†*	29.7†*
18					**18.8**†*	**33.5**†*	**30.9**†*	**27.7**†*	**18.0**†*	**36.6**†*	**32.9**†*	**29.1**†*

所提出的词元级 CEM（橙色点线）的性能最接近采用预言机真实值词错误率数据选择所得的性能（红色点线）。子集选择的最佳百分位工作点约为 70%。

使用上述基于词元级 CEM 选出的前 70% 自适应数据，在 IHM、SDM 和 MDM 条件下，于开发集和评估集上评估所提出的基于置信度分数的 LHUC 自适应，其性能如表 VII 所示。可以观察到几个主要趋势。

a) 基于置信度分数的自适应 LHUC-SAT Conformer 系统（系统 8）持续实现了具有统计显著性的词错误率降低：从**0.7%**（开发集，IHM）到 **2.6%**（评估集，SDM），绝对降低（相对降低**3.4%** 至 **6.3%**），且无需进行数据选择，便优于相应的基线自适应系统（系统 6）。

b)在测试时自适应中进一步引入贝叶斯估计后，与采用非贝叶斯、基于点估计的自适应方法（系统 8）相比，基于置信度分数的贝叶斯自适应 Conformer 系统（系统 9）在两个测试集上均取得了稳定的额外性能提升，绝对词错误率最高降低了 **0.8%**（评估集，MDM）。

c) 对非 SAT 的 SI Conformer 系统进行测试时自适应时，也发现了类似趋势。所提出的基于置信度分数的贝叶斯 LHUC 自适应 Conformer 系统（系统 5），始终优于相应的未进行数据选择的基线自适应系统（系统 2）和基于置信度分数的非贝叶斯自适应系统（系统 4）。

d) 使用外部 Transformer 语言模型进行 n-best 重打分后，性能提升依然稳定。基于置信度分数的贝叶斯自适应 LHUC-SAT 系统（系统 18）取得了最佳性能。与基线说话人无关 Conformer 系统（系统 10）相比，该系统在开发集和评估集上分别实现了总体具有统计显著性的词错误率绝对降低**2.7%** 和**3.3%**（相对降低**8.9%** 和 **10.2%**）。

VII. 结论

本文提出了基于置信度分数的无监督贝叶斯说话人自适应方法，采用多种紧凑且数据高效的说话人相关（SD）参数表示，以支持说话人自适应训练

以及先进 Conformer 自动语音识别系统的测试时无监督说话人自适应。通过基于专门设计的置信度估计模块所生成的置信度分数进行说话人级别的数据选择，降低了自适应性能对监督质量的敏感性。基于置信度分数的说话人数据选择进一步加剧了数据稀疏问题，而贝叶斯学习缓解了这一问题。在基准 300小时 Switchboard 语料库和 233 小时 AMI 数据集上的实验表明，所提出的基于置信度分数的自适应方案始终优于基线说话人无关 Conformer 模型，以及未进行说话人数据选择的传统非贝叶斯点估计自适应方法。经过外部语言模型重打分后，仍保持了类似的性能提升。未来研究将聚焦于 Conformer 及其他端到端自动语音识别系统的快速在线自适应。

致谢

本研究获香港研究资助局优配研究金资助（项目编号：14200021、14200220）、创新及科技基金资助（项目编号：ITS/254/19 和 ITS/218/21）、中国国家重点研发计划（2020YFC2004100）、国家自然科学基金委员会项目（编号：62106255）、广州市民政科技基金项目（编号：2022MZK02）以及广西人机交互与智能决策重点实验室开放研究基金（GXHIID2202）资助。

参考文献

[1] G. E. Dahl 等，“用于大词汇量语音识别的上下文相关预训练深度神经网络”，IEEE/ACM 音频、语音与语言处理汇刊，第20卷，42，2011 1，30第期，第 – 页。[2] O. Abdel-Hamid 等，“用于语音识别的卷积神经网络”，IEEE/ACM 音频、语音与语言处理汇刊，第22卷，第10期，第1533–1545页，2014年。[3] V. Peddinti 等，“用于高效建模长时序上下文的时延神经网络架构”，载于 INTERSPEECH，2015年，第3214–3218页。[4] W. Chan 等，“听、关注与拼写：用于大词汇量会话语音识别的神经网络”，载于 ICASSP，2016年，第4960–4964页。[5] S. Watanabe 等，“用于端到端语音识别的混合 CTC/注意力架构”，IEEE J-STSP，第11卷，第8期，第1240–1253页，2017年。[6] J. Li 等，“大规模语音识别中常见端到端模型的比较”，载于 INTERSPEECH，2020年，第1–5页。[7] L. Dong 等，“Speech-Transformer：用于语音识别的无循环序列到序列模型”，载于 ICASSP，2018年，第5884–5888页。[8]S. Karita 等，“结合连接时序分类和语言模型集成改进基于 Transformer 的端到端语音识别”，载于 INTERSPEECH，2019年，第1408–1412页。[9] P. Bell 等，“基于神经网络的语音识别自适应算法：综述”，《开放信号处理期刊》，第2卷，第33–66页，2020年。[10] Y. Wang 等，“用于混合语音识别的基于 Transformer 的声学建模”，载于 ICASSP，2020年，第6874–6878页。[11] A. Gulati 等，“ Conformer：用于语音识别的卷积增强 Transformer”，载于 INTERSPEECH，2020年，第5036–5040页。[12] P. Guo 等，“ Conformer 推动 ESPnet 工具包的近期发展”，载于 ICASSP，2021年，第5874–5878页。[13] J. Li 等，“端到端自动语音识别的最新进展”，《亚太信号与信息处理协会信号与信息处理汇刊》，第11卷，第1期，2022年。[14] G. Saon 等，“推进用于语音识别的循环神经网络转导器技术”，载于 ICASSP，2021年，第5654–5658页。

[15] J. Deng 等人，《基于置信度分数的 Conformer 说话人自适应语音识别》，INTERSPEECH，第 2623–2627 页，2022。[16] A. Graves 等人，《连接时序分类：使用循环神经网络为未分段序列数据添加标签》，载于 ICML，2006，第 369–376 页。[17] J. K. Chorowski 等人，《基于注意力的语音识别模型》，载于 NeurIPS，2015，第 577–585 页。[18] A. Graves，《使用循环神经网络进行序列转导》，载于 ICML Workshop，2012。[19] K. Rao 等人，《探索使用循环神经网络转导器进行流式端到端语音识别的架构、数据和单元》，载于 ASRU，2017，第 193–199 页。[20] S. Karita 等人，《语音应用中 Transformer 与循环神经网络的比较研究》，载于 ASRU，2019，第 449–456 页。[21] Z. T¨uske 等人，《英语会话语音识别的极限》，载于 INTERSPEECH，2021，第 2062–2066 页。[22] Y. Huang 等人，《Conformer 转导器的快速说话人自适应：你只需要注意力和偏置》，载于 INTERSPEECH，2021，第 1309–1313 页。[23] M. Zeineldeen 等人，《改进基于 Conformer 的鲁棒混合模型训练方案》，载于 INTERSPEECH，2022，第 1036–1040 页。[24] O. Abdel-Hamid 等人，《基于说话人编码判别式学习的混合神经网络／隐马尔可夫模型语音识别快速说话人自适应》，载于 ICASSP，2013，第 7942–7946 页。[25] M. Karafi´at 等人，《基于 i-vector 的自动语音识别判别式自适应》，载于 ASRU，2011，第 152–157 页。[26] G. Saon 等人，《使用 i-vector 对神经网络声学模型进行说话人自适应》，载于 ASRU，2013，第 55–59 页。[27] H. Huang 等人，《研究增强说话人表征以改善基于深度神经网络的语音识别中的说话人归一化》，载于 ICASSP，2015，第 4610–4613 页。[28] V. V. Digalakis 等人，《使用高斯混合模型约束估计进行说话人自适应》，IEEE 语音与音频处理汇刊，第 3 卷，第 5 期，第 357–366 页，1995。[29] M. J. Gales，《基于隐马尔可夫模型的语音识别最大似然线性变换》，《COMPUT SPEECH LANG》，第 12 卷，第 2 期，第 75–98 页，1998。[30] E. Eide 等人，《声道长度归一化的参数化方法》，载于 ICASSP，1996，第 346–348 页。[31] L. Lee 等人，《使用高效频率扭曲方法进行说话人归一化》，载于 ICASSP，1996，第 353–356 页。[32] L. F. Uebel 等人，《声道长度归一化研究》，载于 EUROSPEECH，1999，第 2527–2530 页。[33] S. H. K. Parthasarathi 等人，《基于特征空间最大似然线性回归的深度神经网络声学模型说话人自适应》，载于 INTERSPEECH，2015，第 3630–3634 页。[34] H. Liao，《上下文相关深度神经网络的说话人自适应》，载于 ICASSP，2013，第 7947–7951 页。[35] D. Yu 等人，《基于 KL 散度正则化的深度神经网络自适应，以改进大词汇量语音识别》，载于 ICASSP，2013，第 7893–7897 页。[36] Y. Huang 等人，《用于演讲转录和智能会议助手系统的声学模型自适应》，载于 ICASSP，2020，第 7394–7398 页。[37] S. Xue 等人，《基于判别码的深度神经网络快速自适应语音识别》，IEEE/ACM 音频、语音与语言处理汇刊，第 22 卷，第 12 期，第 1713–1725 页，2014。[38] M. J. Gales，《隐马尔可夫模型的簇自适应训练》，IEEE 语音与音频处理汇刊，第 8 卷，第 4 期，第 417–428 页，2000。[39] C. Wu 等人，《用于语音识别快速自适应的多基自适应神经网络》，载于 ICASSP，2015，第 4315–4319 页。[40] T. Tan 等人，《基于深度神经网络的声学模型簇自适应训练》，IEEE/ACM 音频、语音与语言处理汇刊，第 24 卷，第 3 期，第 459–468 页，2015。[41] P. Swietojanski 等人，《学习隐单元贡献以对神经网络声学模型进行无监督说话人自适应》，载于 SLT，2014，第 171–176 页。[42] P. Swietojanski 等人，《学习隐单元贡献以进行无监督声学模型自适应》，IEEE/ACM 音频、语音与语言处理汇刊，第 24 卷，第 8 期，第 1450–1463 页，2016。[43] X. Xie 等人，《BLHUC：用于深度神经网络说话人自适应的隐单元贡献贝叶斯学习》，载于 ICASSP，2019，第 5711–5715 页。

[44] X. Xie 等人，《深度神经网络自适应的贝叶斯学习》，IEEE/ACM 音频、语音与语言处理汇刊，第 29 卷，第 2096–2110 页，2021。[45] C. Zhang 等人，《使用参数化 Sigmoid 和 ReLU 隐藏层激活函数进行深度神经网络说话人自适应》，载于 ICASSP，2016，第 5300–5304 页。[46] J. P. Neto 等人，《用于连续语音识别的混合 HMM-ANN 系统说话人自适应》，载于 EUROSPEECH，1995，第 2171–2174 页。[47] R. Gemello 等人，《用于混合 ANN/HMM 模型自适应的线性隐藏层变换》，《语音通信》，第 49 卷，第 10–11 期，第 827–835 页，2007。[48] B. Li 等人，《混合 NN/HMM 系统中说话人自适应的判别式输入与输出变换比较》，载于 INTERSPEECH，2010，第 526–529 页。[49] Y. Zhao 等人，《深度神经网络的低秩加对角自适应》，载于 ICASSP，2016，第 5005–5009 页。[50] T. Anastasakos 等人，《用于说话人自适应训练的紧凑模型》，载于 ICSLP，1996，第 1137–1140 页。[51] T. Ochiai 等人，《使用深度神经网络进行说话人自适应训练》，载于 ICASSP，2014，第 6349–6353 页。[52] M. Kitza 等人，《用于说话人自适应的 BLSTM 层特定仿射变换比较》，载于 INTERSPEECH，2018，第 877–881 页。[53] Z. Meng 等人，《通过对抗学习进行说话人不变训练》，载于 ICASSP，2018，第 5969–5973 页。[54] T. Tsuchiya 等人，《通过对抗学习为零资源语言提取说话人不变特征》，载于 ICASSP，2018，第 2381–2385 页。[55] R. Price 等人，《使用输出层层级结构进行深度神经网络说话人自适应》，载于 SLT，2014，第 153–158 页。[56] P. Swietojanski 等人，《带辅助目标的结构化输出层，用于上下文相关声学建模》，载于 INTERSPEECH，2015，第 3605–3609 页。[57] Z. Huang 等人，《通过多任务学习实现深度神经网络快速自适应》，载于 INTERSPEECH，2015，第 3625–3629 页。[58] Z. Huang 等人，《深度模型中网络参数的最大后验自适应》，载于 INTERSPEECH，2015，第 1076–1080 页。[59] Z. Huang 等人，《用于自动语音识别深度模型中激活函数参数的贝叶斯无监督批量与在线说话人自适应》，IEEE/ACM 音频、语音与语言处理汇刊，第 25 卷，第 1 期，第 64–75 页，2016。[60] M. K. Baskar 等人，《基于 wav2vec2 的构音障碍语音识别说话人自适应》，载于 INTERSPEECH，2022，第 3403–3407 页。[61] M. Delcroix 等人，《基于辅助特征的端到端语音识别系统自适应》，载于 INTERSPEECH，2018，第 2444–2448 页。[62] L. Sarı 等人，《使用基于注意力的说话人记忆对端到端语音识别进行无监督说话人自适应》，载于 ICASSP，2020，第 7384–7388 页。[63] Z. Fan 等人，《具备说话人感知能力的 Speech-Transformer》，载于 ASRU，2019，第 222–229 页。[64] Y. Zhao 等人，《具有说话人感知持久记忆的 Speech-Transformer》，载于 INTERSPEECH，2020，第 1261–1265 页。[65] T. Ochiai 等人，《多通道端到端语音识别的说话人自适应》，载于 ICASSP，2018，第 6707–6711 页。[66] K. C. Sim 等人，《用于自动语音识别的鲁棒连续设备端个性化》，载于 INTERSPEECH，2021，第 1284–1288 页。[67] Y. Huang 等人，《使用个性化语音合成和神经语言生成器实现快速 RNN-T 自适应》，载于 INTERSPEECH，2020，第 1256–1260 页。[68] K. Li 等人，《端到端 CTC 模型的说话人自适应》，载于 SLT，2018，第 542–549 页。[69] Z. Meng 等人，《基于注意力的端到端语音识别的说话人自适应》，载于 INTERSPEECH，2019，第 241–245 页。[70] F. Weninger 等人，《倾听、关注、拼写与自适应：说话人自适应序列到序列语音识别》，载于 INTERSPEECH，2019，第 3805–3809 页。[71] K. C. Sim 等人，《移动设备上端到端语音识别的个性化：面向命名实体》，载于 ASRU，2019，第 23–30 页。[72] V. Manohar 等人，《使用无格 MMI 进行声学模型半监督训练》，载于 ICASSP，2018，第 4844–4848 页。[73] G. Sivaraman 等人，《端到端语音识别的无监督模型自适应》，载于 ICASSP，2022，第 6987–6991 页。[74] G. Evermann 等人，《后验概率解码、置信度估计与系统组合》，载于《语音转录》，2000，第 78–81 页。[75] L. Mangu 等人，《在语音识别中寻找共识：词错误最小化及混淆网络的其他应用》，《COMPUT SPEECH LANG》，第 14 卷，第 4 期，第 373–400 页，2000。[76] D. J. MacKay，《反向传播网络的实用贝叶斯框架》，《神经计算》，第 4 卷，第 3 期，第 448–472 页，1992。

[77] A. Graves，《神经网络的实用变分推断》，载于 NeurIPS，2011，第 2348–2356 页。[78] J. J. Godfrey 等，《SWITCHBOARD：用于研究与开发的电话语音语料库》，载于 ICASSP，1992，第 517–520 页。[79] T. Hain 等，《AMI会议转录系统：进展与性能》，载于 MLMI，2006，第 419–431 页。[80] L. Gillick 等，《语音识别算法比较中的若干统计问题》，载于 ICASSP，1989，第 532–535 页。[81] A. Kumar 等，《端到端语音识别的话语置信度度量及其在分布式语音识别场景中的应用》，载于 INTERSPEECH，2020，第 4357–4361 页。[82] Q. Li 等，《用于语音识别的注意力序列到序列模型的置信度估计》，载于 ICASSP，2021，第 6388–6392 页。[83] W. Liu 等，《端到端儿童语音识别的话语级神经置信度度量》，载于 ASRU，2021，第 449–456 页。[84] Q. Li 等，《提升端到端语音识别域外数据上的置信度估计》，载于 ICASSP，2022，第 6537–6541 页。[85] N. Houlsby 等，《面向自然语言处理的参数高效迁移学习》，载于 ICML，2019，第 2790–2799 页。[86] T. Anastasakos 等，《置信度度量在语音识别器无监督自适应中的应用》，载于 ICSLP，1998，第 599–613 页。[87] F. Wallhoff 等，《面向大词汇量连续语音识别的帧判别式与置信度驱动自适应》，载于 ICASSP，2000，第 1835–1838 页。[88] C. Gollan 等，《用于声学模型自适应的置信度分数》，载于 ICASSP，2008，第 4289–4292 页。[89] X. Liu 等，《面向大词汇量连续语音识别系统组合的语言模型交叉自适应》，《计算机语音与语言》，第 27 卷，第 4 期，第 928–942 页，2013。[90] G. Zavaliagkos 等，《利用未转录训练数据提升性能》，载于 ICSLP，1998，第 1007–1011 页。[91] T. Kemp 等，《语音识别器的无监督训练：近期实验》，载于 EUROSPEECH，1999，第 2725–2728 页。[92] K. Yu 等，《大词汇量连续语音识别的无监督训练与定向人工转录》，《语音通信》，第 52 卷，第 7–8 期，第 652–663 页，2010。[93] A. Srinivasamurthy 等，《通过语义知识提取进行半监督学习，以提升空中交通管制中的语音识别效果》，载于 INTERSPEECH，2017，第 2406–2410 页。[94] N. Moreau 等，《利用基于帧级似然比的置信度度量拒绝错误数据》，载于 EUROSPEECH，1999，第 291–294 页。[95] K. Bartkova 等，《语音参数在基于隐马尔可夫模型的语音识别系统拒绝流程中的实用性》，载于 EUROSPEECH，1997，第 267–270 页。[96] S. Cox 等，《Switchboard数据库的置信度度量》，载于 ICASSP，1996，第 511–514 页。[97] C. Uhrik 等，《基于n元语法语言模型回退行为的置信度指标》，载于 EUROSPEECH，1997，第 2771–2774 页。[98] T. Kemp 等，《利用词格估计置信度》，载于 EUROSPEECH，1997，第 827–830 页。[99] F. Wessel 等，《大词汇量连续语音识别的置信度度量》，IEEE《语音与音频处理汇刊》，第 9 卷，第 3 期，第 288–298 页，2001。[100] M. S. Seigel 等，《结合信息源，基于条件随机场模型进行置信度估计》，载于 INTERSPEECH，2011，第 905–908 页。[101] K. Kalgaonkar 等，《使用循环神经网络估计自动语音识别结果的置信度分数》，载于 ICASSP，2015，第 4999–5003 页。[102] A. Kastanos 等，《使用格结构循环神经网络估计黑盒自动语音识别系统的置信度》，载于 ICASSP，2020，第 6329–6333 页。[103] R. Prabhavalkar 等，《少即是多：利用有限标签上下文和路径合并改进 RNN-T解码》，载于 ICASSP，2021，第 5659–5663 页。[104] D. S. Park 等，《SpecAugment：一种简单的自动语音识别数据增强方法》，载于 INTERSPEECH，2019，第 2613–2617 页。[105] G. Sun 等，《采用基于长短期记忆网络的跨话语信息表示的Transformer语言模型》，载于 ICASSP，2021，第 7363–7367 页。[106] M. Kitza 等，《BLSTM声学模型的累积自适应》，载于 INTERSPEECH，2019，第 754–758 页。[107] Z. T¨uske 等，《用于在Switchboard上取得最先进结果的单头注意力序列到序列模型》，载于 INTERSPEECH，2020，第 551–555 页。[108] W. Wang 等，《对端到端语音识别中基于音素的子词单元的研究》，载于 INTERSPEECH，2020，第 1778–1782 页。

[109] M. Zeineldeen 等，《基于Conformer的Switchboard数据集混合自动语音识别系统》，载于 *ICASSP*，2022，第 7437–7441 页。