# 端到端语音识别：综述

Rohit Prabhavalkar，*IEEE* 会员，Takaaki Hori，*IEEE* 高级会员，Tara N. Sainath，*IEEE* 会士，Ralf Schl¨uter，*IEEE* 高级会员，以及 Shinji Watanabe，*IEEE* 会士

arXiv:2303.03329v1 [eess.AS] 3 Mar 2023

摘要——**在过去十年的自动语音识别（ASR）研究中，深度学习的引入使词错误率相较于不采用深度学习的建模方法降低了 50% 以上（相对降幅）。伴随这一转变，研究者提出了多种全神经网络自动语音识别架构。这些所谓的端到端（E2E）模型是高度集成的纯神经网络自动语音识别模型，它们高度依赖通用机器学习知识，能够从数据中更一致地学习，同时较少依赖自动语音识别领域的专门经验。深度学习的成功及其广泛应用，加上更通用的模型架构，使端到端模型如今成为主流的自动语音识别方法。本综述旨在对端到端自动语音识别模型及其相应改进进行分类，并探讨其特性及其与基于经典隐马尔可夫模型（HMM）的自动语音识别架构之间的关系。本文涵盖端到端自动语音识别的所有相关方面：建模、训练、解码和外部语言模型集成，并讨论其性能与部署机会，同时展望未来可能的发展方向。**

索引词—**端到端、自动语音识别**。

## 一、引言

经典[1] 统计架构将自动语音识别系统分解为四个主要组成部分：从语音音频信号中进行声学特征提取、声学建模、语言建模，以及基于贝叶斯决策规则的搜索 [1], [2]。经典声学建模基于隐马尔可夫模型（HMM），以应对语速变化。在经典方法中，深度学习已被引入声学建模和语言建模。在声学建模中，深度学习取代了高斯混合分布（混合 HMM [3], [4]），或扩充了声学特征集（非线性判别/串联方法 [5], [6]）。在语言建模中，深度学习取代了基于计数的方法 [7]–[8], [9]。然而，在引入深度学习时，经典自动语音识别架构尚未受到触及。当今最先进的经典自动语音识别系统由许多彼此独立的组件和知识源构成，尤其包括语音信号预处理、针对录音条件的鲁棒性方法、音素清单和发音词典、语音聚类、对词汇表外词的处理、各种自适应/归一化方法、具有不同目标且包括序列判别式训练在内的精细训练计划等。另一方面，深度学习的潜力催生了将原先彼此分离的建模步骤进行整合的成功方法，例如将语音信号预处理和特征提取整合到声学建模中 [10], [11]。

更进一步地，将深度学习引入自动语音识别也推动了相关研究，即用更为一体化的联合神经网络模型结构取代基于隐马尔可夫模型（HMM）的经典自动语音识别架构 [12], [13]–[14], [15]。这些尝试可被视为以更通用的、面向序列到序列处理的机器学习方法，来替代专用的语音处理模型；这或许类似于自然语言处理中统计方法取代更偏重语言学的模型的方式。对于这些全神经方法，近年来已确立了端到端（E2E） [16], [17], [13], [18] 这一术语。然而，这一术语在许多方面缺乏区分。因此，本综述首先有必要尝试在自动语音识别的上下文中界定端到端 这一术语。

根据《剑桥词典》，形容词“端到端”的定义是：“涵盖一个过程的所有阶段”[19]。这可以从多个角度来理解：

*a)* 联合建模：就自动语音识别而言，端到端这一特性可以理解为将自动语音识别系统的所有组件作为一个计算图进行联合处理。更进一步地，自动语音识别领域对端到端的普遍理解是采用一种单一的联合建模方法，不一定区分各个独立组件；这也可能意味着摒弃将自动语音识别分为声学模型和语言模型的经典做法。

*b)* 单遍搜索：就识别／搜索问题而言，端到端这一特性可以理解为在作出决策之前，整合自动语音识别系统的所有组件（模型、知识源）。这与贝叶斯决策规则相符，因为该规则恰恰要求整合所有可用知识源，作出单一的全局决策。

*c)* 联合训练：在模型训练方面，端到端方法建议使用与当前任务相一致的单一目标函数，对模型所有组件的全部参数进行联合估计；对于自动语音识别而言，这意味着最小化预期词错误率。

*d)* 训练数据：集成模型的联合训练意味着使用单一类型的训练数据；对于自动语音识别而言，这类数据就是转录语音音频数据。然而，在自动语音识别中，通常还可以获得数量更多的纯文本数据，以及可选的未转录语音音频。因此，端到端建模面临的挑战之一是如何利用纯文本数据和纯音频数据 [20], [21]。

*e)* 从头训练：端到端特性也可以用于描述训练过程本身，即要求从头训练，避免使用先验对齐等外部知识，或使用基于不同准则和/或知识源预训练的初始模型。需要注意的是，在某些场景中，预训练和微调策略也很重要，例如模型具有明确的模块化结构，包括自监督学习

[1] 此处的“经典”指的是过去长期处于最先进水平的自动语音识别架构。该架构将声学模型与语言模型分离，并基于隐马尔可夫模型进行声学建模。

，[22] 或前端与语音识别模型的联合训练 [23]。

*f)* 次级知识源：对于自动语音识别而言，标准的次级知识源包括发音词典和音素集，以及语音聚类；在经典的最先进自动语音识别系统中，语音聚类通常基于分类与回归树（CART）[24]。次级知识源和单独训练的组件可能引入错误，可能与整体训练目标不一致，和/或产生额外成本。因此，端到端方法会避免使用这些资源。

*g)* 词汇建模：若不使用发音词典及相应的子词单元，端到端识别的词汇表就只能基于整词模型或字符模型。根据齐夫定律 [26]，整词模型 [25], 需要不切实际的大量转录训练数据才能支持大型词汇表，而许多任务可能无法获得足够的数据。另一方面，基于字符生成子词词汇表的方法，例如目前流行的字节对编码（BPE）方法 [27], ，可能会被视为端到端目标之外的次级方法。

*h)* 通用建模与知识驱动建模：最后，也可以从底层建模的通用性角度来看待端到端：任务特定约束是完全从数据中学习得到的，还是在系统架构建模之初就引入了任务特定知识？例如，自动语音识别中的单调性约束可以像基于注意力的端到端方法那样完全从数据中学习 [15]，也可以像经典的隐马尔可夫模型结构那样直接实现。

因此，端到端自动语音识别可以定义为一种集成式自动语音识别模型，它支持联合训练和识别，并始终以最小化预期词错误率为目标，同时避免使用单独获取的知识源。

然而，端到端方法应用于自动语音识别有哪些潜在优势？开发自动语音识别系统的首要目标是最小化预期词错误率。不过，次要目标包括降低最终解码器的时间和内存复杂度，并在开发预算有限的情况下，提高通用性和建模的简便性。

首先，将自动语音识别系统定义为单一神经网络结构意义上的端到端系统，有助于实现建模通用性，并可能在为新语言/领域构建自动语音识别系统时缩短开发周期。与经典建模相比，由单一神经网络结构定义的自动语音识别模型可能更加精简；由于无需整合多个独立模型，解码过程也变得更加简单。由此带来的内存占用和功耗降低有助于嵌入式自动语音识别应用[28]。

此外，联合训练端到端自动语音识别有助于避免中间训练阶段产生伪最优解。对于难以获得发音词典等资源的语言或领域，避免使用这类次级知识源可能会有所帮助。此外，次级知识源本身也可能存在错误。避免使用这些资源，有助于直接利用数据训练出更优的模型，前提是有足够的任务专用训练数据。

随着人们对端到端自动语音识别模型的兴趣日益高涨，相关研究也日趋多样，本文作者认为，现在正是概述这一快速发展的研究领域的时机。本综述旨在深入介绍端到端自动语音识别系统的当前研究现状，涵盖端到端自动语音识别的所有相关方面，并对不同的端到端和经典自动语音识别架构进行对比讨论。

本文对端到端语音识别的综述结构如下。第 II 节介绍端到端语音识别的历史演变，重点讨论输入/输出对齐，并概述当前具有代表性的基本端到端自动语音识别模型。第 III 节讨论基本端到端模型的改进，包括端到端模型组合、训练损失函数、上下文、编码器/解码器结构以及端点检测。第 IV 节概述端到端自动语音识别模型的训练。第 V 节讨论不同端到端方法所采用的解码算法。第 VI 节讨论（独立）语言模型在端到端自动语音识别中的作用及其整合方式。第 VII 节讨论新型端到端模型与经典的基于隐马尔可夫模型的自动语音识别方法之间的关系。第 VIII 节评述不同端到端方法以及经典自动语音识别方法的实验比较结果。第 IX 节概述端到端自动语音识别方法的应用。第 X 节探讨自动语音识别领域端到端研究的未来方向，最后在第 XI 节给出结论。

## II. 自动语音识别中的端到端模型分类体系

在讨论各种端到端建模方法的细节之前，我们首先介绍所采用的符号。我们将输入语音话语记为 $X$，并假设其已参数化为长度为 $T'$ 的 $D$ 维声学帧（例如，对数梅尔特征）：$X = (\mathbf{x}1, \cdots, \mathbf{x}_{T'})$，其中 $\mathbf{x}_t \in \mathbb{R}^D$。我们将相应的词序列记为 $C$，它可以分解为长度为 $L$ 的适当标签序列：$C = (c_1, \cdots, c_L)$，其中每个标签 $c_j \in \mathcal{C}$。我们的描述与将词序列分解为标签时所采用的具体表示无关；常见的选择包括字符、词或子词序列（例如，BPE [27],词片段 [29]）。所有端到端模型都包含一个编码器模块，记为 $H(X)$，它将长度为 $T'$ 的输入声学帧序列映射为高层表示，$H(X) = (\mathbf{h}1, \cdots, \mathbf{h}_T)$，其长度为 $T$（通常为 $T \leq T'$）。同样，我们的讨论与用于实现函数 $H(X)$ 的神经网络设计中所涉及的具体建模选择无关——例如，该网络可以采用循环神经网络（单向或双向） [30], 卷积网络或 Transformer [31]。给定一组训练样例 $\mathcal{T}$，其中包含 $N$ 对话语及其对应的转录文本：$\mathcal{T} = \{(X_i, C_i)\}_{i=1}^N$，我们的目标是训练一个模型，在给定输入声学信号的情况下，估计所有可能转录文本上的条件分布：$P(C|X) = P(C|H(X))$。

广义而言，使用端到端模型估计$P(C|X)$面临的挑战之一，是需要处理声学帧序列与相应标签序列之间未知的对齐；二者的长度分别为 $T$ 和 $L$。在经典自动语音识别模型中，这些帧级对齐

可以用隐马尔可夫模型（HMM）建模，同时使用生成式高斯混合模型（GMM）或神经网络对声学帧的输出分布进行建模；训练神经网络声学模型所需的帧级对齐，可以通过基于基础高斯混合模型-隐马尔可夫模型系统的强制对齐获得，但也可以进行无需初始对齐的直接序列训练 [32]。像经典自动语音识别系统通常采用的那样，对音素标签序列进行对齐似乎很自然；而端到端系统通常采用基于字符的（子）词单元。不过，字符级对齐的概念似乎不那么明确[2]。因此，在这项工作中，我们根据系统对这种对齐的建模方式——无论是显式还是隐式——对各种端到端建模方法进行分类。

早期的端到端建模方法通过潜变量对齐进行显式建模，并在训练和推理过程中将其边缘化（可能是近似地）。此类方法包括联结主义时序分类（CTC） [12], 、循环神经网络转导器（RNN-T） [13], 、循环神经对齐器（RNA） [33],和混合自回归转导器 [34]（HAT）。正如后续章节将要讨论的，这一类方法中的后续模型对对齐的建模日益复杂，独立性假设也越来越少，因此能力越来越强。与之相对的是基于注意力的编码器-解码器模型，它们最初在机器翻译 [35]领域广受关注。与 CTC 等显式对齐模型不同，基于注意力的编码器-解码器模型使用注意力机制 [36] 来学习整个声学序列与各个标签之间的对应关系。此外，还有一些研究处于这两个极端之间：神经转导器 [37], 或基于单调对齐 [38] 及其变体的模型（例如单调分块对齐（MoChA） [39],、单调无限回看（MILK） [40] 等）采用显式对齐模型，同时也利用注意力机制，使模型能够检查局部声学信息以优化预测。本节余下部分将依次介绍这些模型。

### *A.* 显式对齐端到端方法

CTC、RNN-T 或 RNA 等显式对齐方法，为长度为 $T$ 的编码器输出 $H(X)$ 与长度为 $L$ 的输出序列之间的对齐定义了一个显式潜变量；每种方法对对齐的定义各不相同，进而决定了相应模型所采用的概率条件独立假设，以及模型的训练和解码过程。所有显式对齐模型的一个共同特征是引入了一个额外的 空白符符号，记为 $\langle \mathrm{b} \rangle$，并定义了符号上的输出概率分布

[2] 例如，在英语等语言中，正字法与发音之间的对应关系并不直接：像 knife 这样的单词可以分解为其组成音素 n ay f，这些音素可以与各个声学帧相关联；但如果将其表示为字符序列，那么单词中不发音的字符 k 应对应哪些帧，就不那么明确了。

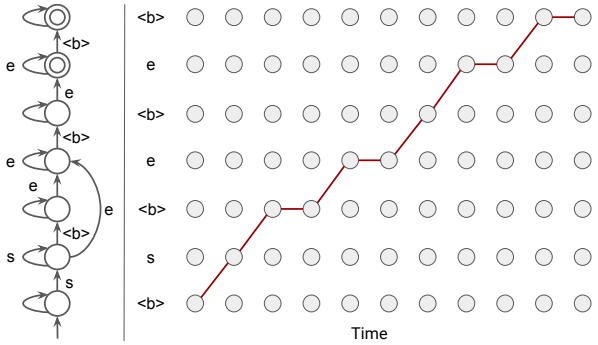


图 1. 目标序列为 $C$ =(s，e，e) 的 CTC模型示例对齐序列（右），以及表示所有有效对齐路径集合的（非确定性）有限状态自动机（FSA） [41]（左）。

设置 $\mathcal{C}_b = \mathcal{C} \cup \{\langle \mathrm{b} \rangle\}$。正如我们将在下文中更详细地讨论的那样，$\langle \mathrm{b} \rangle$ 符号在这些模型中的含义略有不同。现在，只需说明的是，对于一个特定的训练样例（$X$、$C$），每个模型都定义了一组有效对齐，记为 $\mathcal{A}_{(T,C)}$，并通过对所有有效对齐序列进行边缘化来定义条件分布 $P(C|X)$:

$$P(C|X) = P(C|H(X)) = \sum_{A} P(C|A, H(X))P(A|H(X)) = \sum_{A \in \mathcal{A}_{(T=|H(X)|,C)}} P(A|H(X)) \tag{1}$$

其中，根据定义，$P(C|A,\ H(X)) = 1$ 当且仅当 $A \in \mathcal{A}_{(T,C)}$，否则为 0；这也意味着，从对齐到标签序列 $C$ 的映射被定义为唯一映射。我们将在后续章节中讨论这些模型各自的具体形式。

1) 联结主义时序分类（CTC）：Graves 等人提出了联结主义时序分类（CTC） [12]，将输入词元序列映射为相应输出词元序列的一种技术。CTC 通过引入一个特殊的“空白符”标签（记为 $\langle \mathrm{b} \rangle$：$\mathcal{C}_b = \mathcal{C} \cup \{\langle \mathrm{b} \rangle\}$），显式地对编码器输出 $H$（$X$）与标签序列 $C$ 之间的对齐进行建模。因此，对齐 $A \in \mathcal{C}_b^*$ 是由 $\mathcal{C}$ 或 $\langle \mathrm{b} \rangle$ 构成的标签序列。给定一个特定的训练样例（$X$，$C$），我们将所有有效对齐的集合记为 $\mathcal{A}^{\mathrm{CTC}}_{(X,C)} = \{A = (a_1,\ a_2,\ \ldots,\ a_T)\}$，其中每个 $a_t \in \mathcal{C}_b$，并且还需满足以下约束：先合并连续重复的标签，再移除所有空白符号后，$A$ 与 $C$ 相同。例如，若 $T = 10$，且 $C =$（s，e，e），则 $A =$（s $\langle \mathrm{b} \rangle$，$\langle \mathrm{b} \rangle$，e，e $\langle \mathrm{b} \rangle$，e，e $\langle \mathrm{b} \rangle$，$\langle \mathrm{b} \rangle$）$\in \mathcal{A}^{\mathrm{CTC}}_{(X,C)}$，如图1所示。从这个例子可以看出，输出中的重复标签可以通过在其间插入空白符来表示。

根据公式（1），CTC通过以下方式定义给定输入 $X$条件下标签序列 $C$ 的后验概率:

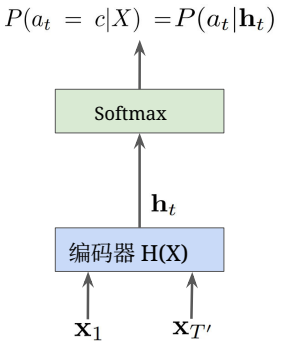


图 2. CTC模型的示意图。该模型由一个编码器和一个Softmax层组成：编码器将输入语音映射为高层表示，Softmax层则预测输出标签集合及空白符的帧级概率。

对所有可能的CTC对齐进行边缘化：

$$\begin{aligned} P_{\text{CTC}}(C|X) &= \sum_{A \in \mathcal{A}^{\text{CTC}}_{(X,C)}} P(A|H(X)) \\ &= \sum_{A \in \mathcal{A}^{\text{CTC}}_{(X,C)}} \prod_{t=1}^{T} P(a_t|a_{t-1}, \cdots, a_1, H(X)) \\ &= \sum_{A \in \mathcal{A}^{\text{CTC}}_{(X,C)}} \prod_{t=1}^{T} P(a_t|\mathbf{h}_t) \end{aligned} \tag{2}$$

关键在于，正如公式 (2) 所示，CTC 作出了一个很强的独立性假设：给定时间 $t$的局部编码器输出后，模型在时间 $t$ 的输出与其他时间步的输出条件独立。

因此，如图 2 所示，CTC模型由一个神经网络组成，该网络在每个步骤对分布 $P(a_t|X)$ 进行建模。例如，可以使用一个神经网络来实现这一点，我们将其称为编码器；常见的选择包括深度卷积网络，或深度循环神经网络（RNN），例如采用长短期记忆网络（LSTM）单元的网络 [30]。编码器（可以是单向或双向的）将输入声学帧 X 转换为编码表示 $H = (\mathbf{h}_1, \cdots, \mathbf{h}_T)$，其中$h_t \in \mathbb{R}^k$。编码器连接到一个Softmax层，其$|\mathcal{C}_b|$ 目标表示公式 (2) 中的各个概率：$P(a_t = c|X) = P(a_t = c|H(X))$。因此，在每个步骤 $t$，模型都会读取一个编码帧 $h_t$ ，并输出一个标签上的分布；换言之，模型“输出”一个单个标签，可以是空白符 $\langle \mathbf{b} \rangle$，也可以是 $\mathcal{C}$ 中的一个目标标签。

*2)* 循环神经网络转换器（*RNN-T*）：循环神经网络转换器（RNN-T） [13], [42] 由 Graves 提出，是对基本 CTC模型 [12], 的改进，它移除了一些我们先前讨论过的条件独立性假设。图 3 所示的 RNN-T 模型，最好通过与 CTC模型对比来理解。与 CTC 一样，RNN-T 模型在输出符号中加入空白符号，从而定义了在 $\mathcal{C}_b$上的标签序列分布。同样，与 CTC 一样，该模型由一个编码器组成，用于处理输入声学帧 X，生成编码表示 $H(X) = (\mathbf{h}_1, \cdots, \mathbf{h}_T)$。

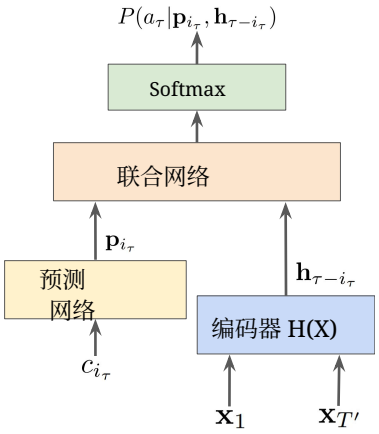


图 3. RNN-T 模型 [13], [42] 由一个编码器和一个预测网络组成。编码器将输入语音帧转换为高层表示，预测网络则对先前已输出的非空白标签序列进行建模。预测网络的输出，

$p_{i_t}$，表示输出先前的非空白标签序列 $c_1, \ldots, c_{i_t}$后的结果。联合网络根据预测网络状态和特定的编码帧，生成关于输出符号（扩展后包含空白符）的概率分布。

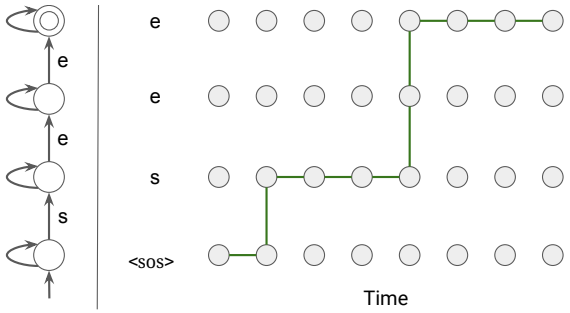


图 4. 目标序列为 $C$ = (s,e,e) 的 RNN-T 模型的示例对齐序列（右）。图中的水平转移对应空白符输出。有限状态自动机（左）表示所有有效 RNN-T 对齐路径的集合。

然而，与 CTC 不同，RNN-T 中的空白符号有着略微不同的含义：对于每个输入编码器帧 $\mathbf{h}_t$，RNN-T 模型会输出 $\mathcal{C}$ 中由单个空白符号终止的零个或多个符号序列。因此，我们可以将 RNN-T 中所有有效对齐序列的集合定义为：$\mathcal{A}^{\text{RNNT}}(X, C) = \{A = (a_1, a_2, \cdots, a_{T+L})\}$，即 $\mathcal{C}_b^*$中所有由 $T + L$ 符号构成的序列；去掉其中所有空白符后，这些序列与 $C$ 相同。最后，对于给定的输出位置 $\tau$，令 $i_\tau$ 表示部分序列 $(a_1, \cdots, a_{\tau-1})$ 中非空白标签的数量。因此，部分序列 $(a_1, \cdots, a_{\tau-1})$ 中空白符的数量为$\tau - i_\tau - 1$。例如，若 $T = 7$，且 $C$ = (s, e, e)，则 $A = (\langle \mathbf{b} \rangle, \text{s}\ \langle \mathbf{b} \rangle, \langle \mathbf{b} \rangle, \langle \mathbf{b} \rangle, \text{e, e}\ \langle \mathbf{b} \rangle, \langle \mathbf{b} \rangle, \langle \mathbf{b} \rangle) \in \mathcal{A}^{\text{RNNT}}_{(X,C)}$。请注意，与 CTC模型不同，如图 4 所示，输出中的重复标签无需特殊处理；$i_1 = i_2 = 0$； $i_3 = i_4 = 1$； $i_{10} = 3$；等等。

我们接着可以将后验概率 $P(C|X)$ 定义为

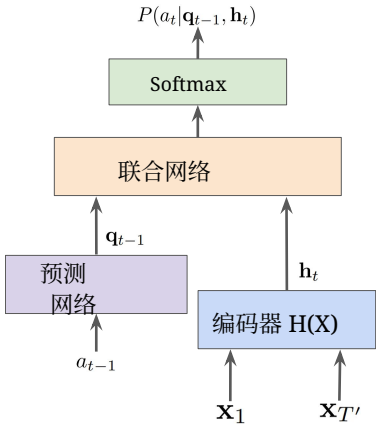


图 5. RNA 模型 [33] 在模型结构上与 RNN-T 模型 [13], [42] 相似。不过，该模型在单个帧中只允许输出一个标签——空白符或非空白标签。与 RNN-T 不同，RNA 模型中的预测网络状态 $\mathbf{q}_{t-1}$取决于整个对齐序列 $a_{t-1},...,a_1$。联合网络根据预测网络状态和特定的编码帧，生成关于输出符号（包含空白符）的概率分布。

之前：

$$
\begin{aligned}
P_{\text{RNNT}}(C|X) &= \sum_{A\in\mathcal{A}^{\text{RNNT}}_{(X,C)}} P(A|H(X)) \\
&= \sum_{A\in\mathcal{A}^{\text{RNNT}}_{(X,C)}} \prod_{\tau=1}^{T+L} P(a_\tau|a_{\tau-1},\ldots,a_1,H(X)) \\
&= \sum_{A\in\mathcal{A}^{\text{RNNT}}_{(X,C)}} \prod_{\tau=1}^{T+L} P(a_\tau|c_{i_\tau},c_{i_\tau-1},\ldots,c_0,\mathbf{h}_{\tau-i_\tau}) \\
&= \sum_{A\in\mathcal{A}^{\text{RNNT}}_{(X,C)}} \prod_{\tau=1}^{T+L} P(a_\tau|\mathbf{p}_{i_\tau},\mathbf{h}_{\tau-i_\tau})
\end{aligned}
\tag{3}
$$

其中，$P=\mathbf{p}_1,\cdots,\mathbf{p}_L$ 表示图 3 所示的预测网络的输出。该网络概括了此前预测的非空白标签序列，并由另一个神经网络实现：$\mathbf{p}_j = NN(\cdot|c_0,\ldots,c_{j-1})$，其中$c_0$ 是一个特殊的句首标签，〈sos〉。因此，正如公式 (2) 所示，RNN-T 减少了 CTC 中的一些独立性假设，因为我们对时间 t 的输出建模时，认为它在条件上依赖于此前非空白预测的序列，但独立于具体的对齐选择（即发出非空白词元的帧的选择）。

最后，我们指出，[43] 中的工作通过允许将有效的帧级对齐表示为任意图，推广了基本的 RNN-T 模型；为计算条件概率分布所需的前向后向算法也可以相应地推广到图上。例如，这种表述允许在 RNN-T 模型中使用类似 CTC 的对齐（即在每一帧输出单个标签——空白符或非空白标签），同时像 RNN-T 模型那样以先前的非空白符号集合为条件。

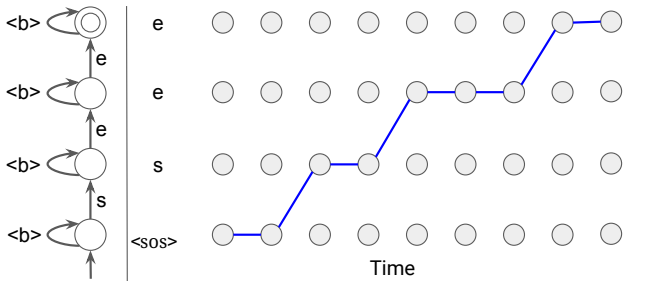


图 6. 目标序列为 $C=$(s，e，e) 的 RNA 模型的示例对齐序列（右）。图中的水平转移对应于输出空白符；对角转移对应于输出非空白符号。有限状态自动机（左）表示 RNA 模型的有效对齐集合。尽管该有限状态自动机与图 4 中对应的 RNN-T 有限状态自动机相同，但 ⟨b⟩ 标签在两种情况下的语义不同。

*3*）循环神经对齐器（*RNA*）：循环神经对齐器（RNA）由 Sak 等人 [33] 提出。RNA 模型通过去除一项条件独立假设对 RNN-T 模型进行了推广。图 5 所示的模型，最好通过考察它与 RNN-T 模型的差异来理解。与 CTC 和 RNN-T 一样，RNA 模型在集合 $\mathcal{C}_b$ 中定义了扩展了空白符的标签上的概率分布，其中 ⟨b⟩ 的语义与 CTC 模型中的相同：在每一帧，模型只能输出单个标签——空白符或非空白标签——然后进入下一帧；与 CTC 不同（但与 RNN-T 一样），模型对每个非空白标签只输出一次。更具体地说，RNA 模型中的有效对齐集合，$\mathcal{A}^{\text{RNA}}_{(X,C)}=(a_1,\cdots,a_T)$，由长度为 $T$ 的 $\mathcal{C}^*_b$ 中的序列组成，这些序列恰好包含 $T-L$ 个空白符号，且移除所有空白符后与 $C$ 相同。因此，空白符号在 RNA 和 RNN-T 模型中有着不同的含义：在 RNN-T 中，输出空白符号会使模型进入下一帧；而在 RNA 中，模型在输出一个空白或非空白标签后便进入下一帧。限制模型在每一帧只输出一个非空白标签，可以通过限制每一帧的模型扩展次数（参见 RNN-T 解码）提高计算效率并简化解码过程。例如，如果 $T=8$，且 $C=$(s，e，e)，那么$A=$(⟨b⟩，s ⟨b⟩，e ⟨b⟩，⟨b⟩，e ⟨b⟩) $\in\mathcal{A}^{\text{RNA}}_{(X,C)}$，如图 6 所示。

RNA 后验概率 $P(C|X)$ 定义如下：

$$
\begin{aligned}
P_{\text{RNA}}(C|X) &= \sum_{A\in\mathcal{A}^{\text{RNA}}_{(X,C)}} P(A|H(X)) \\
&= \sum_{A\in\mathcal{A}^{\text{RNA}}_{(X,C)}} \prod_{t=1}^{T} P(a_t|a_{t-1},\ldots,a_1,H(X)) \\
&= \sum_{A\in\mathcal{A}^{\text{RNA}}_{(X,C)}} \prod_{t=1}^{T} P(a_t|\mathbf{q}_{t-1},\mathbf{h}_t)
\end{aligned}
\tag{4}
$$

其中，如前所述，$i_t$ 表示部分对齐序列（$a_1$，.....，$a_{t-1}$）中的非空白标签数量，而$\mathbf{q}_{t-1}=$NN(·|$a_{t-1}$，$\cdots$，$a_1$)表示神经网络的输出，该网络概括了整个部分对齐

序列，其中 NN(·)表示一个合适的神经网络（ [33]中的长短期记忆网络）。因此，RNA 通过以先前的非空白标签序列以及生成这些标签的对齐为条件，消除了 RNN-T 模型中仅存的一个条件独立假设。

由于 RNA 模型不仅以非空白标签序列（如 RNN-T 中那样）为条件，还以这些标签发出的具体帧为条件，因此无法精确计算式（3）中的对数似然（及相应的梯度）。RNA 转而采用一种简化假设：利用解码器 RNN NN(·) 中与最可能对齐路径相对应的 RNN-T 状态，同时利用模型在每一帧只能输出单个标签这一约束。后一约束——每帧只允许输出单个标签（空白或非空白）——也已在单调 RNN-T 模型 [44] 的背景下进行过探索。最后，我们指出， [45] 中的工作通过在定义状态时使用两个 RNN，进一步推广了 RNA 模型：一个慢速RNN（对应于先前预测的非空白标签序列），以及一个快速RNN（还以非空白标签输出时所在的帧为条件）。

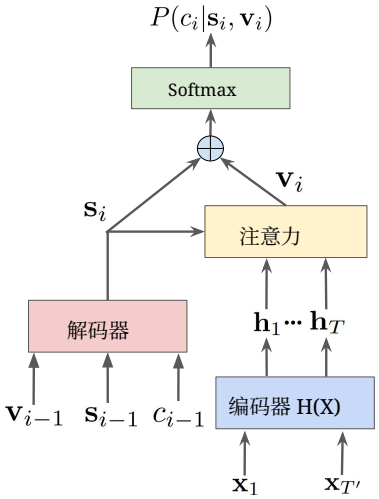


图 7. 基于注意力的编码器-解码器（AED）模型 [14], [15], [46]。输出分布以解码器状态$\mathbf{s}_i$ （概括先前已解码符号）和上下文向量$\mathbf{v}_i$ （根据解码器状态概括编码器输出）为条件。在 Chan 等人的开创性工作中， [15], 例如，这是通过拼接这两个向量来实现的，如图中的 $\bigoplus$ 符号所示。

*B.* 隐式对齐端到端方法

显式对齐方法（如 CTC、RNN-T 或 RNA）的主要优势之一，是它们能够构建易于流式处理 的自动语音识别模型——也就是说，假设用于生成编码器特征 $H(X)$ 的神经网络能够仅依据先前声学特征的序列，并在有限前瞻范围内生成编码帧，那么模型就可以在语音输入系统的同时输出假设。然而，有些应用可能适合离线语音处理：也就是说，在产生任何输出之前，先将全部输入语音送入系统。本节将讨论基于注意力的编码器-解码器（AED）模型（也称为 listen-attend-and-spell（LAS））[14], [15], [46], ，该模型利用注意力机制 [36] 隐式识别并建模输入声学信号中与各个输出单元相关的部分。因此，在基本的 AED 模型中，注意力机制免去了显式建模对齐的需要。

在第 II-A 节介绍的流式显式对齐方法中，推理期间，模型会持续输出符号，直到处理完位于

此时解码过程即告完成；类似地，在训练过程中，前向后向算法会遍历所有可能的对齐序列。由于注意力编码器-解码器模型会一次性处理整个声学序列，因此模型需要一种机制来表明所有输出符号均已发出。实现方式是在输出集合中加入句末符号 ⟨eos⟩，使输出词汇表由集合 $\mathcal{C}_{\text{eos}} = \mathcal{C} \cup \{\langle\text{eos}\rangle\}$ 构成。因此，图 7 所示的注意力编码器-解码器模型由编码器网络和基于注意力的解码器组成：编码器网络将输入声学帧序列 $X = (\mathbf{x}_1, \ldots, \mathbf{x}_{T'})$ 编码为更高层次的表示 $H(X) = (\mathbf{h}_1, \ldots, \mathbf{h}_T)$；基于注意力的解码器则定义了输出符号集合 $\mathcal{C}_{\text{eos}}$ 上的概率分布。因此，给定一个配对训练样例 $(X, C)$，

我们用 $C_e = (c_1, \ldots, c_L, \langle\text{eos}\rangle)$ 表示在长度为 $(L+1)$ 的真实符号序列后添加句末符号 ⟨eos⟩所得的序列。注意力编码器-解码器模型计算添加句末符号 ⟨eos⟩ 后的输出序列的条件概率，如下所示：

$$\begin{aligned} P(C_e|X) &= P(C_e|H(X)) \\ &= \prod_{i=1}^{L+1} P(c_i|c_{i-1}, \ldots, c_0 = \langle\text{sos}\rangle, H(X)) \\ &= \prod_{i=1}^{L+1} P(c_i|c_{i-1}, \ldots, c_0 = \langle\text{sos}\rangle, \mathbf{v}_i) \\ &= \prod_{i=1}^{L+1} P(c_i|\mathbf{s}_i, \mathbf{v}_i) \end{aligned} \tag{5}$$

其中，$\mathbf{v}_i$ 表示一个上下文向量，它概括了编码器输出中与先前预测序列 $c_{i-1}, \ldots, c_0$相关的部分，即 $H(X)$；而$\mathbf{s}_i$表示输出先前符号序列后对应的解码器状态，该状态通过基于先前的上下文向量和输出标签更新解码器状态而生成：

$$\mathbf{s}_i = \text{Decoder}(\mathbf{v}_{i-1}, \mathbf{s}_{i-1}, c_{i-1})$$

符号 $c_0 = \langle\text{sos}\rangle$ 是一种特殊的句子起始符号，在基于注意力的解码器生成任何输出之前，它会作为第一个输入。如公式 (5) 所示，与 CTC 或 RNN-T 等模型相比，注意力编码器-解码器模型的一项重要优势在于，它们不对模型输出与输入声学信号之间作任何独立性假设，因此比隐式对齐模型更通用；同时，由于无需对所有可能的对齐序列进行显式边缘化，其训练和实现也要容易得多。不过，这也有代价：先前生成的上下文向量（类似于显式对齐模型中已解码的部分对齐）

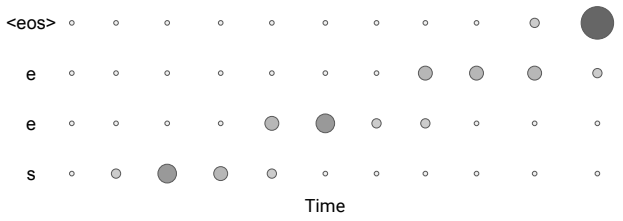


图 8。与 RNN-T 或 CTC 等模型不同，注意力编码器-解码器模型没有显式对齐。不过，可以将特定输出符号 $c_i$ 的注意力权重$\alpha_{t,\ i}$解释为对齐权重；上图展示了目标序列 $C$ =(s、e、e ⟨eos⟩) 对应的对齐权重。在此表示中，圆圈大小和颜色深浅与相应的注意力权重成正比；因此，每一行的总概率质量都相同。如上图所示，最初的几帧对应第一个符号 $c_1$ =s，而后面的帧对应第二个“e”：$c_3$ =e。

在解码过程中不会被修正。换句话说，尽管编码器处理 $H(X)$ 可能是双向的，注意力编码器-解码器模型中的解码过程却呈现出左右不对称性 [47]。

*1*）注意力编码器-解码器模型中的上下文向量计算：如前所述，上下文向量 $\mathbf{v}_i$ 是通过采用注意力机制 [36] 计算得到的。这些方法背后的核心思想是定义一个状态向量$\mathbf{s}_i$ ，该向量对应于模型输出 $c_1$、.....、$c_{i-1}$后的状态。注意力函数 atten($\mathbf{h}_t$，$\mathbf{s}_i$) $\in \mathbb{R}$ 随后定义模型输出 $i-1$ 个前置符号后的状态与 $H(X)$ 中每个编码帧之间的分数。然后，可以使用 softmax函数对这些分数进行归一化，从而定义一组与每个$\mathbf{h}_t$ 对应的权重，如下所示：

$$\alpha_{t,i} = \frac{\exp\{\text{atten}(\mathbf{h}_t, \mathbf{s}_i)\}}{\sum_{t'=1}^{T} \exp\{\text{atten}(\mathbf{h}_{t'}, \mathbf{s}_i)\}}$$

直观地说，权重 $\alpha_{t,i}$ 表示在模型已经输出符号 $c_1, \ldots, c_{i-1}$ 之后，输出下一个符号$c_i$时，某个特定编码帧$\mathbf{h}_t$ 的相关性，如图 8 所示。上下文向量根据计算得到的注意力权重对编码器输出进行概括：

$$\mathbf{v}_i = \sum_t \alpha_{t,i} \mathbf{h}_t$$

文献中探讨了多种可能的注意力机制：最常见的形式包括点积注意力 [15] 和加性注意力 [36]，，此外，文献中也探讨了许多其他变体（例如高斯注意力 [48], 等）。除了选择具体的注意力机制外，一种常见的提升性能的方法是采用多个独立的注意力头——$\mathbf{v}_i^1, \ldots, \mathbf{v}_i^K$——然后将它们拼接起来，得到最终的上下文向量 $\mathbf{v}_i = \left[\mathbf{v}_i^1; \ldots; \mathbf{v}_i^K\right]$，这就是所谓的多头注意力方法 [31],; 另一种做法则是将多个基于注意力的层堆叠起来，如 Vaswani 等人提出的 Transformer 解码器 [31]。

### *C.* 基于注意力并采用对齐建模的端到端方法

注意力编码器-解码器模型不作任何条件独立性假设，功能极其强大，表现往往优于 CTC 或 RNN-T [49] 等显式注意力端到端方法。不过，这些模型也存在一些显著缺点，最突出的是它们本质上无法进行流式处理——也就是说，必须处理完所有声学帧后才能生成任何输出假设。另一个相关问题是，这些模型对声学序列的长度极为敏感 [50], ，因此处理长音频时需要进行特殊处理 [50]。

已有许多研究探讨了构建流式 AED 模型的技术。一般来说，这些模型的动机来自这样一种观察：语音（不同于机器翻译等任务）在编码帧（假设编码器是单向的）与输出单元之间呈现“局部”关系。因此，与通过对所有输入帧 $\mathbf{v}_i$求和来计算上下文向量的通用注意力编码器-解码器模型不同，各种提出的模型将这一求和限制在部分帧上进行，从而实现流式解码。就本文的阐述而言，最容易将这些模型理解为具有底层对齐（无论对齐已知与否），并可利用该对齐执行流式推理。

神经转导器（NT） [37] 会将输入编码器帧明确划分为 $T^W$ 长度为$W$的非重叠块：$H_1^W = [\mathbf{h}_1, \ldots\ldots, \mathbf{h}_W]$；$\cdots$；$H_{T^W}^W = [\mathbf{h}_{T^W+1}, \ldots\ldots, \mathbf{h}_{T^W W}]$，其中 $T^W = \left\lceil \frac{T}{W} \right\rceil$，且当 $t > T$时$\mathbf{h}_t = \mathbf{0}$。与计算上下文向量时会考察所有编码帧的注意力编码器-解码器模型不同，NT 模型每次只能处理一个块；只有在输出特殊的块结束符号（类似于注意力编码器-解码器模型中的 ⟨eos⟩ ）后，模型才会推进到下一个块；当模型在最后一个块中输出块结束符号 $H_{T^W}^W$时，推理便终止。如果未知真实输出序列 $C$相对于长度为 $W$的块的对齐情况，则可以先使用一种粗略的初始对齐来训练系统，将符号平均分配到 $T^W$ 个块中；随后通过根据当前模型参数计算最可能的输出对齐 [37] ，进行迭代细化，这与基于隐马尔可夫模型的系统中的强制对齐类似。另一种方法 [51] 是使用单独的系统（例如传统的混合系统）获取初始对齐（例如词级对齐），再据此将子词单元分配到各个块中。

Raffel 等人提出了另一种方法 [38],，通过显式引入一个对齐模块来修改标准注意力编码器-解码器模型。该模块从左到右扫描编码器帧 $H$（$X$），以确定当前帧是否应被用于生成输出，这一过程被建模为一个伯努利随机变量。如果某一帧 $\tau$被选中，则模型根据局部编码器帧$\mathbf{h}_\tau$生成输出。生成下一个输出后，模型从当前选中的帧开始重复这一过程，从而允许在同一帧生成多个输出。由于模型受到流式处理方式下生成输出的约束，这就形成了输入语音与输出标签之间的硬单调对齐。一种单调分块注意力（MoChA）模型 [39]在 Raffel 等人工作的基础上进行了改进，使模型能够利用在所选帧

左侧局部帧窗口上通过注意力计算得到的上下文向量来生成下一个输出 $\tau$：$\mathbf{h}_{\tau-W+1}, \ldots, \mathbf{h}_{\tau}$。因此，MoChA 模型由两个层次的过程组成：确定应生成输出的帧 [38], ，随后在所选帧左侧的帧上运行一个注意力编码器-解码器模型。Arivazhagan 等人提出了对 MoChA 模型的改进 [40] ——单调无限回溯（MILK）注意力模型——该模型在每一步都基于所选帧左侧的所有帧 $\tau$（即 $\mathbf{h}_1, \ldots, \mathbf{h}_{\tau}$）计算上下文向量。文献 [52]提出了另一种实现流式处理的两阶段方法，称为触发式注意力：使用 CTC 网络触发，即控制一个具有有限解码器延迟的注意力编码器-解码器模型的激活。在文献 [53], 中比较了多种局部单调注意力变体。文献 [54] 讨论了分段注意力变体，而文献 [55] 研究了相关解码以及分段长度建模的重要性，从而改善了对长序列的泛化能力。分段注意力模型与转导器模型[56]相关。然而，分段式端到端自动语音识别模型并不局限于基于注意力机制实现；它们不仅可能与直接隐马尔可夫模型 [57], 相关，而且还被证明等价于神经转导器建模 [58]。

## III. 基本端到端模型的架构改进

在本节中，我们将进一步介绍前文所述的基础端到端模型。具体而言，我们描述了对这些模型进行的各种算法改进；与传统自动语音识别系统相比，这些改进对于提升性能至关重要。首先，我们介绍组合不同互补端到端模型、从而提升性能的多种方法。接下来，我们介绍将上下文融入这些模型的方法，以提升模型对罕见专有名词实体的识别性能。随后，我们描述改进后的编码器和解码器架构，这些架构能够更充分地利用 TPU 上的众多核心。最后，我们讨论如何通过集成式端到端语音端点检测器降低模型延迟。下文将进一步讨论这些主题的详细内容。

### *A.* 模型的组合

不同的端到端模型之间存在互补性，因此已有大量尝试着眼于组合这些方法。

例如， [59] 发现，基于注意力的模型在处理较长或含噪语句时表现很差，主要原因在于模型能够看到完整语句，因此在预测对齐时具有过大的灵活性。相比之下，CTC 等模型在解码过程中具有从左到右的约束，在这些情况下表现好得多。该论文研究了一种同时采用 CTC 和基于注意力的模型的多任务学习策略；在 WSJ 和 CHiME 任务上，与基于注意力的模型相比，该策略使词错误率相对降低了 5%–14%。

此外， [60] 探讨了如何结合 RNN-T 和注意力编码器-解码器的优势。具体来说，RNN-T 以从左到右的方式解码话语，这也已显示出对较长且含噪话语的改进。另一方面，由于注意力编码器-解码器能够看到整个话语，因此在需要借助周围上下文来预测当前词语的话语中，它往往表现更好，例如，数字（1.15）。为结合 RNN-T 和注意力编码器-解码器，作者尝试先用 RNN-T 生成第一遍解码，再用注意力编码器-解码器进行重打分。为减少计算量，作者尝试让 RNN-T 和注意力编码器-解码器共享编码器。作者发现，在一项语音搜索任务中，RNN-T+注意力编码器-解码器与单独使用 RNN-T 相比，词错误率相对降低了 17%–22%。此外，作者还探索了其他流式第一遍解码后接基于注意力的第二遍重打分的方法，例如审议 [61]。

重打分存在的一个问题是，我们受限于第一遍系统生成的格。为了解决这一问题，人们也探索了运行第二遍束搜索的方法。为了与传统自动语音识别系统具有竞争力，第二遍处理必须采用流式处理，以便能够在实时内完成。基于这一点，CascadedEncoder[62],Y架构 [63]以及通用自动语音识别 [64], 便是这样的例子。

### *B.* 融入上下文

针对特定领域进行上下文偏置，包括用户的歌曲名称、应用名称和联系人姓名，是任何生产级自动语音识别（ASR）系统的重要组成部分。在端到端模型中，上下文偏置尤其具有挑战性，因为这些模型在束搜索期间只保留少量候选项，而且对专有名词的识别效果也较差，而专有名词正是偏置短语的主要来源。文献中已经提出了一些融入上下文的方法。

一种方法是构建一个独立的加权有限状态转换器（FST），以表示偏置短语。在束搜索过程中，将偏置 FST 的得分与端到端模型的得分进行插值，并特别注意确保不会对短语产生过度或不足的偏置。这种技术被称为浅层融合上下文偏置 [65]。

另一种方法则考察以全神经方式将偏置短语注入模型。例如，[66] 通过嵌入向量表示一组偏置短语。这些向量作为额外输入馈送给基于注意力的模型，随后该模型可以选择关注这些短语，从而提高预测这些短语的概率。此外， [67] 研究了以全神经方式向对话上下文进行偏置。除此之外， [68]考察了利用音素和字素信息来嵌入偏置短语。最后， [69] 使用话语级上下文向量，类似于通过对子网络获得的逐帧隐藏状态向量进行池化而计算得到的 i-vector（该子网络称为序列摘要网络）。

### *C.* 编码器/解码器结构

随着时间推移，端到端模型的编码器架构不断改进。最早的端到端模型在编码器和解码器中都采用了长短期记忆循环神经网络（长短期记忆网络）。这些序列模型的主要缺点是，每一帧都依赖于前一帧的计算，因此无法并行批处理多帧。

随着硬件的改进，尤其是配备数千个核心的端侧边缘张量处理单元（TPU），人们开始探索能够更充分利用这些

硬件的架构。这类架构包括基于卷积的架构，例如 ContextNet [70]。在《用于自动语音识别的Transformer》中，研究人员探索了使用自注意力取代长短期记忆网络中的序列递归 [71], [72]。最后，人们还探索了将自注意力与卷积相结合（称为 Conformer [73], ），或与多层感知机相结合 [74], 。Transformer 和 Conformer 在许多任务上的性能均已显示出与长短期记忆网络相当的竞争力 [75], [76]。

在解码器方面，针对转导器模型的研究表明，可以用一个简单的嵌入查找表取代大型长短期记忆网络解码器，该查找表仅关注模型 [34], [77], [78], [79], [80]中前面少数几个词元。这表明，端到端模型的大部分能力都来自编码器，这一直是端到端模型和经典混合隐马尔可夫模型的共同特点。

### *D.* 集成式端点检测

流式语音识别系统的一个重要特点是必须快速进行端点检测，以便确定自动语音识别结果，并将其发送到服务器执行相应操作。通常，端点检测由外部语音活动检测器完成。由于端点检测既涉及声学模型，也涉及语言模型决策，近期有关流式 RNN-T模型 [81], [82]的研究尝试在话语末尾预测一个麦克风关闭词元 ⟨eos⟩ 。例如，我们的句子可能是“今天天气怎么样 ⟨eos⟩”。研究表明，将端点检测决策作为模型的一部分进行预测，可以改善延迟与词错误率之间的权衡。此外， [83], [84]还探索了使用 CTC 空白符号进行端点检测。

## IV. 端到端模型训练

总体而言，端到端模型的训练遵循深度学习方案 [85], [86],，同时特别考虑自动语音识别中需要处理的序列结构和潜在对齐问题。尽管可能需要精心设计训练计划并进行大规模数据增强，端到端自动语音识别模型仍可采用端到端方式训练。端到端模型的一大优势是，它们不假设输入帧之间相互条件独立。给定训练集 $\mathcal{T} = \{(X_i,\ C_i)\}_{i=1}^{N}$，需要最小化的训练准则 $\mathcal{L}$ （这等价于最大化总条件对数似然）可写为： $\mathcal{L} = \sum_{i=1}^{N} \log\ P(C_i|X_i)$ 。

### *A.* 训练中的对齐

RNN-T 和 CTC 等端到端模型会引入一个额外的空白词元 ⟨b⟩ 用于对齐。因此，优化意味着要对所有对齐进行边缘化，这需要借助前向后向算法 [87], [88] 高效计算训练准则及其梯度。对于 CTC、RNN-T 和 RNA 模型，以及与这些模型各自定义的对齐差异相对应的经典（全求和）混合人工神经网络/隐马尔可夫模型，只需对该算法稍作修改即可。相比之下，注意力编码器-解码器模型没有用于显式对齐的潜变量。感兴趣的读者可参阅相关论文，了解训练算法的更多细节
[89], [90], [12], [13], [42], [33], [14], [15], [46]。

在训练过程中，所有用于自动语音识别的端到端模型以及经典隐马尔可夫模型都提供了解决底层序列对齐问题的机制：既可以像 CTC、RNN-T、RNA 以及混合人工神经网络/隐马尔可夫模型那样，通过相应的潜变量显式解决；也可以像注意力编码器-解码器模型那样，隐式解决。此外，还需要考虑语音与静音的区别；这可以通过将静音作为潜在标签引入来显式处理（混合人工神经网络/隐马尔可夫模型），也可以像目前几乎所有端到端模型的标准做法那样，完全不对静音进行标注，从而隐式处理。

端到端模型也可以利用分层训练计划。这些计划可能包括多个彼此分开的训练轮次，以及显式生成的初始对齐；在若干轮维特比式 [91], [92], [93] 训练后，先将其固定，再重新启用端到端式全求和训练，对所有可能的对齐进行边缘化。Zeyer 等人 [45], 采用了这种替代方法：先使用一个初始的全求和 RNN-T 模型生成对齐，然后继续进行逐帧交叉熵训练。通过将式（4）中对所有可能对齐的求和替换为与所生成对齐序列相对应的单项，该方法大大简化了训练过程。近期， [94] 也提出了类似流程，同样仅采用端到端模型。在这项工作中，使用 CTC 初始化训练并生成初始对齐，随后进行中间阶段的维特比式 RNN-T 训练，最后进行全求和微调。与仅采用全求和训练的方法相比，该流程提高了收敛性。

值得注意的是，与 RNN-T 和 RNA 标签拓扑不同，CTC 不要求每个标签位置都具有单个标签发射的对齐。然而，训练 CTC 模型最终确实会使每个假设标签对应单个标签发射。关于 CTC 训练这一特性的分析（通常称为尖峰行为）可见 [95] 及其中引用的文献。 [96] 甚至提出了一种不含非空白循环转移的 CTC 变体。

### *B.* 使用外部语言模型进行训练

端到端自动语音识别模型通常在序列级进行归一化。因此，采用最大互信息准则 [97]进行序列训练，与标准交叉熵／条件似然训练相同。然而，一旦在训练阶段纳入外部语言模型，就需要显式地进行序列归一化，从而形成最大互信息序列判别训练。训练阶段已利用这种方法，进一步将端到端模型与基于纯文本数据训练的外部语言模型相结合 [98], [99], [100]。

### *C.* 最小错误训练

由于语音识别的目标是最小化词错误率（WER），越来越多的研究开始将其纳入目标函数。这些方法称为序列判别学习，在传统自动语音识别[101], [102], [103], [104], [105],中取得了显著改进，随后也在端到端模型中得到了探索。通常，这些损失通过以下方式构造：

以“束搜索”模式而非教师强制模式运行，并根据束中的候选假设所产生的错误构造损失。因此，这种训练首先需要训练模型以优化 $P(C|X)$，从而学习一组适用于束搜索的良好权重。不过，也有研究提出了直接方法，避免这种分离，从头开始进行判别式训练[32], [106]。

探讨惩罚词错误的论文包括最小词错误率（MWER）训练 [107], ，其损失函数经过构造，使得预期词错误数达到最小。后续研究还包括用于 RNN-T 和自注意力-T 的 MWER [108], ，以及使用前缀搜索而非 n-best 的 MWER [109]。此外，也有研究从强化学习的角度考察 MWER [110], [111]。最后，最优补全蒸馏（OCD） [112] 旨在通过高效的动态规划算法最小化总编辑距离。另有一类序列训练研究会在训练时引入单独的外部语言模型， [113], 这同样可以通过近似格重组[99] 以及无格方法 [100]高效完成。

### *D.* 预训练

所有端到端模型以及用于自动语音识别的经典隐马尔可夫模型都提供了整体式模型，原则上能够支持从头开始训练。然而，为了通过应用预训练和模型增长 [114], [115]使训练过程达到最佳性能并实现高效收敛，存在许多用于初始化和引导训练过程的策略。监督式逐层预训练已成功应用于经典 [4]，[116], 以及基于注意力的自动语音识别模型 [117], ，并且可以与中间层子采样方案[118], 和模型增长 [119]结合使用。利用未转写音频、大规模半监督数据和/或多语种数据的预训练方法 [120], [121], [122], [123], [124], [125], [126], [127], [128], [129], [130], [131]值得进行一篇独立的综述；同样，这些方法也适用于混合式深度神经网络/隐马尔可夫模型和端到端方法——本文不再对此展开讨论。

### *E.* 训练计划与课程

人们已经制定了专门的训练计划，以指导优化过程，并在此过程中显式或隐式地实现适当的对齐行为 [118], [45], [94]。

学习率控制存在许多方法 [132]，[133]：NewBob [134], [135] 及其改进方法 [133]，、全局与逐参数学习率控制（指数衰减、幂衰减等） [136], 、学习率预热 [31]，、热重启/余弦退火 [137], 、权重衰减与逐步减小批量大小 [138], 、微调 [139] 或基于群体的训练 [140]。关于元学习的综述，另请参阅 [141]。

序列学习方法也会考虑课程学习 [142], [143], 例如，先考虑较短的序列[144], [145],暂时增加下采样 [118] 初期采用更多下采样，或者在多说话人自动语音识别训练中进行排序按信噪比对混合语音进行排序，并从能量均衡、性别混合的说话人开始 [146]。

### *F.* 优化与正则化

优化通常基于随机梯度下降[147], 和动量 [148], [149]，，以及一系列相应的自适应方法，其中最突出的是 Adam [150]及其变体 [150], [116], [151]。

增加训练轮次似乎能够带来改进 [45]，，表 8]，，而且据报道，对多个训练轮次进行平均也有助于提升效果 [75]。关于双重下降效应及其与训练数据量、标签噪声和早停之间关系的讨论，另请参阅 [152]。

正则化对训练性能有显著贡献，包括 L2 和权重衰减 [153], [137],权重噪声 [154],自适应均值 L1/L2 [155], 梯度噪声 [156], Dropout[157], [158], [159], 层丢弃法 [160], [161], [162], dropcon- nect [163],Zoneout [164], 注意力分数平滑 [14],标签平滑 [165], 计划采样 [166], 辅助损失 [165], [167], 可变时间反向传播 [168]，[169], Mixup [170], 提高帧率 [151],或批归一化 [171]。

### *G.* 数据增强

端到端自动语音识别模型的训练也受益于数据增强方法，这些方法也可被视为正则化方法。然而，其多样性及其对性能的影响，使得有必要单独进行综述。

大多数数据增强方法通过利用语音信号变化的某些维度来进行数据扰动：语速扰动 [172], [173],声道长度扰动[174], [172], 频率轴失真 [172],序列噪声注入 [175],SpecAugment [176], 或语义掩码 [177]。此外，还可利用纯文本数据，通过特征级 [178] 或信号级 [179]的文本转语音（TTS）生成数据。在一项针对基于 TTS 的数据增强对不同端到端自动语音识别架构影响的比较中， [179], 注意力编码器-解码器似乎是唯一能够显著受益于 TTS 生成数据的架构。

在近期的一项研究 [145] 及其后续工作[151]，中，本文列举的许多正则化和数据增强方法被联合采用，使单头注意力编码器-解码器模型在 Switchboard 任务上达到了最先进水平。

## V. 端到端模型的解码

本节主要介绍几种端到端语音识别的解码算法。端到端自动语音识别的基本解码算法旨在从所有可能的序列中估计出最可能的序列 $C$，如下所示：

$$\hat{C} = \arg\max_{C \in \mathcal{U}^*} P(C|X)$$

其中，$^*$ 表示克莱尼闭包，用于表示由词元词汇表 $\mathcal{U}$中的词元组成的所有可能序列的集合。 $P(C|X)$是由端到端自动语音识别系统得到的概率分布函数。下一节将介绍如何获得ˆ识别结果 $C$。

### A. 贪心搜索

贪心搜索算法主要用于CTC，它忽略输出标签之间的依赖关系，如下所示：

$$\hat{A} = \prod_{t=1}^{T} \left( \arg\max_{a_t} P(a_t|X) \right)$$

其中， $a_t$ 是第 II-A1 节中引入的对齐标记。通过将ˆ对齐标记序列 $A$ 转换为相应的标记序ˆ列 $C$，即可获得原始字符序列。argmax 操作可以在输入帧 $t$上并行执行，从而实现快速解码 [12], [180], ，但由于缺乏输出依赖关系，与基于注意力和 RNN-T 的方法相比，其性能通常相对较差。

由于 Transformer[31], [71], [75] 及其变体 [73], [76]在帧维度上的全部计算均已并行化 [161], [181], 因此 CTC 的快速解码得到了进一步提升。例如，包括 Im- puter [182], 、Mask-CTC [181], 、基于插入的建模 [183],、连续积分-触发（CIF） [184] 以及其他变体[185], [186] 在内的非自回归模型，已作为 CTC 的替代非自回归模型得到积极研究。 [186] 表明，通过使用 Intel(R) Xeon(R) Silver 4114 CPU、2.20GHz，CTC 贪心搜索及其变体可实现 0.06 的实时因子（RTF：（解码时间）/（输入长度））。该论文还表明，与采用束搜索的注意力/RNN-T 方法相比，非自回归模型的性能下降并不显著（使用自条件 CTC [185] 时为 19.7%，而使用注意力/RNN-T 时为 18.5/18.9%）。

贪心搜索算法也用作基于注意力、RNA、CTC 和 RNN-T 的方法的近似解码，如下所示：

$$\hat{c}_i = \arg\max_{c_i} P(c_i|\hat{C}_{1:i-1}, X) \text{ for } i = 1, \ldots, N$$

$$\hat{a}_t = \arg\max_{a_t} P(a_t|\hat{A}_{1:t-1}, X) \text{ for } t = 1, \ldots, T$$

与下文所述的束搜索算法相比，贪心搜索算法不会考虑序列中的备选假设。然而，已知贪心搜索算法造成的性能下降并不严重 [15],[33], ，尤其是在模型于匹配条件下得到充分训练时[3]。

### B. 束搜索

束搜索算法旨在通过近似方式˜在解码过程中从所有可能的假设 $\mathcal{U}^*$ 中考虑一个子集 $\mathcal{C}$ ，即 $\mathcal{C} \subset \mathcal{U}^*$ 。预测的ˆ ˜输出序列 $C$ 是从假设子集 $\mathcal{C}$中选出的，而非从所有可能的假设 $\mathcal{U}^*$中选出，即，

$$\hat{C} = \arg\max_{C \in \tilde{\mathcal{C}}} P(C|X) \tag{6}$$

束搜索算法旨在找到一组可能的假˜设 $\mathcal{C}$，从而高效地纳入有希望的假设同时避免遍历所有可能假设 $\mathcal{U}^*$时遇到的组合爆炸。

束搜索主要分为两类：1) 帧同步 束搜索和 2) 标签同步 束搜索。它们的主要区别在于，是否对每个输入帧 $t$ 或每个输出标记 $i$执行假设剪枝。以下各节将介绍这两种算法。

### C. 标签同步束搜索

假设我们有一组截至第 $(i-1)$˜个标记 $\mathcal{C}_{1:i-1}$的部分假设。所有截至第 $i$˜个标记 $\mathcal{C}_{1:i}$ 的可能部分假设集合可由 $\mathcal{C}_{1:i-1}$ 按如下方式扩展：

$$\mathcal{C}_{1:i} = \{(\tilde{\mathcal{C}}_{1:i-1}, c_i = c)\}_{c \in \mathcal{U}} \tag{7}$$

˜假设的数量 $|\mathcal{C}_{1:i}|$ 最多为 $|\mathcal{C}_{1:i-1}| \times |\mathcal{U}|$ 。束搜索算法会从 $\mathcal{C}_{1:i}$ 中剪去概率分数较低的假设，只保留其中一定数量（束宽 $\Delta$）的假设 $i$ ，具体如下：

$$\tilde{\mathcal{C}}_{1:i} = \text{NBEST}_{C_{1:i} \in \mathcal{C}_{1:i}} P(C_{1:i}|X), \text{ where } |\tilde{\mathcal{C}}_{1:i}| = \Delta \tag{8}$$

请注意，NBEST(·) 是一种操作，用于根据概率分数 $P(C_{1:i}|X)$ 提取排名靠前的 $\Delta$ 个假设。该概率分数由端到端神经网络计算得出，也可以是第 VI-B 节所述的多个分数的融合。

在标签同步束搜索中，输出序列的长度（$N$）是未知的。因此，在此剪枝过程中，我们还将到达话语末尾的假设（即预测句末符号⟨eos⟩）作为有希望的假设加入到式（6）中的假设集合 $\mathcal{C}$ 。

由于标签同步束搜索不显式依赖对齐信息，因此，与帧同步束搜索不同，长度相同的序列假设可能覆盖完全不同数量的编码器帧，正如 [58]所指出的那样。结果是，我们观察到，非常短和非常长的片段假设的分数往往会落入相同的范围，束搜索也会错误地选择此类假设。 [190] 展示了这类极端情况的一个示例，分别导致短片段假设和长片段假设出现大量删除错误和插入错误。因此，标签同步束搜索需要借助启发式方法限制输出序列长度，以避免输出序列过长或过短。通常，最小和最大长度阈值根据输入帧长度 $|X|$ 按比例确定，并通过可调参数 $\rho_{\min}$ 和 $\rho_{\max}$ 设置为 $L_{\min} = \lfloor \rho_{\min}|X| \rfloor$， $L_{\max} = \lfloor \rho_{\max}|X| \rfloor$。尽管这些是控制假设长度相当直观的方法，但输出长度的最小值和最大值取决于每种语言所使用的词元单位或文字系统类型。另一种启发式方法是提供与输出长度或注意力权重相关的附加分数，包括长度惩罚和覆盖项[191], [187]。端点检测 [192]也用于自动估计假设长度。 [188] 重新定义了注意力解码器的隐式长度模型，以纳入

[3] 另一方面，在注意力编码器-解码器模型中，扩大搜索空间并不能持续提升语音识别性能 [187], [188], ； 神经机器翻译中也观察到了这一现象 [189]。

束搜索，从而在增大束大小时保持一致的行为且不会出现性能下降。

CTC 的标签同步束搜索算法通过对每个标签假设 [12] 的所有可能对齐进行边缘化来实现。

### *D.* 帧同步束搜索

与公式（8）中的标签同步情况不同，帧同步束搜索算法会在每个输入帧 $t$ 执行剪枝，如下所示：

$$\tilde{\mathcal{C}}_{1:i(t)} = \mathrm{NBEST}_{C_{1:i(t)}} P(C_{1:i(t)}|X), \text{ where } |\tilde{\mathcal{C}}_{1:i(t)}| = \Delta$$

其中，$C_{1:i(t)}$ 是由对齐 $A_{1:t}$ 得到的长度为 $i$（$t$）的标签序列，该对齐在第 II-A 节中介绍。$P$（$C_{1:i(t)}|X$）是对所有可能对齐 $A_{1:t} \in \mathcal{A}_{(X,C_{1:i(t)})}$ 求和得到的。对齐集合 $\mathcal{A}$ 取决于显式对齐方法，包括 CTC、RNN-T 和 RNA。$\mathcal{C}_{1:i(t)}$ 是扩展到输入帧 $t$ 为止的部分假设，与公式（7）类似。

与标签同步算法相比，帧同步算法需要在束搜索算法中处理额外的输出标记转换。得益于输入帧和输出标记的显式对齐信息，帧同步算法可以轻松扩展到在线和/或流式解码。

针对隐马尔可夫模型的束搜索经典方法，以及 CTC 和 RNN-T 的变体，都基于加权有限状态转换器（WFST）[41], [193], [194] 或词法前缀树[195], [196], [79]。这些方法被归类为帧同步束搜索。它们通常与 N-gram 语言模型或全上下文神经语言模型[197], [198]结合使用。RNN-T [13], [199] 和 CTC 前缀搜索 [200]可以通过在标签转移状态中纳入语言模型分数来处理神经语言模型。触发式注意力方法 [201], [202] 使我们能够在帧同步束搜索中将注意力编码器-解码器模型与 CTC 和神经语言模型结合使用，并利用注意力编码器-解码器模型和语言模型分数，对 CTC前缀搜索给出的假设进行在线重评分。

### *E.* 分块解码

另一种束搜索实现方式为输入特征采用固定长度的块单元。在这种分块处理中，我们可以使用基于 BLSTM、输出延迟单向 LSTM 或 Transformer（及其变体）的非因果编码器网络，从而利用块内的未来上下文。这些未来上下文信息可以避免完全因果网络的性能下降。在这种设置中，块大小成为控制延迟和准确率之间权衡的因素。这项技术同时应用于 RNN-T [203], [204], [205] 和注意力编码器-解码器模型 [39], [206], [207]。分块处理对注意力编码器-解码器模型尤为重要，因为它能够在输入特征与输出标签之间提供分块单调对齐约束，并实现分块流式解码。

### *F.* 解码期间的模型融合

与经典的基于 HMM 的束搜索类似，我们将从不同模块获得的各种分数结合起来，包括主要的端到端自动语音识别分数和语言模型分数。

*1*）同步分数融合：当多个模块的分数同步时，会执行最简单的分数融合。在这种情况下，我们可以在每个帧 $t$ 或标签 $i$处直接将多个分数相加。最广为人知的分数组合是语言模型浅层融合。

语言模型浅融合：如第 VI 节所述，各种神经语言模型都可以与端到端自动语音识别集成。最简单的集成方式是采用语言模型浅融合[208][209][210]„ 如第 VI-B1 节所述，在式（8）的束搜索过程中，将语言模型 $P_{\mathsf{lm}}(C_{1:i})$ 的分数与端到端自动语音识别分数 $P(C_{1:i}|X)$ 按对数线性方式相加，具体如下：

$$\log P(C_{1:i}|X) \rightarrow \log P(C_{1:i}|X) + \gamma \log P_{\mathsf{lm}}(C_{1:i})$$

其中，$\gamma$ 是语言模型权重。当然，我们也可以结合其他分数，例如长度惩罚和覆盖度项，正如第 V-C 节所述。

*2*）异步分数融合：如果我们结合依赖于帧的分数函数 $P(a_t|\cdot)$（例如 CTC、RNN-T）和依赖于标签的分数函数 $P(c_i|\cdot)$（例如注意力、语言模型），就必须处理帧时间索引与标签时间索引 $i$ 和 $t$之间的不匹配。

在时间同步束搜索中，这种融合是通过将语言模型分数纳入标签转移状态来实现的 [211], [212], [213]。 [200] 还通过纳入由词分隔符（空格）符号触发的语言模型分数，将基于词的语言模型与基于词元的CTC模型相结合。

在标签同步束搜索中，我们首先根据假设的标签序列，对所有可能的对齐进行边缘化，从而由依赖帧的分数函数计算出依赖标签的分数。CTC/注意力联合解码 [190]是一个典型示例，其中CTC分数是通过CTC前向算法 [180]对所有可能的对齐进行边缘化计算得出的。这种方法消除了标签同步束搜索中的错误对齐问题以及难以找到正确句末的问题 [190], [214]。

请注意，束搜索中的模型融合方法可以实现简单的单遍解码，但它要求模型的时间单位相同，或者需要通过额外的动态规划来调整不同的时间单位，尤其是在标签同步束搜索中。当话语长度增加时，这种动态规划计算量会显著增大，因此需要采用一些启发式方法来降低计算成本 [215]。

### *G.* 分数融合中的词汇约束

传统上，我们使用基于词的语言模型来捕捉词元单位的上下文信息，同时也会考虑自动语音识别中的基于词的词汇约束。然而，端到端自动语音识别通常使用字母或词元单位，这会在束搜索过程中造成进一步的单位不匹配。如前文所述，将词汇约束从词元单位整合到词元单位的经典方法基于加权有限状态转换器。这种方法首先构建一个由词元（T）、词汇表（L）和基于词的语言转换器（G）组成的 TLG转换器 [194]。该 TLG转换器同时用于 CTC [194] 和基于注意力的方法 [46]。

时间同步束搜索中采用的另一种方法，是在词分隔符（空格）符号触发时插入基于词的语言模型分数 [216]。为了使基于词的语言模型与基于字符的端到端自动语音识别同步， [217] 使用前缀树表示，将基于词和基于字符的语言模型结合起来；而 [192], [218]则使用前瞻词概率来预测后续字符，而非使用基于字符的语言模型。前缀树表示也用于子词词元单元的情况 [219], [220]。

### *H.* 多遍融合

此前的融合方法是在束搜索过程中执行的，因此能够采用单遍算法。常见的替代方法基于多遍算法，不必考虑同步问题，而是在整个话语范围内考虑完整上下文，进行 n-best 或格评分。 [15] 采用 N-best重评分技术来整合基于词的语言模型。 [47]在多遍解码框架中结合前向和后向搜索，以整合双向 LSTM 解码器网络。近期，研究人员考察了切换不同端到端自动语音识别系统的两遍算法，包括 RNN-T $\rightarrow$ 注意力[221] CTC $\rightarrow$ 注意力 [222], [223]。这旨在通过第一遍提供流式输出，并在第二遍利用注意力重新评分来优化先前的输出，从而满足实时接口要求并提升识别性能。

除了上述讨论中的 N-best 输出之外，为了更好地进行多遍解码，人们也强烈希望生成格输出，因为格中包含更丰富的假设信息。格输出还可用于口语术语检测、口语语言理解和词后验概率。然而，由于缺少马尔可夫假设，RNN-T 和注意力编码器-解码器无法合并假设，也无法生成格，这与基于隐马尔可夫模型或 CTC 的系统不同。为解决这一问题，一些研究通过将输出依赖关系限制在固定长度内（即有限历史） [224], [34], 来修改这些模型；另一些研究则保留原始 RNN-T 结构，但在束搜索过程中合并相似的假设 [80]。

### *I.* 在假设和话语两个层面进行向量化

我们可以在束搜索过程中对多个假设进行向量化，从而加速解码过程：用针对向量化假设的向量-矩阵运算，替代对每个假设进行的分数累积步骤。这种方法已在 RNN-T [212], [213], [225] 和基于注意力的模型[215]中得到研究。通过这一改进，我们能够利用多核CPU和GPU的并行计算能力，大幅提升速度，同时还可以在一个批次中同时处理多个话语。

主要的深度神经网络和端到端自动语音识别工具包都支持这种向量化。例如，TensorFlow[4][226], 和 FAIRESEQ[5][227] 为通用的序列到序列任务提供了向量化束搜索接口，并且它可以用于基于注意力的端到端自动语音识别。包括 ESPnet[6][215], 、ESPRESSO[7][218], 和 RETURNN[8][228] 在内的端到端自动语音识别工具包也支持向量化束搜索算法。

## VI. 语言模型集成

长期以来，语言模型（LM）一直与声学模型结合用于自动语音识别，以提高识别准确率 [2]。语言模型的主要作用是帮助自动语音识别系统生成语言上最有可能的句子假设。实际应用中，语言模型通常设计为概率模型，例如 $N$ 元模型 [2] 或神经网络语言模型 [8], [229], ，并使用大量示例句子进行训练。语言模型提供一个概率分布$P(C)$，定义在句子这一随机变量 $C$上，用于为句子假设评分，并促使系统选择更可能的假设。

如第七节所述，混合式自动语音识别系统通常在推理阶段使用独立的语言模型，而端到端系统基本上使用单一的自动语音识别模型，其中包含一个承担语言模型作用的网络组件，且整个网络经过联合训练。例如，RNN-T 的预测网络和注意力编码器-解码器模型的解码器网络都承担语言模型的作用。因此，端到端自动语音识别似乎不再需要独立的语言模型。然而，许多研究表明，独立的语言模型有助于提高端到端自动语音识别的识别准确率。

端到端自动语音识别仍需要外部语言模型，可能有三个原因：

*a)* 弥补泛化能力不足：端到端模型需要学习比经典的基于模块的模型（如声学模型）更复杂的映射函数。因此，如果训练数据量不足，端到端模型往往会面临过拟合问题。预训练语言模型有望弥补端到端模型预测泛化能力不足的问题。

*b)* 外部文本数据的使用：端到端模型需要使用配对语音和文本数据进行训练，而语言模型只需使用文本数据即可训练。通常，文本数据比配对数据更容易收集。对于相同数量的句子，语言模型的训练速度也比端到端模型更快。因此，利用外部文本数据可以更有效地改进语言模型，从而进一步提升自动语音识别系统的性能。

*c)* 领域自适应：将端到端模型应用于特定领域时，领域自适应有助于提高识别准确率。然而，端到端模型的领域自适应需要目标领域的一定数量的配对数据。此外，在假设存在多个领域的情况下，为系统支持的各个领域维护多个端到端模型可能成本高昂。如果有目标领域的预训练语言模型，则无需更新端到端模型，也能更轻松地提高领域特定词汇和说话风格的识别准确率。

[4] https://www.tensorflow.org/api docs/python/tf/contrib/seq2seq/BeamSearchDecoder [5] https://github.com/pytorch/fairseq/blob/master/fairseq/sequence generator.py

[6]https://github.com/espnet/espnet [7]https://github.com/freewym/espresso [8] https://github.com/rwth-i6/returnn

本节回顾了用于端到端自动语音识别的各种语言模型，以及将其整合到主端到端网络中的融合技术。

### A. 语言模型

语言模型提供先验概率分布 $P(C)$。如果句子 $C$ 可以分解为由字符、子词和单词等构成的词元序列，则可以根据链式法则计算其概率分布：

$$P(C) = \prod_{i=1}^{L+1} P(c_i|c_{0:i-1})$$

其中，$c_i$ 表示 $C$的第 $i$个词元，$c_{0:i-1}$ 表示词元序列 $c_0$，$c_1$，...，$c_{i-1}$，并假设 $c_0 = \langle \text{sos} \rangle$ 和 $c_{L+1} = \langle \text{eos} \rangle$。

大多数语言模型旨在提供条件概率 $P(c_i|c_{0:i-1})$，也就是说，它们被建模为根据前面的词元序列预测下一个词元。这里简要回顾这类语言模型。

*1）N元语言模型*： $N$元语言模型长期以来一直用于自动语音识别 [2]。早期的端到端系统 [194], [46], [187] 也采用了 $N$元语言模型。 $N$元模型依赖马尔可夫假设，即下一个词元的概率分布仅取决于前面的 $N-1$ 个词元，即$P(c_i|c_{0:i-1}) \approx P(c_i|c_{i-N+1:i-1})$，其中 $N$ 对于基于词的模型通常为 3 到 5，对于基于子词和基于字符的模型则取更大的值。$N$元概率的极大似然估计根据训练数据集中$N$ 个连续词元的计数确定，如下所示

$$P(c_i|c_{i-N+1:i-1}) = \frac{\mathcal{K}(c_{i-N+1}, \ldots, c_i)}{\sum_{c_i} \mathcal{K}(c_{i-N+1}, \ldots, c_i)}$$

其中，$\mathcal{K}(\cdot)$表示每个词元序列的计数。由于数据规模有限，因此必须采用平滑技术，避免根据稀有词元序列的零计数或极小计数来估计概率。这些技术会根据计数 [230]的大小，使用低阶模型（例如（$N-1$）元模型）对 $N$元概率进行补偿。

$N$元模型的优点是简单，尽管其表现不如最先进的神经语言模型。在训练过程中，主要步骤只是统计数据集中的 $N$ 元组，而且只需执行一次。在解码时，可以通过表查找快速获取语言模型概率，也可以预先将其附加到解码图中，例如加权有限状态转换器。

*2) 前馈神经网络语言模型*：前馈神经网络（FNN）语言模型由 [8], 提出，它使用神经网络估计 $N$元概率。该网络接收 $N-1$ 个词元，并预测下一个词元，如下所示：

$$\begin{aligned} P(c_i|c_{i-N+1:i-1}) &= \text{softmax}(W_o h_i + b_o) \\ h_i &= \tanh(W_h e_i + b_h) \\ e_i &= \text{concat}(E(c_{i-N+1}), \ldots, E(c_{i-1})) \end{aligned}$$

其中，$W_o$ 和 $W_h$ 是权重矩阵，$b_o$ 和 $b_h$ 是偏置向量。 $E(y)$ 提供 $c$的嵌入向量，并且 concat(·)操作将给定的向量 [9]连接起来。该模型首先将每个输入词元映射到嵌入空间，然后获得隐藏向量 $h_i$ 作为上下文向量，用以表示前面的 $N-1$ 个词元。最后，它通过 Softmax 层输出下一个词元的概率分布。尽管这种语言模型仍然依赖马尔可夫假设，但其性能优于上一节介绍的经典 $N$-gram语言模型。前馈神经网络语言模型的优越性能主要源于对每个词元及上下文的分布式表示。该语言模型学会表示词元/上下文向量，使语义相似的词元/上下文在嵌入空间和上下文空间中彼此靠近。由于这种表示的平滑效果优于 $N$-gram语言模型所采用的基于计数的表示，因此前馈神经网络语言模型在预测下一个词元时比$N$-gram语言模型具有更好的泛化能力。基于神经网络的语言模型基本上都利用了这种表示方式。

*3) RNN语言模型*：引入了循环神经网络（RNN）语言模型，利用循环连接来捕捉超过 $N-1$ 个先前词元的更长距离上下文信息 [229]。与前馈神经网络语言模型不同，隐藏向量的计算方式如下：

$$\begin{aligned} h_i &= \text{recurrence}(e_i, h_{i-1}) \\ e_i &= E(c_{i-1}) \end{aligned}$$

其中，recurrence($e_i$，$h_{i-1}$) 表示一个递归函数，该函数接收先前的隐藏向量 $h_{i-1}$ 和输入 $e_i$，并输出下一个隐藏向量 $h_i$。对于简单的（Elman 型）RNN，该函数可计算如下：

$$\text{recurrence}(e, h) = \tanh(W_h e + W_r h + b_h)$$

其中，$W_r$ 是用于循环连接的权重矩阵，作用于先前的隐藏向量 $h$。这种循环结构使隐藏向量能够保留上下文信息，而不必将上下文限制在 $N-1$ 个词元之内。然而，随着词元通过该递归过程不断处理，过去的信息会呈指数级衰减。因此，目前循环部分更广泛地采用堆叠的长短期记忆网络层，它们具有独立的内部记忆单元和门控机制，以保留长距离上下文信息[231]。借助这一机制，RNN语言模型在许多任务中的表现优于其他基于 $N$-gram 的模型。

*4）卷积语言模型*：卷积语言模型 [232]使用门控时序卷积替代了循环神经网络语言模型中使用的循环连接。隐藏向量计算如下：

$$\begin{aligned} h_i &= h'_i \otimes \sigma(g_i) \\ h'_i &= e_{i-k+1:i} * W + b \\ g_i &= e_{i-k+1:i} * V + c \end{aligned}$$

其中，$\otimes$ 表示逐元素乘法，$*$ 表示时域卷积操作，$k$ 表示补丁大小。 $\sigma(g_i)$表示卷积激活 $h'_i$的门控函数，其中 $\sigma(\cdot)$是Sigmoid函数。 $W$ 和 $V$ 是卷积矩阵，$b$ 和 $c$ 是偏置向量。卷积块和门控块通常会通过残差连接多次堆叠。在 [233], 中，采用了一个包含14个块的卷积语言模型（ConvLM）用于端到端自动语音识别。卷积语言模型的优势在于其计算

[9] 为简化起见，我们省略了从嵌入层到Softmax层的可选直接连接 [8]

效率高，即使上下文长度固定且仅包含字母和词片等短词元，其性能也与循环神经网络语言模型相近。具体而言，其高度可并行化的架构有助于利用大型训练数据集训练模型。

*5*）*Transformer*语言模型：Transformer架构 [31] 已应用于语言模型 [234]，并用于自动语音识别 [235], [75],，其中语言模型被设计为无需来自编码器等其他模块输入的Transformer解码器。隐藏向量的计算方式如下：

$$h_i = \mathrm{FFN}(h'_i) + h'_i$$
$$h'_i = \mathrm{MHA}(e_i, e_{1:i}, e_{1:i}) + e_i$$

其中，FFN(·) 和 MHA(·，·，·) 分别表示前馈网络和多头注意力模块。多头注意力和前馈模块通常会多次堆叠，例如堆叠 6 次 [75], ，以获得最终的隐藏向量。Transformer语言模型的优势在于，它们能够通过自注意力机制考虑所有先前的词元，而无需像循环神经网络语言模型那样将其概括为固定大小的记忆。因此，可以充分利用较长的先前上下文来预测下一个词元，从而获得比循环神经网络语言模型更好的性能。然而，计算复杂度会随着序列长度呈二次方增长。因此，上下文长度通常会限制为固定大小，或限制在单个句子范围内。为克服这一限制，Transformer-XL [236] 会重用已计算的激活值，其中包含更早先词元的信息，并使用截断时间反向传播（时间反向传播）算法训练模型。压缩Transformer [237] 将这一方法进一步扩展，通过引入压缩步骤，在固定大小的记忆网络中保留较早但重要的信息，从而利用更长的上下文信息。

### *B.* 融合方法

将语言模型融入端到端自动语音识别有多种方法。研究人员探索了各种语言模型融合方法，以在不同情况下提高自动语音识别的准确率。

*1)* 浅层融合：浅层融合是在推理阶段结合预训练端到端模型和语言模型的最常用方法。正如我们在第 V-F 节所述，浅层融合只需通过对数线性组合，将端到端模型分数与语言模型分数结合起来：

$$Score(C|X) = \log P(C|X) + \gamma \log P(C) \tag{9}$$

其中，$\gamma$ 是语言模型的缩放因子 [208][209][210]。这种方法的优点是，当源域与目标领域之间没有显著差异时，它简单且有效。

*2*）深层融合：深层融合 [238] 是一种利用联合网络将语言模型与端到端模型相结合的方法。所有网络参数都经过联合训练，使模型能够更好地协作，从而提高识别准确率；其中，联合网络通过门控机制结合端到端模型状态和语言模型状态，并根据当前状态控制语言模型的贡献。

*3*）冷融合：冷融合 [239] 是另一种结合语言模型的方法，与深层融合类似，但端到端模型是在冻结语言模型参数的同时，借助预训练语言模型进行训练的。由于端到端模型在整个训练过程中都能利用语言模型，因此它学会了利用语言模型减少特定于语言的信息，并仅提取将源序列映射到目标序列所需的相关信息。这种机制降低了语言模型在端到端模型中的作用，并缓解了训练数据的语言偏差。因此，端到端模型能够更好地应对训练数据与目标领域之间的域不匹配。与深层融合不同，冷融合可以将端到端模型与针对目标领域的预训练语言模型相结合，从而提高识别准确率。组件融合 [240] 将冷融合扩展为：在端到端模型中使用预训练语言模型和训练数据的转录文本，更侧重于减少训练数据带来的偏差。

*4)* 内部语言模型估计：还有一种通过浅层融合减少训练数据中语言偏差的方法。当源域（训练数据）与目标领域（测试数据）之间存在较大的领域不匹配时，语言偏差就会成为问题，因为端到端模型分数强烈依赖于源域中的语言先验。为了从分数中消除这种偏差，我们可以显式估计代表语言先验的语言模型，称为内部语言模型，并从式 (9) 的自动语音识别分数中减去该语言模型分数，即：

$$Score(C|X) = \log P_{\varphi}(C|X) - \gamma_{\varphi} \log P_{\varphi}(C) + \gamma_{\tau} \log P_{\tau}(C)$$

其中，下标 $\varphi$ 和 $\tau$ 分别表示源领域和目标领域。$\gamma_{\varphi}$ 和 $\gamma_{\tau}$ 是它们的缩放因子。减去内部语言模型分数，相当于近似声学概率密度 $P_{\varphi}(X|C)$，因为在 $X$ 固定时，$P_{\varphi}(X|C) \propto P_{\varphi}(C|X)/P_{\varphi}(C)$ 成立，此时自动语音识别分数可视为经典混合式自动语音识别系统。因此，减去的端到端模型分数起到了声学模型的作用，使其在语言方面更不受领域影响，并与外部语言模型 $P_{\tau}(C)$ 结合后实现更高的识别准确率。

密度比方法 [241] 使用训练数据的转录文本训练内部语言模型。混合自回归转导器（HAT） [34] 扩展了 RNN-T，使模型在消除编码器输出（即将其设为零）时成为内部语言模型。该方法利用预测网络作为内部语言模型，从而简化了框架，避免了训练额外的语言模型以及在推理阶段使用该模型。在 [242],的工作中，研究人员提出了一种类似于 HAT 的方法，分别在标准 RNN-T 和基于注意力的编码器-解码器模型之上构建内部语言模型。在 [98], 中，研究人员针对注意力编码器-解码器模型提出了几种估计内部语言模型的技术：将估计得到的偏置向量而非零向量输入语言模型。偏置向量可以通过对编码器状态或上下文向量取平均来估计，也可以仅根据解码器标签上下文，通过一个用于预测上下文向量的小型长短期记忆网络来估计。这些用于估计内部语言模型的技术也在 [243]中针对 RNN-T 进行了评估。

### C. 大规模预训练语言模型的使用

近年来，使用大规模文本数据训练的语言模型已可用于各种自然语言处理任务。BERT [244]和 GPT-2[245] 是基于Transformer语言模型的代表性模型。这类语言模型也以不同方式应用于端到端自动语音识别系统，例如 N-best重打分 [246]和对话上下文嵌入 [247]。

## VII. 与经典自动语音识别的关系

端到端系统具有若干特性，使其有别于传统自动语音识别系统。理想情况下，端到端系统定义的自动语音识别模型会将所有知识源整合到一个全局联合模型中，不使用次级知识源，并避免像经典方法那样将声学模型和语言模型分开。这些全局联合模型完全从头开始训练，采用基于单一类型（转录）训练数据的全局训练准则，因此所需的自动语音识别领域知识较少，前提是有充足的训练数据。类似地，识别过程可以通过单次处理完成，且不需要使用显式词典。

端到端系统与传统自动语音识别系统共享部分特性，而出于性能方面的考虑，且取决于可用训练数据的数量，端到端系统的其他特性则会有所取舍。下文将讨论端到端方法与经典自动语音识别方法之间可以观察到的各种关系。

首先，端到端模型为自动语音识别引入了替代性的对齐建模方法。注意力机制提供了一种全新的方法，可将声学观测序列映射为标签序列；而转导器方法[12], [13], [33], [211] 则以一种可解释为类似于具有特定模型拓扑结构的隐马尔可夫模型的方式处理对齐问题，其中包括对对齐的边缘化 [90], [248], [249]。在解码过程中，CTC模型也可以以类似隐马尔可夫模型的方式使用 [194]。此外，转导器方法等价于分段模型/直接隐马尔可夫模型 [58]。

E2E 与经典统计自动语音识别系统共有的最显著特性，是采用单遍解码策略，在作出最终决策 [93]之前整合所有涉及的知识源（模型、组件）。这包括在 E2E 系统 [180], [44], [187], [250], [145], [219], [251][252], 中使用完整标签上下文依赖，以及在经典系统中通过全上下文语言模型实现相同功能 [253], [197], [198], [254]。在经典自动语音识别系统中，甚至可能在搜索过程中保留隐马尔可夫模型对齐路径求和 [255]。E2E 和经典自动语音识别系统在解码时都采用束搜索。然而，与经典搜索方法相比，E2E 束搜索通常高度简化，束大小非常小，约为 1 到 100 [14][15], [187], [118]。非常小的束大小也在一定程度上掩盖了基于注意力的编码器-解码器模型所表现出的长度偏差 [256], [257], ，从而在模型错误与搜索错误之间进行权衡 [258]。文献[188]概述了在自动语音识别中除使用小束大小之外处理长度偏差的方法。

E2E 系统的另一个显著特征是直接进行字符级建模，避免采用基于音素的建模和发音词典 [18], [216], [194], [15], [259], [187], [260], [261], [262], [112], [263]，其中一些系统甚至朝着整词建模发展 [259], [25]。然而，字符级建模同样适用于经典的混合隐马尔可夫模型架构 [264] ，并且已有研究表明，即使在不使用神经网络的标准隐马尔可夫模型中也能发挥作用 [265]。

许多经典自动语音识别搜索范式都基于多遍方法，通过依次应用日益复杂的声学模型和/或语言模型来生成搜索空间表示 [266], [267], [196]。然而，使用端到端模型时也会采用多遍策略，但这会削弱端到端概念。解码器模型组合采用两遍方法，同时仍保持与 [60]中相同的延迟约束。此外，多遍方法还包括端到端自适应方法 [268], [269], [270], [271]。

经典自动语音识别系统的架构将声学模型和语言模型分开。与此相反，端到端模型避免了这种分离，并定义了一个联合模型。虽然这使得模型可以通过单一目标进行训练，但也将（隐式）先验的训练限制在音频训练数据的转录文本上。尽管如此，为了利用更多纯文本训练数据，通常会将独立语言模型与端到端模型结合使用。然而，由于端到端模型具有隐式先验，即内部语言模型，因此与独立语言模型结合并非易事，需要相应的内部语言模型估计和补偿方法，例如 [241], [34], [242], [98][272]。至少从识别准确率的角度来看，尚不清楚经典自动语音识别架构中声学建模与语言建模的明确分离究竟是劣势（因为采用独立训练目标），还是优势（因为可以轻松使用纯文本训练数据）。此外，人们发现，语言模型的训练目标，即语言模型困惑度，与词错误率之间具有很强的相关性 [273], [274], [275], [276]。此外，语言建模的判别式方法 [277], [278]也可视为迈向联合建模的一步。

尽管采用逐帧交叉熵进行自动语音识别的标准混合人工神经网络/隐马尔可夫模型训练已经具有判别性，但尚未达到序列判别性，需要先验对齐，而且训练时也未考虑（外部）语言模型。不过，采用序列判别式训练准则[97]和无格训练方法 [32]，或许可以弥补这些潜在不足。

与严格的端到端系统不同，经典自动语音识别架构会使用主要训练数据之外的次级知识源，即用于声学模型训练的（转录）语音音频，以及用于语言模型训练的文本数据。其中最突出的是使用发音词典并定义音素集。在低资源场景下，发音词典等次级资源可能会有所帮助。然而，生成这些资源往往成本高昂，甚至可能引入错误，例如词典中的发音与实际观察到的发音不符。因此，当训练资源足够丰富时，次级知识源可能会变得过时 [279]，甚至

有害，因为它们可能会引入错误信息 [280],[281]。

经典自动语音识别模型通常采用逐步训练的方式：将先前训练的模型中获得的知识注入后续训练阶段，例如以隐马尔可夫模型状态对齐的形式。然而，经典自动语音识别中的此类方法也可以视为特定的训练计划。使用基于高斯混合模型的声学模型所得到的隐马尔可夫模型对齐来初始化深度学习模型，也可作此理解，其中高斯混合模型充当初始浅层模型。在经典自动语音识别中，也有人提出从头训练深度神经网络、同时避免对高斯模型进行中间训练的方法 [282], [283], [284],，并将其与字符级建模结合使用 [264]。推动传统系统实现更集成化训练的另一步骤，是采用判别式训练准则，避免使用表示竞争词序列的中间表示（通常基于格） [285], [32], [286], [287], [106]。

传统自动语音识别系统的训练通常会采用辅助目标来解决语音聚类等子任务。分类回归树（CART）方法用于聚类三音素隐马尔可夫模型状态 [24], [288]。较新的方法提出在神经网络建模框架内进行聚类，同时仍保留辅助聚类目标 [289], [283]。不过，端到端方法也会使用辅助目标，最显著的是用于子词生成，例如通过字节对编码实现 [27]。此外，还可以间接利用现有发音词典来辅助端到端系统的子词生成 [290], [291],，而相关研究表明，这种方法的表现优于字节对编码。在传统自动语音识别系统中，也可以通过直接对上下文中的音素进行建模来完全避免语音聚类 [292]。

值得注意的是，具体而言，基于注意力的编码器-解码器模型需要更多的训练轮次才能达到较高的性能，例如，比较在 Switchboard 300小时上训练的系统时，请参见[293]中的表5。此外，研究表明，基于注意力的编码器-解码器模型会受到训练资源不足的影响 [294], [295],，而这一问题可以通过多种方法加以改善，包括正则化技术 [145] 以及使用 SpecAugment [296] 和文本转语音（TTS） [21]进行数据增强。

此外，SpecAugment 也被证明能够改进经典的混合隐马尔可夫模型 [297]。另一方面，文本转语音显著改进了在有限资源上训练的基于注意力的编码器-解码器模型，但其性能仍不及其他端到端方法或混合隐马尔可夫模型；而文本转语音对后两者的改进并不显著 [179]。多语言方法也有助于改进低资源任务的自动语音识别开发，无论是经典方法 [298], ，还是端到端系统 [299], [300]。

语音识别研究一直受到国际评测活动（例如 NIST）及相应基准任务的推动。经典方法与端到端方法之间的竞争，在广泛使用的 Librispeech [301] 和 Switchboard [302] 任务中得到了很好的体现，这些任务表明端到端模型正日益受到重视。在 Librispeech 上，目前已发表的最佳经典混合系统的词错误率约为 2.3%（test-clean）和 4.9%（test-other） [303], [294],

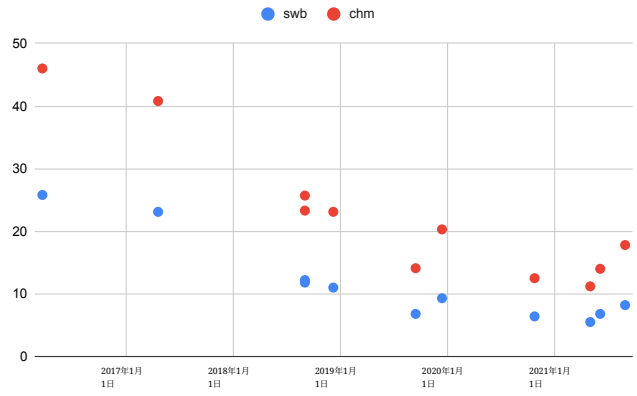


图9. Switchboard 任务中的端到端自动语音识别性能提升。

与此同时，已有若干端到端系统取得了相近的性能 [296], [176], [304], [177], ，其中一些端到端系统明显超越了此前的最先进结果，词错误率分别低至 1.8%（test-clean）和 3.7%（test-other） [305]； [73], [70]也报告了类似结果。融合基于经典隐马尔可夫模型的方法与单调 RNN-T 的见解，在有限的训练预算下也取得了同样出色的结果 [94]。最后，在 Switchboard 300小时数据集上训练时，目前的最佳结果由端到端系统取得 [151] ，在 HUB5'00 Switchboard测试集上的词错误率为 5.4%，而最佳混合系统的结果为 6.6% [306] 。

## VIII. 常见基准测试中端到端方法的总体性能趋势

本节总结了各种技术在常见自动语音识别基准上的表现：图10展示了基于 Librispeech [301] 的结果，图9展示了基于 Switchboard（SWBD） [302] 的结果，以考察端到端自动语音识别技术的发展轨迹。我们选择这两个数据库，是因为它们在语音和机器学习领域社区中广泛使用，并涵盖了自发语音（SWBD）和朗读语音（Librispeech）两种说话风格。图9和图10显示，基于端到端模型的早期尝试 [261], [118] 带来了显著的性能提升，所有任务的错误率都降至原来的一半以下 [10]。

尽管整体趋势表明，随着时间推移，自动语音识别性能稳步提升，但仍有几项显著进展。2019年年中，两个基准测试中都观察到一项重要进展，这得益于数据增强方法 SpecAugment [176], [177], ，如第 IV-G 节所述。之后的进展主要来自对新型端到端架构的探索，包括在 SpecAugment 基础上采用 Transformer [75], [307], 、Conformer [73], [76], 和 ContextNet[70] ，如第 III-C 节所述。为提升自动语音识别性能，语言建模领域也进行了类似探索 [235], [75]。2021年 Librispeech 基准测试中的最后一项进展，则基于自监督学习 [22], [308] 和半监督

[10] 想了解这些基准测试最新更新的读者，也可以查看 https://github.com/syhw/wer arewe 和 https://github.com/thu-spmi/ASR-Benchmarks/blob/main/README.md。

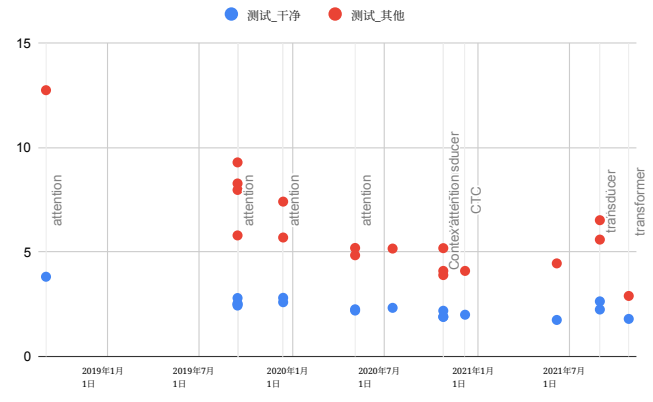


图10。Librispeech任务中的端到端自动语音识别性能提升。

学习 [304], [309]。这些技术利用了大量无标签域内语音数据（例如，Libri-Light 60K 小时 [310]）。

## IX. 端到端模型的部署

本文讨论的许多思路都已由各行业研究实验室进行探索 [311], [312], [313]，[314], [315] ，其中包括谷歌等。在本节中，我们以谷歌设备端生产级系统的开发作为部署的典型案例进行回顾。首个部署到生产环境的流式端到端模型于2019年随 Pixel 4智能手机推出 [212], [316]。该模型采用流式RNN-T第一遍系统，并在第二遍中使用注意力编码器-解码器系统对第一遍的假设进行重新评分。此外，模型还采用了基于FST的上下文偏置 [65] ，这对于针对多样化查询获得准确结果至关重要。该模型在中央处理器上运行，速度远快于实时处理。

2020年，针对Pixel 5智能手机 [317]，，系统得到进一步改进，以降低用户感知延迟（即用户开始说话到词出现在设备上的时间间隔）。改进内容包括端到端端点检测 [82]等方面的进展，以促使麦克风更快关闭；以及FastEmit [64] ，以促使模型更早发出词元。

最后，在2021年，该模型针对 Pixel 6 智能手机进行了进一步改进 [318]，，以利用设备上的张量处理单元（TPU） [319] 。改进包括：在编码器中使用 Conformer层 [73]；在解码器中使用小型嵌入预测网络 [77]；使用两遍级联编码器运行第二遍束搜索 [62]；以及使用神经语言模型重评分器，帮助提高长尾命名实体的准确率。该模型是谷歌迄今发布的最佳自动语音识别系统，在质量和延迟方面均表现出色，并且显著优于云端的80GB模型。

## X. 未来工作的方向

目前，端到端模型主导着自动语音识别领域的学术讨论。但这至少在一定程度上尚未（？ ）体现在端到端自动语音识别架构的相应商业部署中。端到端模型尚不能完美适用于所有自动语音识别条件，要充分发挥端到端建模的优势，还需要进一步研究。

当训练数据充足时，端到端模型似乎表现出色，但在低资源条件下却难以扩展。同样，领域变化需要灵活地切换语言模型；对于基于声学模型和语言模型分离的经典自动语音识别模型而言，这种做法很自然。关于在端到端模型中使用外部语言模型以及进行内部语言模型估计的持续研究已初见成效，预计未来还会取得进一步进展。

顶尖端到端自动语音识别系统通常需要比性能相当的传统自动语音识别系统多几个数量级的训练轮次，因此还需要进一步研究高效、稳健的优化方法和训练计划。

端到端模型高度集成，也意味着模块化程度有所降低，而模块化可能有助于提升模型的可解释性和可复用性。此外，更高效的训练计划也可能受益于模块化。端到端模型的一个公认优势是，所有内容都从数据中学习，从而无需使用次级知识源（例如发音词典和音素集）。然而，罕见事件仍然带来挑战，例如自动语音识别中的稀有词，这方面仍需进一步研究。

由于声学模型和语言模型不再分离，随之而来的问题是：如何在端到端模型训练中利用纯文本资源——除了使用文本转语音生成训练数据之外，我们还能想到其他解决方案吗？值得注意的是，近期已有不少研究探索了结合语音与文本模态的方法，尝试将二者隐式或显式地映射到共享空间中 [320], [321], [322], [323], [324], [325]，[130], [326]。此外，针对自动语音识别等判别式问题以及文本转语音等生成式问题，都已有高性能的端到端解决方案；如何将二者结合利用，以转录语音音频为基础，并借助纯文本和／或纯音频数据进行半监督训练 [20]，[327]？

更具体地说，对于 AED架构，我们观察到长度偏差。尽管已有许多用于解决 AED 中长度偏差的启发式方法，但我们仍缺乏对此现象有充分依据的解释，也缺少对原始模型相应的补救方法。

其他尚待解决的研究问题包括说话人自适应，以及对录音条件的鲁棒性，尤其是在失配情况下。

端到端原则也为解决多通道自动语音识别提供了一个很有前景的思路，即通过端到端方案联合处理声源分离、说话人分离和语音识别问题 [328], [329]。

最后，我们需要探究端到端是否是一个合适的指导原则，以及不同的端到端自动语音识别模型彼此之间、以及它们与传统自动语音识别方法之间存在怎样的关系。自动语音识别研究与开发最重要的指导原则一直是性能，而广泛使用的基准任务和国际评测活动也极大地推动了自动语音识别的发展。面对当前传统模型和端到端模型的多样性，我们还需要解决这样一个问题：如今自动语音识别领域的最先进水平由什么构成？未来我们是否可以期待一种通用的最先进自动语音识别架构？

## XI. 结论

在本文中，我们详细概述了端到端自动语音识别方法。此类模型近年来越来越受欢迎，其主张采用高度集成的神经网络组件，通过基于字符的输出单元，直接将输入语音转换为输出文本序列。因此，这类模型摒弃了由声学模型、发音模型和语言模型组成的传统模块化自动语音识别架构，转而采用单一的紧凑结构，并依靠数据实现有效学习。这些设计选择使高准确率的设备端语音识别模型得以部署（见第 IX 节），但也带来了一些缺点，这些问题仍是当前研究的重点（见第 X 节）。最后，我们指出，Jinyu Li 的优秀综述文章 [330],与本文相互补充，为端到端自动语音识别提供了另一种视角。

## 致谢

作者谨向 Julian Dierkes、Yifan Peng、Zolt´an T¨uske、Albert Zeyer 和 Wei Zhou 致谢，感谢他们帮助完善本文。

## 参考文献

[1] T. Bayes, 《论概率学说中的一个问题的求解》, 《伦敦皇家学会哲学汇刊》, 第53卷, 第370–418页, 1763年。[2] F. Jelinek, 《语音识别统计方法》。马萨诸塞州剑桥：麻省理工学院出版社，1997年。[3] H. A. Bourlard 和 N. Morgan, 《联结主义语音识别：一种混合方法》。马萨诸塞州诺韦尔：克吕维尔学术出版社，1993年。[4] F. Seide、G. Li 和 D. Yu, 《使用上下文相关深度神经网络进行会话语音转录》, 载于*Interspeech*会议论文集，意大利佛罗伦萨，2011年8月，第437–440页。[5] V. Fontaine、C. Ris 和 H. Leich, 《用于改进语音识别的非线性判别分析》, 载于*Eurospeech*会议论文集，希腊罗得岛，1997年9月，第1–4页。[6] H. Hermansky、D. Ellis 和 S. Sharma, 《用于传统隐马尔可夫模型系统的Tandem联结主义特征提取》, 载于*IEEE ICASSP*会议论文集，第3卷，土耳其伊斯坦布尔，2000年6月，第1635–1638页。[7] M. Nakamura 和 K. Shikano, 《基于神经网络的英语词类别预测研究》, 载于*IEEE ICASSP*会议论文集，英国格拉斯哥，1989年5月，第731–734页。[8] Y. Bengio、R. Ducharme 和 P. Vincent, 《神经概率语言模型》, 载于*NIPS*会议论文集，第13卷，科罗拉多州丹佛，2000年11月，第932–938页。[9] H. Schwenk 和 J.-L. Gauvain, 《面向大词汇量连续语音识别的联结主义语言建模》, 载于*IEEE ICASSP*会议论文集，佛罗里达州奥兰多，2002年5月，第765–768页。[10] Z. T¨uske、P. Golik、R. Schl¨uter 和 H. Ney, 《使用原始时域信号和深度神经网络进行大词汇量连续语音识别的声学建模》, 载于*Interspeech*会议论文集，新加坡，2014年9月，第890–894页。[11] 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》, 载于*IEEE ASRU*会议论文集，亚利桑那州斯科茨代尔，2015年12月，第30–36页。[12] A. Graves、S. Fern´andez、F. Gomez 和 J. Schmidhuber, 《联结主义时序分类：使用循环神经网络标注未分段序列数据》, 载于*ICML*会议论文集，宾夕法尼亚州匹兹堡，2006年6月，第369–376页。[13] A. Graves, 《使用循环神经网络进行序列转导》, 2012年11月，arXiv:1211.3711。 [在线]。网址：https://arxiv.org/abs/ 1211.3711[14] J. K. Chorowski、D. Bahdanau、D. Serdyuk、K. Cho 和 Y. Bengio, 《用于语音识别的基于注意力的模型》, 载于*NIPS*会议论文集，第28卷，加拿大魁北克省拉瓦尔，2015年12月，第577–585页。

[15] W. Chan, N. Jaitly, Q. Le, 和 O. Vinyals, 《Listen, Attend and Spell：用于大词汇量会话语音识别的神经网络》, 载于 IEEE ICASSP 论文集，中国上海，2016年3月，第4960–4964页。[16] P. Liang, A. Bouchard-Cˆot´e, D. Klein, 和 B. Taskar, 《一种端到端的机器翻译判别式方法》, 载于 ACL 论文集，澳大利亚悉尼，2006年7月，第761–768页。[17] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, 和 P. Kuksa, 《（几乎）从零开始的自然语言处理》, 《机器学习研究期刊》, 第12卷，第2493–2537页，2011年。[18] A. Graves 和 N. Jaitly, 《使用循环神经网络实现端到端语音识别》, 载于 ICML 论文集，中国北京，2014年6月，第1764–1772页。[19] 《剑桥词典》, https://dictionary.cambridge.org/dictionary/english/end-to-end，访问日期：2020年2月21日。[20] A. Tjandra, S. Sakti, 和 S. Nakamura, 《边说边听：深度学习语音链》, 载于 IEEE ASRU 论文集。日本冲绳：IEEE，2017年12月，第301–308页。[21] M. K. Baskar, S. Watanabe, R. Astudillo, T. Hori, L. Burget, 和 J. ˇCernock´y, 《使用非配对语音和文本的半监督序列到序列自动语音识别》, 载于 Interspeech 论文集，奥地利格拉茨，2019年9月，第3790–3794页。[22] A. Baevski, Y. Zhou, A. Mohamed, 和 M. Auli, 《wav2vec 2.0：语音表示自监督学习框架》, 载于 NeurIPS 论文集，加拿大不列颠哥伦比亚省温哥华，2020年12月，第12 449–12 460页。[23] X. Chang, T. Maekaku, Y. Fujita, 和 S. Watanabe, 《语音识别、语音增强与自监督学习表示的端到端集成》, 2022年4月，arXiv:2204.00540。 [在线]。可访问：https://arxiv.org/abs/2204.00540[24] L. Breiman, J. Friedman, C. Stone, 和 R. Olshen, 《分类与回归树》。加利福尼亚州贝尔蒙特：Taylor & Francis，1984年。[25] H. Soltau, H. Liao, 和 H. Sak, 《神经语音识别器：用于大词汇量语音识别的声学到词的长短期记忆模型》, 2016年10月，arXiv:1610.09975。 [在线]。可访问：https://arxiv.org/abs/1610.09975[26] G. K. Zipf, 《人类行为与最省力原则》。马萨诸塞州波士顿：Addison-Wesley Press，1949年。[27] R. Sennrich, B. Haddow, 和 A. Birch, 《使用子词单元进行稀有词的神经机器翻译》, 载于 ACL 论文集，德国柏林，2015年8月，第1715–1725页。[28] R. Pang, T. N. Sainath, R. Prabhavalkar, S. Gupta, Y. Wu, S. Zhang, 和 C.-c. Chiu, 《端到端模型的压缩》, 载于 Interspeech 论文集，印度海得拉巴，2018年9月，第27–31页。[29] M. Schuster 和 K. Nakajima, 《日语和韩语语音搜索》, 载于 IEEE ICASSP 论文集，日本京都，2012 年 3 月，第 5149–5152 页。[30] S. Hochreiter 和 J. Schmidhuber, 《长短期记忆》, 《神经计算》, 第9卷，第8期，第1735–1780页，1997年。[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, 和 I. Polosukhin, 《注意力就是一切》, 载于 NIPS 论文集，加利福尼亚州洛杉矶，2017年12月，第5998–6008页。[32] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, 和 S. Khudanpur, 《基于无格最大互信息的纯序列训练神经网络，用于自动语音识别》, 载于 Interspeech 论文集。加利福尼亚州旧金山：国际语音通信协会，2016年9月，第2751–2755页。 [在线]。可访问：https://doi.org/10 21437 2016 595 。/Interspeech。-[33] H. Sak, M. Shannon, K. Rao, 和 F. Beaufays, 《循环神经对齐器：用于序列到序列映射的编码器-解码器神经网络模型》, 载于 Interspeech 论文集，第8卷，瑞典Stockhol，2017年8月，第1298–1302页。[34] E. Variani, D. Rybach, C. Allauzen, 和 M. Riley, 《混合自回归转导器（HAT）》, 载于 IEEE ICASSP 论文集，西班牙巴塞罗那，2020年5月，第6139–6143页。[35] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, 和 J. Dean, 《谷歌神经机器翻译系统：弥合人工与机器翻译之间的差距》, 2016年10月，arXiv:1609.08144。 [在线]。可访问：https://arxiv.org/abs/ 1609.08144[36] D. Bahdanau, K. Cho, 和 Y. Bengio, 《通过联合学习对齐与翻译实现神经机器翻译》, 2014年5月，arXiv:1409.0473。[在线]。可访问：https://arxiv.org/abs/1409.0473

[37] N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, 和 S. Bengio, 《使用部分条件化的在线序列到序列模型》，载于 NIPS 论文集，西班牙巴塞罗那，2016年12月，第5067–5075页。[38] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, 和 D. Eck, 《通过强制单调对齐实现在线线性时间注意力》，载于 ICML 论文集，澳大利亚悉尼，2017 年 8 月，第 2837–2846 页。[39] C.-C. Chiu 和 C. Raffel, 《单调分块注意力》，载于 ICLR 论文集，2018年4月，arXiv:1712.05382。[40] N. Arivazhagan, C. Cherry, W. Macherey, C.-C. Chiu, S. Yavuz, R. Pang, W. Li, 和 C. Raffel, 《用于同步机器翻译的单调无限回溯注意力》，载于 ACL 论文集，意大利佛罗伦萨，2019年6月，第1313–1323页。[41] M. Mohri, F. Pereira, 和 M. Riley, 《语音识别中的加权有限状态换能器》，《Computer Speech & Language》，第16卷，第1期，第69–88页，2002年。[42] A. Graves, A.-r. Mohamed, 和 G. Hinton, 《使用深度循环神经网络进行语音识别》，载于 IEEE ICASSP 论文集，加拿大不列颠哥伦比亚省温哥华，2013年5月，第6645–6649页。[43] N. Moritz, T. Hori, S. Watanabe, 和 J. Le Roux, 《基于图监督的序列转导》，载于 IEEE ICASSP 论文集，新加坡，2022年5月，第7212–7216页。[44] A. Tripathi, H. Lu, H. Sak, 和 H. Soltau, 《单调循环神经网络转导器与解码策略》，载于 IEEE ASRU 论文集，新加坡圣淘沙，2019年12月，第944–948页。[45] A. Zeyer, A. Merboldt, R. Schl¨uter, 和 H. Ney, 《改进神经转导器的新训练流程》，载于 Interspeech 论文集，中国上海，2020年10月，第2812–2816页。[46] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, 和 Y. Bengio, 《端到端基于注意力的大词汇量语音识别》，载于 IEEE ICASSP 论文集，中国上海，2016年3月，第4945–4949页。[47] M. Mimura, S. Sakai, 和 T. Kawahara, 《前向后向注意力解码器》，载于 Interspeech 论文集，印度海得拉巴，2018年9月，第2232–2236页。[48] A. Graves, 《使用循环神经网络生成序列》，2017年8月。 2013 [ ] 1308 0850 [49] 在线。可访问：https://arxiv.org/abs/。R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, 和 N. Jaitly, 《用于语音识别的序列到序列模型比较》，载于 Interspeech 论文集，瑞典Stockhol，2017年8月，第939–943页。[50] C.-C. Chiu, W. Han, Y. Zhang, R. Pang, S. Kishchenko, P. Nguyen, A. Narayanan, H. Liao, S. Zhang, A. Kannan, R. Prabhavalkar, Z. Chen, T. Sainath, 和 Y. Wu, 《长篇语音识别端到端模型比较》，载于 IEEE ASRU 论文集，新加坡圣淘沙，2019年12月，第889–896页。[51] T. N. Sainath, C.-C. Chiu, R. Prabhavalkar, A. Kannan, Y. Wu, P. Nguyen, 和 Z. Chen, 《提升在线神经转导器模型的性能》，载于 IEEE ICASSP 论文集，加拿大艾伯塔省卡尔加里，2018年4月，第5864–5868页。[52] N. Moritz, T. Hori, 和 J. Le Roux, 《用于端到端语音识别的触发式注意力》，载于 IEEE ICASSP 论文集，英格兰布莱顿，2019年5月，第5666–5670页。[53] A. Merboldt, A. Zeyer, R. Schl¨uter, 和 H. Ney, 《局部单调注意力变体分析》，载于 Interspeech 论文集，奥地利格拉茨，2019年9月，第1398–1402页。[54] A. Zeyer, R. Schl¨uter, 和 H. Ney, 《潜在单调注意力变体研究》，2021年3月，arXiv:2103.16710。 [在线]。可访问：http://arxiv.org/abs/ 2103 16710 。A. Zeyer, R. Schmitt, W. Zhou, R. Schl¨uter, 和 H. Ney, 《用于自动语音识别的单调分段注意力》，卡塔尔多哈，2023年1月，即将发表；arXiv:2210.14742。 [在线]。可访问：https://arxiv.org/abs/ 2210 14742 。Z. Tian, J. Yi, Y. Bai, J. Tao, S. Zhang, 和 Z. Wen, 《用于端到端语音识别的同步Transformer》，载于 IEEE ICASSP 论文集，西班牙巴塞罗那，2020年5月，arXiv:1912.02958。 [在线]。可访问：https://arxiv.org/abs/ 1912 02958 。E. Beck, M. Hannemann, P. Doetsch, R. Schl¨uter, 和 H. Ney, 《用于大词汇量自动语音识别的分段编码器-解码器模型》，载于 Interspeech 论文集，印度海得拉巴，2018年9月。[58] W. Zhou, A. Zeyer, A. Merboldt, R. Schl¨uter, 和 H. Ney, 《分段建模与神经转导器建模的等价性：概念验证》，载于 Interspeech 论文集，捷克布尔诺，2021年8月，第2891–2895页。[59] S. Kim, T. Hori, 和 S. Watanabe, 《基于联合CTC-注意力与多任务学习的端到端语音识别》，载于 IEEE ICASSP 论文集，路易斯安那州新奥尔良，2017年3月，第4835–4839页。[60] T. N. Sainath、R. Pang、D. Rybach、Y. He、R. Prabhavalkar、W. Li、M. Visontai、Q. Liang、T. Strohman、Y. Wu、I. McGraw 和 C. Chung- Cheng, 《两遍端到端语音识别》，载于 Interspeech 论文集，英国布莱顿，2019年5月，第2773–2777页。[61] K. Hu、T. N. Sainath、R. Pang 和 R. Prabhavalkar, 《基于审议模型的两遍端到端语音识别》，载于 IEEE ICASSP 论文集。西班牙巴塞罗那：IEEE，2020年5月，第7799–7803页。[62] A. Narayanan、T. N. Sainath、R. Pang、J. Yu、C.-C. Chiu、R. Prab- havalkar、E. Variani 和 T. Strohman, 《用于统一流式和非流式自动语音识别的级联编码器》，载于 IEEE ICASSP 论文集，加拿大安大略省多伦多，2021年6月，第5629–5633页。[63] A. Tripathi、J. Kim、Q. Zhang、H. Lu 和 H. Sak, 《Transformer 转导器：统一流式与非流式语音识别的单一模型》，2020年10月，arXiv:2010.03192。 [在线]。可访问：https://arxiv.org/abs/ 2010 03192 。J. Yu、W. Han、A. Gulati、C.-C. Chiu、B. Li、T. N. Sainath、Y. Wu 和 R. Pang, 《通用自动语音识别：通过全上下文建模统一并改进流式自动语音识别》，2020年10月，arXiv:2010.06030。 [在线]。可访问：https://arxiv.org/abs/ 2010 06030 。D. Zhao、T. N. Sainath、D. Rybach、P. Rondon、D. Bhatia、B. Li 和 R. Pang, 《浅融合端到端上下文偏置》，载于 Interspeech 论文集，奥地利格拉茨，2019年9月，第1418–1422页。[66] G. Pundak、T. N. Sainath、R. Prabhavalkar、A. Kannan 和 D. Zhao, 《深度上下文：端到端上下文语音识别》，载于 IEEE SLT 论文集，希腊雅典，2018年12月，第418–425页。[67] S. Kim 和 F. Metze, 《对话上下文感知端到端语音识别》，载于 IEEE SLT 论文集，希腊雅典，2018年12月，第434–440页。[68] A. Bruguier、R. Prabhavalkar、G. Pundak 和 T. N. Sainath, 《Phoebe：端到端语音识别的发音感知上下文化》，载于 IEEE ICASSP 论文集，英国布莱顿，2019年5月，第6171–6175页。[69] M. Delcroix、S. Watanabe、A. Ogawa、S. Karita 和 T. Nakatani, 《基于辅助特征的端到端自动语音识别系统自适应》，载于 Interspeech 论文集，印度海得拉巴，2018年9月，第2444–2448页。[70] W. Han、Z. Zhang、Y. Zhang、J. Yu、C.-C. Chiu、J. Qin、A. Gulati、R. Pang 和 Y. Wu, 《ContextNet：利用全局上下文改进自动语音识别的卷积神经网络》，载于 Interspeech 论文集，中国上海，2020年10月，第3610–3614页。[71] L. Dong、S. Xu 和 B. Xu, 《Speech-Transformer：用于语音识别的无循环序列到序列模型》，载于 IEEE ICASSP 论文集，加拿大艾伯塔省卡尔加里，2018年4月，第5884–5888页。[72] Q. Zhang、H. Lu、H. Sak、A. Tripathi、E. McDermott、S. Koo 和 S. Kumar, 《Transformer 转导器：采用 Transformer 编码器和 RNN-T 损失的可流式语音识别模型》，载于 IEEE ICASSP 论文集，西班牙巴塞罗那，2020年5月，第7829–7833页。[73] A. Gulati、J. Qin、C.-C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 和 R. Pang, 《Conformer：卷积增强的语音识别 Transformer》，载于 Interspeech 论文集，中国上海，2020年10月，第5036–5040页。[74] Y. Peng、S. Dalmia、I. Lane 和 S. Watanabe, 《Branchformer：用于捕获语音识别与理解中局部和全局上下文的并行多层感知机-注意力架构》，载于 ICML 论文集。马里兰州巴尔的摩：PMLR，2022年7月，第17 627–17 643页。[75] S. Karita、N. Chen、T. Hayashi、T. Hori、H. Inaguma、Z. Jiang、M. Someki、N. E. Y. Soplin、R. Yamamoto、X. Wang、S. Watanabe、T. Yoshimura 和 W. Zhang, 《语音应用中 Transformer 与 RNN 的比较研究》，载于 IEEE ASRU 论文集，新加坡圣淘沙，2019年12月，第449–456页。[76] P. Guo、F. Boyer、X. Chang、T. Hayashi、Y. Higuchi、H. Inaguma、N. Kamo、C. Li、D. Garcia-Romero、J. Shi、J. Shi、S. Watanabe、K. Wei、W. Zhang 和 Y. Zhang, 《Conformer 推动的 ESPNET 工具包最新进展》，载于 IEEE ICASSP 论文集。加拿大安大略省多伦多：IEEE，2021年6月，第5874–5878页。[77] R. Botros、T. Sainath、R. David、E. Guzman、W. Li 和 Y. He, 《绑定与精简的 RNN-T 解码器》，载于 Interspeech 论文集，捷克布尔诺，2021年9月，第4563–4567页。[78] M. Ghodsi、X. Liu、J. Apfel、R. Cabrera 和 E. Weinstein, 《采用无状态预测网络的 RNN 转导器》，载于 IEEE ICASSP 论文集，西班牙巴塞罗那，2020年5月，第7049–7053页。[79] W. Zhou、S. Berger、R. Schl¨uter 和 H. Ney, 《用于大词汇量语音识别的基于音素的神经转导器》，载于 IEEE ICASSP 论文集，加拿大安大略省多伦多，2021年6月，第5644–5648页。[80] R. Prabhavalkar、Y. He、D. Rybach、S. Campbell、A. Narayanan、T. Strohman 和 T. N. Sainath, 《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 IEEE ICASSP 论文集 ，多伦多 ，安大略省，加拿大，2021年6月，第5659–5663页。

[81] S.-Y. Chang、B. Li 和 G. Simko，“采用多任务和多领域训练的统一端点检测器”，载于 IEEE ASRU 会议论文集，圣淘沙，新加坡，2019年12月，第100–106页。[82] B. Li、S.-y. Chang、T. N. Sainath、R. Pang、Y. He、T. Strohman 和 Y. Wu，“迈向快速、准确的流式端到端自动语音识别”，载于 IEEE ICASSP 会议论文集，巴塞罗那，西班牙，2020年5月，第6069–6073页。[83] T. Yoshimura、T. Hayashi、K. Takeda 和 S. Watanabe，“集成基于CTC的语音活动检测的端到端自动语音识别”，载于 IEEE ICASSP 会议论文集，巴塞罗那，西班牙，2020年5月，第6999–7003页。[84] Y. Fujita、T. Wang、S. Watanabe 和 M. Omachi，“采用非自回归插入式模型实现流式自动语音识别”，载于 Interspeech 会议论文集，布尔诺，捷克，2021年9月，第3740–3744页。[85] Y. Bengio，“深层架构基于梯度的训练实用建议”，2012年6月，arXiv:1206.5533。 [在线]。可访问：https://arxiv.org/abs/ 1206 5533 。J. Schmidhuber，“神经网络中的[86] 深度学习：综述”，Neural Networks，第61卷，第85–117页，2015年1月，arXiv:1404.7828。 [ ] 1404 7828 在线。[87] 可访问：https://arxiv.org/abs/。L. Baum，“马尔可夫过程概率函数统计估计中的一个不等式及相关最大化技术”，Inequalities，第3卷，第1–8页，1972年。[88] L. Rabiner 和 B.-H. Juang，“隐马尔可夫模型简介”，IEEE声学、语音与信号处理汇刊，第3卷，第1期，第4–16页，1986年。[89] Y. Bengio、R. De Mori、G. Flammia 和 R. Kompe，“用于语音识别或密度估计的神经网络-高斯混合模型混合架构”，载于 NIPS 会议论文集，第4卷，科罗拉多州，1991年12月，第175–182页。[90] P. Haffner，“采用全局MMI算法的联结主义语音识别”，载于 Eurospeech 会议论文集，柏林，德国，1993年12月，第1929–1932页。[91] R. E. Bellman，《动态规划》。新泽西州普林斯顿：普林斯顿大学出版社，1957年。[92] A. Viterbi，“卷积码的错误界及渐近最优解码算法”，IEEE信息论汇刊，第13卷，第260–269页，1967年。[93] H. Ney，“单阶段动态规划算法在连续词识别中的应用”，IEEE声学、语音与信号处理汇刊，第32卷，第2期，第263–271页，1984年。[94] W. Zhou、W. Michel、R. Schl¨uter 和 H. Ney，“用于语音识别的神经转导器高效训练”，2022年4月，arXiv:2204.10586。 [在线]。可访问：https://arxiv.org/abs/ 2204.10586[95] A. Zeyer、R. Schl¨uter 和 H. Ney，“ 为何 CTC 会产生尖峰行为？ ”，预印本 arXiv:2105.14849，2021 年5月，arXiv:2105.14849。 [ ] 2105 14849 在线。可访问：http://arxiv.org/abs/。A. Laptev、S. Majumdar 和 B. Ginsburg，“通过新的WFST拓扑结构探索CTC变体”，载于 Interspeech 会议论文集，仁川，韩国，sep 2022 [ ] 10 21437 。[96] [97] 在线。可访问：https://doi.org/。X. He、L. Deng 和 W. Chou，“序列模式识别中的判别式学习——面向优化的语音识别统一综述”，IEEE信号处理杂志，第25卷，第5期，第14–36页，2008年。[98] M. Zeineldeen、A. Glushko、W. Michel、A. Zeyer、R. Schl¨uter 和 H. Ney，“探究改进基于注意力的编码器-解码器自动语音识别模型语言模型集成的方法”，载于 Interspeech 会议论文集，布尔诺，捷克，2021年8月，第2856–2860页。[99] N.-P. Wynands、W. Michel、J. Rosendahl、R. Schl¨uter 和 H. Ney，“采用近似重组的注意力模型高效序列训练”，载于 IEEE ICASSP 会议论文集，新加坡，2022年5月，arXiv:2110.09245。[100] Z. Yang、W. Zhou、R. Schl¨uter 和 H. Ney，“基于音素的神经转导器的无格序列判别式训练”，载于 IEEE ICASSP 会议论文集，罗得岛，希腊，2022年10月，已提交，预印本：arXiv:NNNN.NNNNN。[101] V. Valtchev、J. J. Odell、P. C. Woodland 和 S. J. Young，“大词汇量识别系统的MMIE训练”，Speech Communication，第22卷，第4期，第303–314页，1997年。[102] D. Povey 和 P. Woodland，“大词汇量连续语音识别的改进判别式训练技术”，载于 IEEE ICASSP 会议论文集，盐湖城，犹他州，2001年5月，第45–48页。[103] R. Schl¨uter、W. Macherey、B. M¨uller 和 H. Ney，“语音识别判别式训练准则和优化方法的比较”，Speech Communication，第34卷，第3期，第287–310页，2001年5月，获EURASIP最佳论文奖。

[104] B. Kingsbury，《用于神经网络声学建模的序列分类准则的基于格的优化》，载于 IEEE ICASSP 会议论文集，台北，台湾，2009年4月，第3761–3764页。[105] G. Heigold、R. Schl¨uter、H. Ney 和 S. Wiesler，《自动语音识别的判别式训练：建模、准则、优化、实现与性能》，《IEEE信号处理杂志》，第29卷，第6期，第58–69页，2012年11月。[106] W. Michel、R. Schl¨uter 和 H. Ney，《大词汇量连续语音识别中无格与基于格的序列判别式训练准则比较》，载于 Interspeech 会议论文集，格拉茨，奥地利，2019年9月，第1601–1605页。 [ ] 1907 01409 在线。获取地址：[107] http://arxiv.org/abs/。R. Prabhavalkar、T. N. Sainath、Y. Wu、P. Nguyen、Z. Chen、C.-C. Chiu 和 A. Kannan，《注意力序列到序列模型的最小词错误率训练》，载于 IEEE ICASSP 会议论文集，卡尔加里，艾伯塔省，加拿大，2018年4月，第4839–4843页。[108] C. Weng、C. Yu、J. Cui、C. Zhang 和 D. Yu，《端到端语音识别中 RNN- 转导器的最小贝叶斯风险训练》，载于 Interspeech 会议论文集，上海，中国，2020年10月，第966–970页。[109] M. K. Baskar、L. Burget、S. Watanabe、M. Karafi´at、T. Hori 和 J. H. ˇ Cernock`y，《用于序列到序列自动语音识别的高精度前缀提升方法》，载于 IEEE ICASSP 会议论文集。布莱顿，英国：IEEE，2019年5月，第5646–5650页。[110] A. Tjandra、S. Sakti 和 S. Nakamura，《通过强化学习优化序列到序列自动语音识别》，载于 IEEE ICASSP 会议论文集。卡尔加里，艾伯塔省，加拿大：IEEE，2018年4月，第5829–5833页。[111] S. Karita、A. Ogawa、M. Delcroix 和 T. Nakatani，《使用策略梯度训练编码器-解码器模型以实现端到端语音识别》，载于 IEEE ICASSP 会议论文集。卡尔加里，艾伯塔省，加拿大：IEEE，2018年4月，第5839–5843页。[112] S. Sabour、W. Chan 和 M. Norouzi，《序列学习的最优补全蒸馏》，载于 ICLR 会议论文集，新奥尔良，路易斯安那州，2019年5月，arXiv:1810.01398。[113] W. Michel、R. Schl¨uter 和 H. Ney，《使用局部与全局对数线性组合进行早期语言模型集成》，载于 Interspeech 会议论文集，上海，中国，2020年10月，第3605–3609页。 [在线]。获取地址：http://https://arxiv.org/abs/ 2005 10049 。G. E. Hinton、S. Osindero 和[114] Y.-W. Teh，《深度置信网络的快速学习算法》，《神经计算》，第18卷，第7期，第1527–1554页，2006年7月。[115] Y. Bengio、P. Lamblin、D. Popovici 和 H. Larochelle，《深度网络的贪心逐层训练》，载于 NIPS 会议论文集，巴塞罗那，西班牙，2006年12月，第153–160页。[116] A. Zeyer、P. Doetsch、P. Voigtlaender、R. Schl¨uter 和 H. Ney，《用于语音识别声学建模的深度双向长短期记忆循环神经网络综合研究》，载于 IEEE ICASSP 会议论文集，新奥尔良，路易斯安那州，2017年3月，第2462–2466页。[117] A. Zeyer、T. Alkhouli 和 H. Ney，《RETURNN：一种通用灵活的神经网络工具包及其在机器翻译和语音识别中的应用》，载于 ACL 会议论文集，墨尔本，澳大利亚，2018年7月，第128–133页。[118] A. Zeyer、K. Irie、R. Schl¨uter 和 H. Ney，《改进语音识别端到端注意力模型的训练》，载于 Interspeech 会议论文集，海得拉巴，印度，2018年9月，第7–11页。[119] A. Zeyer、A. Merboldt、R. Schl¨uter 和 H. Ney，《注意力模型的全面分析》，载于 NIPS 会议论文集，蒙特利尔，加拿大，2018年12月。[120] Y. Chung、C. Wu、C. Shen、H. Lee 和 L. Lee，《音频 Word2Vec：使用序列到序列自编码器对音频片段表示进行无监督学习》，计算研究知识库，第abs/ 1603 00982, 卷。在线。获2016 [ ] 1603 取地址：http://arxiv.org/abs/。 00982 [121] Y.-C. Chen、S.-F. Huang、H.-y. Lee、Y.-H. Wang 和 C.-H. Shen，《音频 Word2Vec：用于音频分割和表示无监督学习的序列到序列自编码》，《IEEE/ACM 音频、语音与语言处理汇刊》，第27卷，第9期，第1481–1493页，2019年。[122] S. Scanzio、P. Laface、L. Fissore、R. Gemello 和 F. Mana，《关于多语种神经网络前端的使用》，载于 Interspeech 会议论文集，布里斯班，澳大利亚，2008年9月，第2711–2714页。[123] Z. T¨uske、J. Pinto、D. Willett 和 R. Schl¨uter，《匹配与不匹配声学条件下跨语言和多语言多层感知机特征的研究》，载于 IEEE 国际声学、语音与信号处理会议，温哥华，加拿大，2013年5月，第7349–7353页。[124] S. Zhou、S. Xu 和 B. Xu，《低资源语言上的单一 Transformer 多语种端到端语音识别》，2018年6月，arXiv:1806.05059。 [在线]。获取地址：https://arxiv.org/abs/1806.05059

[125]O. Adams、M. Wiesner、S. Watanabe 和 D. Yarowsky，《大规模多语言对抗式语音识别》，2019年4月，arXiv:1904.02210。 [在线]。获取地址：https://arxiv.org/abs/1904.02210[126] W. Hou、Y. Dong、B. Zhuang、L. Yang、J. Shi 和 T. Shinozaki，《通过多任务学习实现大规模端到端多语言语音识别与语言识别》，载于 Interspeech 会议论文集，上海，中国，2020年10月，第1037–1041页。[127] V. Pratap、A. Sriram、P. Tomasello、A. Hannun、V. Liptchinsky、G. Synnaeve 和 R. Collobert，《大规模多语言自动语音识别：50种语言、1个模型、10亿参数》，2020年，arXiv:2007.03001。

[128] 2007 03001 在线。获取地址：https://arxiv.org/abs/。B. Li、R. Pang、T. N. Sainath、A. Gulati、Y. Zhang、J. Qin、P. Haghani、W. R. Huang、M. Ma 和 J. Bai，《扩展大规模多语言自动语音识别的端到端模型》，载于 IEEE ASRU 会议论文集，2021年，第1011–1018页。[129] Y. Zhang、D. S. Park、W. Han、J. Qin、A. Gulati、J. Shor、A. Jansen、Y. Xu、Y. Huang、S. Wang、Z. Zhou、B. Li、M. Ma、W. Chan、J. Yu、Y. Wang、L. Cao、K. C. Sim、B. Ramabhadran、T. N. Sainath、F. Beaufays、Z. Chen、Q. V. Le、C.-C. Chiu、R. Pang 和 Y. Wu，《BigSSL：探索大规模半监督学习在自动语音识别中的前沿》，《IEEE信号处理精选主题期刊》，第16卷，第6期，第1519–1532页，2022年10月。

[130] 2109 13226 在线。获取地址：https://arxiv.org/abs/。Z. Chen、Y. Zhang、A. Rosenberg、B. Ramabhadran、P. J. Moreno、A. Bapna 和 H. Zen，《MAESTRO：通过模态匹配获得匹配的语音文本表示》，载于 Interspeech 会议论文集，仁川，韩国，2022年9月，arXiv:2204.03409。 [在线]。获取地址：https://arxiv.org/abs/2204 03409 。A. Radford、J. W. Kim、C.

[131] McLeavey、P. Mishkin、T. Xu、G. Brockman 和 I. Sutskever，《介绍 Whisper——通过大规模弱监督实现鲁棒语音识别》，2022年9月。 [在线]。获取地址：https://openai.com/blog/whisper/ [132] T. P. Vogl、J. Mangis、A. Rigler、W. Zink 和 D. Alkon，《加速反向传播方法的收敛》，《生物控制论》，第59卷，第4期，第257–263页，1988年。[133] N. S. Keskar 和 G. Saon，《深度神经网络 SGD 训练的非单调学习率策略》，载于 IEEE ICASSP 会议论文集。昆士兰，澳大利亚：IEEE，2015年4月，第4974–4978页。[134] S. Renals、N. Morgan、H. Bourlard、C. Wooters 和 P. Kohn，《联结主义语音识别：现状与展望》，ICSI，1991年，技术报告 TR-OI-070。[135] D. Johnson、D. Ellis、C. Oei、C. Wooters 和 P. Faerber，《QuickNet》，ICSI，伯克利，2004年。 [在线]。获取地址：http://www.icsi.berkeley.edu/Speech/qn.html [136] A. Senior、G. Heigold、M. Ranzato 和 K. Yang，《深度神经网络语音识别学习率的实证研究》，载于 IEEE ICASSP 会议论文集。温哥华，不列颠哥伦比亚省，加拿大：IEEE，2013年5月，第6724–6728页。[137] I. Loshchilov 和 F. Hutter，《解耦权重衰减正则化》，载于 ICLR 会议论文集，新奥尔良，路易斯安那州，2019年5月，arXiv:1711.05101。[138] S. L. Smith、P.-J. Kindermans、C. Ying 和 Q. V. Le，《不要衰减学习率，而要增加批量大小》，载于 ICLR 会议论文集，新奥尔良，路易斯安那州，2018年5月，arXiv:1711.00489。[139] J. Howard 和 S. Ruder，《用于文本分类的通用语言模型微调》，载于 ACL 会议论文集，墨尔本，澳大利亚，2018年6月，第328–339页。[140] M. Jaderberg、V. Dalibard、S. Osindero、W. M. Czarnecki、J. Donahue、A. Razavi、O. Vinyals、T. Green、I. Dunning、K. Simonyan、C. Fernando 和 K. Kavukcuoglu，《神经网络的基于群体的训练》，2017年11月，arXiv:1711.09846。 [在线]。获取地址：https://arxiv.org/abs/1711 09846 1 。v T. Hospedales、A. Antoniou、P.

[141] Micaelli 和 A. Storkey，《神经网络中的元学习：综述》，《IEEE模式分析与机器智能汇刊》，第PP卷，第1–20页，2021年。[142] J. L. Elman，《神经网络中的学习与发展：从小处着手的重要性》，Cognition，第48卷，第1期，第71–99页，1993年。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/0010027793900584 [143] Y. Bengio、J. Louradour、R. Collobert 和 J. Weston，《课程学习》，载于 ICML 会议论文集，蒙特利尔，魁北克省，加拿大，2009年6月，第41–48页。[144] D. Amodei、R. Anubhai、E. Battenberg、C. Case、J. Casper、B. Catan-zaro、J. Chen、M. Chrzanowski、A. Coates、G. Diamos、E. Elsen、J. Engel、L. Fan、C. Fougner、T. Han、A. Hannun、B. Jun、P. LeGresley、L. Lin、S. Narang、A. Ng、S. Ozair、R. Prenger、J. Raiman、S. Satheesh、D. Seetapun、S. Sengupta、Y. Wang、Z. Wang、C. Wang、B. Xiao、D. Yogatama、J. Zhan 和 Z. Zhu，“Deep Speech 2：英语和普通话端到端语音识别”，载于 ICML 会议论文集，纽约市，纽约州，2016年6月，第 173–182 页。[145] Z. T¨uske、G. Saon、K. Audhkhasi 和 B. Kingsbury，“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 会议论文集，中国上海，2020年10月，第 551–555 页。[146] W. Zhang、X. Chang、Y. Qian 和 S. Watanabe，“改进端到端单通道多说话人语音识别”，IEEE/ACM 音频、语音与语言处理汇刊，第 28 卷，第 1385–1394 页，2020年。[147] B. Polyak，“加快迭代方法收敛的一些方法”，《苏联计算数学与数学物理》，第 4 卷，第 5 期，第 1–17 页，1964年。 [在线]。网址：https://www.sciencedirect.com/science/article/pii/0041555364901375 [148] Y. Nesterov，“一种求解凸规划问题的方法，其收敛速率为 O( $\frac{1}{k^2}$ )”，《苏联数学文献》，第 27 卷，第 372–376 页，1983年。[149] I. Sutskever、J. Martens、G. Dahl 和 G. Hinton，“初始化和动量在深度学习中的重要性”，载于 ICML 会议论文集，佐治亚州亚特兰大，2013年6月，第 1139–1147 页。[150] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，2014年4月，arXiv:2104.02387。 [在线]。网址：https://arxiv.org/abs/1412 6980 。Z. Tuske¨、G. Saon 和 B. Kingsbury，[151] “英语会话语音识别的极限”，载于 Interspeech 会议论文集，捷克布尔诺，2021年9月，第 2062–2066 页。[152] P. Nakkiran、G. Kaplun、Y. Bansal、T. Yang、B. Barak 和 I. Sutskever，“深度双重下降：更大的模型和更多的数据何以有害”，2019年12月，arXiv:1912.02292。 [在线]。网址：https://arxiv.org/abs/1912 02292 。A. Krogh 和 J. Hertz，“简单的权重衰减即

[153] 可改善泛化”，载于神经信息处理系统进展（NIPS），科罗拉多州丹佛，1991年12月，第 950–957 页。[154] A. F. Murray 和 P. J. Edwards，“训练期间加入突触权重噪声可提升多层感知机性能和容错能力”，IEEE 神经网络汇刊，第 5 卷，第 5 期，第 792–802 页，1994年9月。[155] A. Graves，“神经网络的实用变分推断”，《神经信息处理系统进展》，第 24 卷，2011年。[156] A. Neelakantan、L. Vilnis、Q. V. Le、I. Sutskever、L. Kaiser、K. Kurach 和 J. Martens，“加入梯度噪声可改善极深网络的学习”，2015年11月，arXiv:1511.06807。 [在线]。网址：https://arxiv.org/abs/1511 06807 。G. E.

[157] Hinton、N. Srivastava、A. Krizhevsky、I. Sutskever 和 R. R. Salakhutdinov，“通过防止特征检测器协同适应来改进神经网络”，2012年7月，arXiv:1207.0580。 [在线]。1207 0580 在线。网址：https://arxiv.org/abs/。A. Krizhevsky、I. Sutskever 和 G. E. Hinton，

[158] “使用深度卷积神经网络进行 ImageNet 分类”，载于神经信息处理系统进展（NIPS），第 25 卷，内华达州太浩湖，2012年12月。[159] Y. Gal 和 Z. Ghahramani，“将 Dropout 作为贝叶斯近似：表示深度学习中的模型不确定性”，载于 ICML 会议论文集，纽约市，纽约州，2016年6月，第 1050–1059 页。[160] G. Huang、Y. Sun、Z. Liu、D. Sedra 和 K. Q. Weinberger，“采用随机深度的深度网络”，载于欧洲计算机视觉会议，荷兰阿姆斯特丹，2016年10月，第 646–661 页。[161] N.-Q. Pham、T.-S. Nguyen、J. Niehues、M. M¨uller 和 A. Waibel，“用于端到端语音识别的极深自注意力网络”，载于 Interspeech 会议论文集，奥地利格拉茨，2019年9月，第 66–70 页。[162] J. Lee 和 S. Watanabe，“基于 CTC 的语音识别的中间损失正则化”，载于 IEEE ICASSP 会议论文集，加拿大安大略省多伦多，2021年6月，第 6224–6228 页。[163] L. Wan、M. Zeiler、S. Zhang、Y. Le Cun 和 R. Fergus，“使用 DropConnect 对神经网络进行正则化”，载于 ICML 会议论文集，2013年，第 1058–1066 页。[164]D. Krueger、T. Maharaj、J. Kram´ar、M. Pezeshki、N. Ballas、N. R. Ke、A. Goyal、Y. Bengio、A. Courville 和 C. Pal，“Zoneout：通过随机保留隐藏激活来正则化循环神经网络”，2016年6月，arXiv:1606.01305。 [在线]。网址：https://arxiv.org/abs/1606 01305 。C. Szegedy、V.

[165] Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna，“重新思考用于计算机视觉的 Inception 架构”，载于 IEEE 计算机视觉与模式识别会议，内华达州拉斯维加斯，2016年6月，第 2818–2826 页。

[166] S. Bengio、O. Vinyals、N. Jaitly 和 N. Shazeer，“使用循环神经网络进行序列预测的计划采样”，NIPS 会议论文集，第 28 卷，2015年12月。[167] T. Trinh、A. Dai、T. Luong 和 Q. Le，“使用辅助损失学习 RNN 中的长期依赖关系”，载于 ICML 会议论文集，瑞典斯德哥尔摩，2018年7月，第 4965–4974 页。[168] R. J. Williams 和 J. Peng，“一种用于在线训练循环网络轨迹的高效基于梯度的算法”，IEEE《神经计算》，第 2 卷，第 4 期，第 490–501 页，1990年。[169] S. Merity、N. S. Keskar 和 R. Socher，“多种尺度下的神经语言建模分析”，2018年3月，arXiv:1803.08240。

[170] [ ] 1803 08240 在线。网址：https://arxiv.org/abs/。L. Meng、J. Xu、X. Tan、J. Wang、T. Qin 和 B. Xu，“MixSpeech：面向低资源自动语音识别的数据增强”，载于 IEEE ICASSP 会议论文集。加拿大安大略省多伦多：IEEE，2021年6月，第 7008–7012 页。[171] S. Ioffe 和 C. Szegedy，“批归一化：通过减少内部协变量偏移加速深度网络训练”，载于 ICML 会议论文集，法国里尔，2015年7月，第 448–456 页。[172] N. Kanda、R. Takeda 和 Y. Obuchi，“用于低资源语音识别的深度神经网络弹性频谱失真”，载于 IEEE ASRU 会议论文集，捷克共和国奥洛穆茨，2013年12月，第 309–314 页。[173] T. Ko 、 V. Peddinti 、 D. Povey 和 S. Khudanpur，“ 用于语音识别的音频增强”，载于 Interspeech 会议论文集，德国德累斯顿，2015年9月。[174] N. Jaitly 和 G. E. Hinton，“声道长度扰动（VTLP）可改善语音识别”，载于 ICML 会议论文集，第 117 卷，2013年6月，第 21 页。[175] G. Saon、Z. T¨uske、K. Audhkhasi 和 B. Kingsbury，“用于端到端语音识别的序列噪声注入训练”，载于 IEEE ICASSP 会议论文集，英格兰布莱顿，2019年5月，第 6261–6265 页。[176] D. S. Park、Y. Zhang、C.-C. Chiu、Y. Chen、B. Li、W. Chan、Q. V. Le 和 Y. Wu，“大规模数据集上的 SpecAugment”，载于 IEEE ICASSP 会议论文集，英国布莱顿，2019年5月，第 6879–6883 页。[177] C. Wang、Y. Wu、Y. Du、J. Li、S. Liu、L. Lu、S. Ren、G. Ye、S. Zhao 和 M. Zhou，“用于基于 Transformer 的端到端语音识别的语义掩码”，载于 Interspeech 会议论文集，中国上海，2020年10月，第 971–975 页。[178] T. Hayashi、S. Watanabe、Y. Zhang、T. Toda、T. Hori、R. Astudillo 和 K. Takeda，“用于端到端自动语音识别的回译式数据增强”，载于 IEEE SLT 会议论文集。希腊雅典：IEEE，2018年12月，第 426–433 页。[179] N. Rossenbach、M. Zeineldeen、B. Hilmes、R. Schl¨uter 和 H. Ney，“合成训练数据对各种自动语音识别架构的收益比较”，载于 IEEE ASRU 会议论文集，哥伦比亚卡塔赫纳，2021年12月，arXiv:2104.05379。[180] A. Graves，“联结主义时序分类”，载于《使用循环神经网络进行有监督序列标注》。德国海德堡：施普林格，2012年，第 61–93 页，第 章“联结主义时序分类”。[181] Y. Higuchi、S. Watanabe、N. Chen、T. Ogawa 和 T. Kobayashi，“掩码 CTC：结合 CTC 和掩码预测的非自回归端到端自动语音识别”，载于 Interspeech 会议论文集，中国上海，2020年10月，第 3655–3659 页。[182] W. Chan、C. Saharia、G. Hinton、M. Norouzi 和 N. Jaitly，“Imputer：通过插补和动态规划进行序列建模”，载于 ICML 会议论文集。PMLR，2020年7月，第 1403–1413 页。[183] Y. Fujita、S. Watanabe、M. Omachi 和 X. Chang，“用于端到端自动语音识别的基于插入的建模”，载于 Interspeech 会议论文集，中国上海，2020年10月，第 3660–3664 页。[184] L. Dong 和 B. Xu，“Cif：用于端到端语音识别的连续积分-触发”，载于 IEEE ICASSP 会议论文集，西班牙巴塞罗那，2020年5月，第 6079–6083 页。[185] J. Nozaki 和 T. Komatsu，“通过以中间预测为条件，放宽基于 CTC 的自动语音识别中的条件独立假设”，载于 Interspeech 会议论文集，捷克布尔诺，2021年9月，第 3735–3739 页。[186] Y. Higuchi、N. Chen、Y. Fujita、H. Inaguma、T. Komatsu、J. Lee、J. Nozaki、T. Wang 和 S. Watanabe，“语音转文本生成中非自回归建模的比较研究”，2021年10月，arXiv:2110.05249。 [在线]。网址：https://arxiv.org/abs/2110.05249[187] J. Chorowski 和 N. Jaitly，“探索序列到序列模型中更优的解码和语言模型集成”，载于 Interspeech 会议论文集，Stockhol，瑞典，2017年8月，第 523–527 页。

[188] W. Zhou、R. Schl¨uter 和 H. Ney，“无长度偏差的基于编码器—解码器注意力的鲁棒束搜索语音识别”，载于《Interspeech论文集》，中国上海，2020年10月，第1768–1772页。[189] P. Koehn 和 R. Knowles，“神经机器翻译面临的六项挑战”，载于《神经机器翻译首届研讨会论文集》。加拿大不列颠哥伦比亚省温哥华：计算语言学协会，2017年8月，第28–39页。[190] T. Hori、S. Watanabe 和 J. Hershey，“用于端到端语音识别的联合CTC注意力解码”，载于《ACL论文集》，加拿大不列颠哥伦比亚省温哥华，2017年7月，第518–529页。[191] Z. Tu、Z. Lu、Y. Liu、X. Liu 和 H. Li，“神经机器翻译的覆盖度建模”，载于《ACL论文集》，德国柏林，2016年5月，第76–85页。[192] T. Hori、J. Cho 和 S. Watanabe，“使用基于词的RNN语言模型进行端到端语音识别”，载于《IEEE SLT论文集》。希腊雅典：IEEE，2018年12月，第389–396页。[193] T. Hori 和 A. Nakamura，《使用加权有限状态转换器的语音识别算法》。加利福尼亚州圣拉斐尔：Morgan & Claypool Publishers，2013年。[194] Y. Miao、M. Gowayyed 和 F. Metze，“EESEN：使用深度RNN模型和基于WFST的解码进行端到端语音识别”，载于《IEEE ASRU论文集》，亚利桑那州斯科茨代尔，2015年12月，第167–174页。[195] R. Haeb-Umbach 和 H. Ney，“万词连续语音识别中的束搜索改进”，《IEEE语音与音频处理汇刊》，第2卷，第2期，第353–356页，1994年。[196] H. Ney 和 S. Ortmanns，“大词汇量连续语音识别动态规划搜索的进展”，《IEEE会刊》，第88卷，第8期，第1224–1240页，8月。

[197] 2000 [ ] 10 1109 5 880081 。在线。可访问：http://dx.doi.org/ 。/。T. Hori、Y. Kubo 和 A. Nakamura，“用于语音识别的循环神经网络语言模型实时单遍解码”，载于《IEEE ICASSP论文集》，意大利佛罗伦萨，2014年5月，第6364–6368页。[198] E. Beck、W. Zhou、R. Schl¨uter 和 H. Ney，“用于大词汇量连续语音识别首遍解码和格重评分的LSTM语言模型”，2019年7月，arXiv:1907.01030。 [在线]。可访问：https://arxiv.org/abs/1907.01030[199] G. Saon、Z. T¨uske 和 K. Audhkhasi，“RNN 转导器的对齐长度同步解码”，载于《IEEE ICASSP论文集》，西班牙巴塞罗那，2020年5月，第7804–7808页。[200] A. Y. Hannun、A. L. Maas、D. Jurafsky 和 A. Y. Ng，“使用双向循环DNN进行大词汇量连续语音识别的首遍解码”，2014年12月，arXiv:1408.2873。 [在线]。可访问：https://arxiv.org/abs/ 1408 2873 。N. Moritz、T. Hori 和 J. Le Roux，“用于端到端语音识别的触发式注意力”，载于《IEEE ICASSP论文集》。英国布莱顿：IEEE，2019年5月，第5666–5670页。[202] N. Moritz、T. Hori 和 J. Le，“使用Transformer模型的流式自动语音识别”，载于《IEEE ICASSP论文集》。西班牙巴塞罗那：IEEE，2020年5月，第6074–6078页。[203] M. Jain、K. Schubert、J. Mahadeokar、C.-F. Yeh、K. Kalgaonkar、A. Sri-ram、C. Fuegen 和 M. L. Seltzer，“用于延迟受控自动语音识别并改进束搜索的RNN-T”，2019年11月，arXiv:1911.01629。

[201] [ ] 1911 01629 在线。可访问：https://arxiv.org/abs/ 。C.-F. Yeh、J. Mahadeokar、K. Kalgaonkar、Y. Wang、D. Le、M. Jain、K. Schubert、C. Fuegen 和 M. L. Seltzer，“Transformer-转导器：使用自注意力的端到端语音识别”，载于《IEEE ICASSP论文集》，英国布莱顿，2019年5月，第7829–7833页。[205] L. Lu、C. Liu、J. Li 和 Y. Gong，“探索Transformer在大规模语音识别中的应用”，载于《Interspeech论文集》，中国上海，2020年10月，第5041–5045页。[206] T. Wang、Y. Fujita、X. Chang 和 S. Watanabe，“基于分块非自回归模型的流式端到端自动语音识别”，载于《Interspeech论文集》，捷克布尔诺，2021年9月，第3755–3759页。[207] E. Tsunoo、Y. Kashiwagi 和 S. Watanabe，“采用分块同步束搜索的流式Transformer自动语音识别”，载于《IEEE SLT论文集》。中国深圳：IEEE，2021年6月，第22–29页。[208] K. Hwang 和 W. Sung，“使用分层循环神经网络进行字符级语言建模”，载于《IEEE ICASSP论文集》。路易斯安那州新奥尔良：IEEE，2017年3月，第5720–5724页。[209] T. Hori、S. Watanabe、Y. Zhang 和 W. Chan，“基于深度CNN编码器和RNN语言模型的联合CTC注意力端到端语音识别进展”，载于《Interspeech论文集》，瑞典Stockhol，2017年8月，第949–953页。[210] A. Kannan、Y. Wu、P. Nguyen、T. N. Sainath、Z. Chen 和 R. Prab-havalkar，“将外部语言模型融入序列到序列模型的分析”，载于《IEEE ICASSP论文集》，艾伯塔省卡尔加里

，加拿大，2018年4月，第1–5828页。

[211]R. Collobert、C. Puhrsch 和 G. Synnaeve，“Wav2Letter：基于卷积网络的端到端语音识别系统”，2016年9月，arXiv:1609.03193。 [在线]。可访问：https://arxiv.org/abs/ 1609.03193[212] Y. He、T. N. Sainath、R. Prabhavalkar、I. McGraw、R. Alvarez、D. Zhao、D. Rybach、A. Kannan、Y. Wu、R. Pang、Q. Liang、D. Bhatia、Y. Shang- guan、B. Li、G. Pundak、K. C. Sim、T. Bagby、S.-y. Chang、K. Rao 和 A. Gruenstein，“面向移动设备的流式端到端语音识别”，载于 IEEE ICASSP 会议论文集，英国布莱顿，2019年5月，第6381–6385页。[213] G. Saon、Z. T¨uske、D. Bolanos 和 B. Kingsbury，“推进用于语音识别的 RNN 转导器技术”，载于 IEEE ICASSP 会议论文集。加拿大安大略省多伦多：IEEE，2021年6月，第5654–5658页，arXiv:2103.09935。 [在线]。可访问：https://arxiv.org/ abs/2103.09935[214] 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。[215] H. Seki、T. Hori、S. Watanabe、N. Moritz 和 J. Le Roux，“ 用于基于 CTC 注意力的语音识别的向量化束搜索 ”，载于 Interspeech 会议论文集，英国布莱顿，2019年5月，第3825–3829页。[216] A. Hannun、C. Case、J. Casper、B. Catanzaro、G. Diamos、E. Elsen、R. Prenger、S. Satheesh、S. Sengupta、A. Coates 和 A. Y. Ng，“Deep Speech：扩展端到端语音识别”，2014年12月，arXiv:

1412 5567 [ ] 1412 5567 . . 在线 .

[217]可访问：https://arxiv.org/abs/ . T. Hori、S. Watanabe 和 J. R. Hershey，“面向开放词汇端到端语音识别的多级语言建模与解码”，载于 IEEE ASRU 会议论文集。日本冲绳：IEEE，2017年12月，第287–293页。[218] Y. Wang、T. Chen、H. Xu、S. Ding、H. Lv、Y. Shao、N. Peng、L. Xie、S. Watanabe 和 S. Khudanpur，“Espresso：快速端到端神经语音识别工具包”，载于 IEEE ASRU 会议论文集，新加坡圣淘沙，2019年12月，第136–143页。[219] Z. T¨uske、K. Audhkhasi 和 G. Saon，“推进基于序列到序列的语音识别”，载于 Interspeech 会议论文集，奥地利格拉茨，2019年9月，第3780–3784页。[220] J. Drexler 和 J. Glass，“端到端自动语音识别的子词正则化与束搜索解码”，载于 IEEE ICASSP 会议论文集。英国布莱顿：IEEE，2019年5月，第6266–6270页。[221] T. N. Sainath、R. Pang、D. Rybach、Y. He、R. Prabhavalkar、W. Li、M. Visontai、Q. Liang、T. Strohman、Y. Wu、I. McGraw 和 C.-C. Chiu，“两遍端到端语音识别”，载于 Interspeech 会议论文集，奥地利格拉茨，2019年9月，第2773–2777页。[222] Z. Yao、D. Wu、X. Wang、B. Zhang、F. Yu、C. Yang、Z. Peng、X. Chen、L. Xie 和 X. Lei，“WeNet：面向生产的流式与非流式端到端语音识别工具包”，捷克布尔诺，第4054–4058页，2021年9月。[223] D. Wu、B. Zhang、C. Yang、Z. Peng、W. Xia、X. Chen 和 X. Lei，“U2++：用于语音识别的统一两遍双向端到端模型”，2021年12月，arXiv:2106.05642。 [在线]。可访问：https://arxiv.org/abs/ 2106 05642 M.

[224]Zapotoczny、P. Pietrzak、A. Lancucki 和 J. Chorowski，“基于注意力的语音识别模型中的格生成”，载于 Interspeech 会议论文集，奥地利格拉茨，2019年9月，第2225–2229页。[225] J. Kim、Y. Lee 和 E. Kim，“通过自适应扩展搜索加速 RNN 转导器推理”，IEEE信号处理快报，第27卷，第2019–2023页，2020年。[226] M. Abadi、P. Barham、J. Chen、Z. Chen、A. Davis、J. Dean、M. Devin、S. Ghemawat、G. Irving、M. Isard、M. Kudlur、J. Levenberg、R. Monga、S. Moore、D. G. Murray、B. Steiner、P. Tucker、V. Vasudevan、P. Warden、M. Wicke、Y. Yu 和 X. Zheng，“TensorFlow：大规模机器学习系统”，载于 OSDI 会议论文集，佐治亚州萨凡纳，2016年11月，第265–283页。[227] M. Ott、S. Edunov、A. Baevski、A. Fan、S. Gross、N. Ng、D. Grangier 和 M. Auli，“FAIRSEQ：快速且可扩展的序列建模工具包”，载于 NAACL 会议论文集，明尼苏达州明尼阿波利斯，2019年6月，第48–53页。[228] P. Doetsch、A. Zeyer、P. Voigtlaender、I. Kulikov、R. Schl¨uter 和 H. Ney，“RETURNN：用于通用循环神经网络的 RWTH可扩展训练框架”，载于 IEEE ICASSP 会议论文集。路易斯安那州新奥尔良：IEEE，2017年3月，第5345–5349页。[229] T. Mikolov、M. Karafi´at、L. Burget、J. ˇCernock`y 和 S. Khudanpur，“基于循环神经网络的语言模型”，载于 Inter- speech 会议论文集，日本幕张，2010年9月，第1045–1048页。[230] S. F. Chen 和 J. Goodman，“语言建模平滑技术的实证研究”，载于 ACL 会议论文集，加利福尼亚州圣克鲁斯，1996年6月，第310–318页。

[231] M. Sundermeyer、R. Schl¨uter 和 H. Ney，《用于语言建模的长短期记忆网络》，载于 Interspeech 会议论文集，波特兰（俄勒冈州），2012年9月，第194–197页。[232] Y. N. Dauphin、A. Fan、M. Auli 和 D. Grangier，《使用门控卷积网络进行语言建模》，载于第34届国际机器学习大会论文集，第70卷。机器学习研究期刊，2017年，第933–941页。[233] N. Zeghidour、Q. Xu、V. Liptchinsky、N. Usunier、G. Synnaeve 和 R. Collobert，《全卷积语音识别》，2018年2月，arXiv:1812.06864。 [在线]。获取地址：https://arxiv.org/abs/ 1812.06864[234] R. Al-Rfou、D. Choe、N. Constant、M. Guo 和 L. Jones，《使用更深层自注意力进行字符级语言建模》，载于 AIII 会议论文集，第33卷，檀香山（夏威夷），2019年2月，第3159–3166页。[235] K. Irie、A. Zeyer、R. Schl¨uter 和 H. Ney，《使用深层Transformer进行语言建模》，载于 Interspeech 会议论文集，格拉茨（奥地利），2019年9月，第3905–3909页。[236] Z. Dai、Z. Yang、Y. Yang、J. G. Carbonell、Q. Le 和 R. Salak hutdinov，《Transformer-XL：超越固定长度上下文的注意力语言模型》，载于 ACL 会议论文集，佛罗伦萨（意大利），2019年7月，第2978–2988页。[237] J. W. Rae、A. Potapenko、S. M. Jayakumar 和 T. P. Lillicrap，《用于长程序列建模的压缩Transformer》，《神经信息处理系统进展》，第33卷，第6154–6158页，2020年。[238] C. Gulcehre、O. Firat、K. Xu、K. Cho、L. Barrault、H.-C. Lin、F. Bougares、H. Schwenk 和 Y. Bengio，《关于在神经机器翻译中使用单语语料库》，2015年6月，arXiv:1503.03535。

[ ] 1503 03535 在线。获取地址：

[239]https://arxiv.org/abs/。A. Sriram、H. Jun、S. Satheesh 和 A. Coates，《冷融合：与语言模型联合训练序列到序列模型》，载于 Interspeech 会议论文集，海得拉巴（印度），2018年9月，第387–391页。[240] C. Shan、C. Weng、G. Wang、D. Su、M. Luo、D. Yu 和 L. Xie，《组件融合：为端到端语音识别系统学习可替换的语言模型组件》，载于 IEEE ICASSP 会议论文集。布莱顿（英国）：IEEE，2019年5月，第5361–5635页。[241] E. McDermott、H. Sak 和 E. Variani，《端到端自动语音识别中语言模型融合的密度比方法》，载于 IEEE ASRU 会议论文集，圣淘沙（新加坡），2019年12月，第434–441页。[242] Z. Meng、S. Parthasarathy、E. Sun、Y. Gaur、N. Kanda、L. Lu、X. Chen、R. Zhao、J. Li 和 Y. Gong，《领域自适应端到端语音识别的内部语言模型估计》，载于 IEEE SLT 会议论文集，深圳（中国），2020年12月，第243–250页。[243] W. Zhou、Z. Zheng、R. Schl¨uter 和 H. Ney，《基于RNN转导器的语音识别中的语言模型集成》，载于 IEEE ICASSP 会议论文集，新加坡，2022年5月，第8407–8411页，arXiv:

2110 06841 [ ] 2110 06841 。在线。

[244]获取地址：http://arxiv.org/abs/。J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向Transformer预训练》，载于 ACL 会议论文集，佛罗伦萨（意大利），2019年7月，第4171–4186页。[245] A. Radford、J. Wu、R. Child、D. Luan、D. Amodei 和 I. Sutskever，《语言模型是无监督多任务学习器》，2019年，OpenAI博客。[在线]。获取地址：https://cdn.openai.com/better-language-models/language models are unsupervised multitask learners.pdf[246] J. Salazar、D. Liang、T. Q. Nguyen 和 K. Kirchhoff，《掩码语言模型评分》，载于 ACL 会议论文集，2020年7月，第2699–2712页。[247] S. Kim、S. Dalmia 和 F. Metze，《端到端语音识别中的门控嵌入：用于对话上下文融合》，载于 ACL 会议论文集，佛罗伦萨（意大利），2019年7月，第1131–1141页。[248] A. Zeyer、E. Beck、R. Schl¨uter 和 H. Ney，《广义全和HMM训练中的CTC》，载于 Interspeech 会议论文集，斯德哥尔摩（瑞典），2017年8月，第944–948页。[249] T. Raissi、W. Zhou、S. Berger、R. Schl¨uter 和 H. Ney，《用于自动语音识别的HMM与CTC：基于从头全和训练的比较》，载于 IEEE SLT 会议论文集，多哈（卡塔尔），2023年1月，arXiv:

2210 09951 [ ] 2210 09951 。在线。

[250]获取地址：http://arxiv.org/abs/。A. Hannun、A. Lee、Q. Xu 和 R. Collobert，《使用时间深度可分离卷积的序列到序列语音识别》，载于 Interspeech 会议论文集，格拉茨（奥地利），2019年9月，第3785–3789页。[251] M. Li、M. Liu 和 H. Masanori，《使用自适应计算步数的端到端语音识别》，载于 IEEE ICASSP 会议论文集，布莱顿（英国），2019年5月，第6246–6250页。[252] P. Bahar、N. Makarov、A. Zeyer、R. Sch¨uter 和 H. Ney，《探索基于潜在注意力的零阶直接HMM自动语音识别》，载于 IEEE ICASSP 会议论文集，巴塞罗那（西班牙），2020年5月，第7854–7858页。

[253] Z. Huang、G. Zweig 和 B. Dumoulin，《用于首遍语音识别的基于缓存的循环神经网络语言模型推理》，载于 IEEE ICASSP 会议论文集，佛罗伦萨（意大利），2014年5月，第6354–6358页。[254] J. Jorge、A. Gim´enez、J. Iranzo-S´anchez、J. Civera、A. Sanchis 和 A. Juan，《使用LSTM语言模型的实时单遍语音识别解码器》，载于 Interspeech 会议论文集，格拉茨（奥地利），2019年9月，第3820–3824页。[255] W. Zhou、R. Schl¨uter 和 H. Ney，《使用LSTM语言模型对基于混合HMM的语音识别进行全和解码》，载于 IEEE ICASSP 会议论文集，巴塞罗那（西班牙），2020年5月，第7834–7838页。[256] P. Sountsov 和 S. Sarawagi，《编码器-解码器模型中的长度偏差及全局条件化方法》，载于 EMNLP 会议论文集，奥斯汀（德克萨斯州），2016年11月，第1516–1525页。[257] K. Murray 和 D. Chiang，《纠正神经机器翻译中的长度偏差》，载于 WMT 会议论文集，布鲁塞尔（比利时），2018年10月，第212–223页。[258] F. Stahlberg 和 B. Byrne，《关于神经机器翻译搜索错误和模型错误：猫把你的舌头叼走了吗？》，载于 EMNLP 会议论文集。香港（中国）：计算语言学协会，2019年11月，第3354–3360页。[259] L. Lu、X. Zhang 和 S. Renals，《大词汇量端到端语音识别中循环神经网络编码器-解码器的训练》，载于 IEEE ICASSP 会议论文集，上海（中国），2016年3月，第5060–5064页。[260] Y. Zhang、W. Chan 和 N. Jaitly，《用于端到端语音识别的超深卷积网络》，载于 IEEE ICASSP 会议论文集，新奥尔良（路易斯安那州），2017年3月，第4845–4849页。[261] S. Toshniwal、H. Tang、L. Lu 和 K. Livescu，《基于编码器-解码器的语音识别中的低层辅助任务多任务学习》，2017年4月，arXiv:1704.01631。 [在线]。获取地址：https://arxiv.org/abs/1704 01631 。A. Renduchintala、S. Ding、M. Wiesner 和 S. Watanabe，《用于端到端自动语音识别的多模态数据增强》，载于 Interspeech 会议论文集，海得拉巴（印度），2018年3月，第2394–2398页。[263] C. Weng、J. Cui、G. Wang、J. Wang、C. Yu、D. Su 和 D. Yu，《改进用于端到端英语会话语音识别的基于注意力的序列到序列模型》，载于 Interspeech 会议论文集，海得拉巴（印度），2018年9月，第761–765页。[264] D. Le、X. Zhang、W. Zheng、C. F¨ugen、G. Zweig 和 M. L. Seltzer，《从音素状态到Chenone：用于混合式语音识别的绑定上下文相关字素》，载于 IEEE ASRU 会议论文集，圣淘沙（新加坡），2019 年 12 月，第 457–464 页。[265] S. Kanthak 和 H. Ney，《在大词汇量语音识别中使用字素进行上下文相关声学建模》，载于 IEEE ICASSP 会议论文集，第2卷，奥兰多（佛罗里达州），2002年5月，第845–848页。[266] N. Deshmukh、A. Ganapathiraju 和 J. Picone，《大词汇量会话语音识别的分层搜索：探索解码问题的解决方案》，《IEEE信号处理杂志》，第16卷，第5期，第84–107页，1999年。[267] L. Nguyen 和 R. Schwartz，《用于语法定向搜索的单树方法》，载于 IEEE ICASSP 会议论文集，第2卷，菲尼克斯（亚利桑那州），1999年3月，第613–616页。[268] L. Sarı、N. Moritz、T. Hori 和 J. Le Roux，《利用基于注意力的说话人记忆进行端到端自动语音识别的无监督说话人自适应》，载于 IEEE ICASSP 会议论文集，巴塞罗那（西班牙），2020年5月，第2–6页。[269] F. Weninger、J. Andr´es-Ferrer、X. Li 和 P. Zhan，《听、关注、拼写与自适应：说话人自适应序列到序列自动语音识别》，载于 Interspeech 会议论文集。格拉茨（奥地利）：国际语音通信协会，2019年9月，第3805–3809页。[270] Z. Meng、Y. Gaur、J. Li 和 Y. Gong，《基于注意力的端到端语音识别的说话人自适应》，载于 Interspeech 会议论文集。格拉茨（奥地利）：国际语音通信协会，2019年9月，第241–245页。[271] N. Tomashenko 和 Y. Est`eve，《端到端声学模型特征空间说话人自适应的评估》，载于 LREC 会议论文集。宫崎（日本）：欧洲语言资源协会，2018年5月，第3163–3170页。[272] A. Zeyer、A. Merboldt、W. Michel、R. Schl¨uter 和 H. Ney，《带有内部语言模型先验校正的LibriSpeech转导器模型》，载于 Interspeech 会议论文集，布尔诺（捷克共和国），2021年4月，第2052–2056页。[273] L. R. Bahl、F. Jelinek 和 R. L. Mercer，《连续语音识别的最大似然方法》，《IEEE模式分析与机器智能汇刊》，第5卷，第2期，第179–190页，1983年3月。[274] J. Makhoul 和 R. Schwartz，《连续语音识别的最新进展》，《美国国家科学院院刊》，第92卷，第22期，第9956–9963页，1995年10月。[275] D. Klakow 和 J. Peters，“测试词错误率与困惑度之间的相关性”，Speech Communication，第 38 卷，2002 年，第 1 期，页码 –19。[276] M. Sundermeyer、H. Ney 和 R. Schl¨uter，“从前馈神经网络到用于语言建模的循环 LSTM 神经网络”，IEEE/ACM 音频、语音与语言处理汇刊，第 23 卷，第 3 期，第 517–529 页，2015年3月。[277] Y. Tachioka 和 S. Watanabe，“循环神经网络语言模型的判别式方法”，载于 IEEE ICASSP 会议论文集，澳大利亚布里斯班，2015年4月，第 5386–5390 页。[278] T. Hori、C. Hori、S. Watanabe 和 J. R. Hershey，“用于语音识别的长短期记忆循环神经网络语言模型的最小词错误训练”，载于 IEEE ICASSP 会议论文集，中国上海，2016年3月，第 5990–5994 页。[279] T. N. Sainath、R. Prabhavalkar、S. Kumar、S. Lee、A. Kannan、D. Rybach、V. Schogol、P. Nguyen、B. Li、Y. Wu、Z. Chen 和 C.-C. Chiu，“词典是否必不可少？评估发音词典在端到端模型中的价值”，载于 IEEE ICASSP 会议论文集，加拿大艾伯塔省卡尔加里，2018年4月，第 5859–5863 页。[280] C. Wooters 和 A. Stolcke，“说话人无关语音理解系统中的多发音词汇建模”，载于国际口语语言处理会议论文集，日本横滨，1994年9月，第 1363–1366 页。[281] I. McGraw、I. Badr 和 J. R. Glass，“使用发音混合模型从语音中学习词典”，IEEE/ACM 音频、语音与语言处理汇刊，第 21 卷，第 2 期，第 357–366 页，2012年。[282] A. Senior、G. Heigold、M. Bacchiani 和 H. Liao，“无 GMM 的 DNN 声学模型训练”，载于 IEEE ICASSP 会议论文集，意大利佛罗伦萨，2014年5月，第 5602–5606 页。 [在线]。获取地址：https://doi.org/10 1109 2014 6854675 。/ICASSP。[283] G. Gosztolya、T. Grósz 和 L. Tóth，“无 GMM 的平坦起始序列判别式 DNN 训练”，载于 Interspeech 会议论文集，N. Morgan 编。旧金山，加利福尼亚州：国际语音通信协会，2016年9月，第 3409–3413 页。 [在线]。获取地址：https://doi.org/10 21437 2016 391 。/Interspeech。[284] H. Hadian、H. Sameti、D. Povey 和 S. Khudanpur，“用于自动语音识别的平坦起始单阶段判别式训练隐马尔可夫模型”，IEEE/ACM 音频、语音与语言处理汇刊，第 26 卷，第 11 期，第 1949–1961 页，2018年。[285] H. Soltau、B. Kingsbury、L. Mangu、D. Povey、G. Saon 和 G. Zweig，“IBM 2004 年会话式电话系统：实现丰富转录”，载于 IEEE ICASSP 会议论文集，美国宾夕法尼亚州费城，2005年3月，第 205–208 页。[286] H. Hadian、D. Povey、H. Sameti、J. Trmal 和 S. Khudanpur，“利用无约束监督信息改进自动语音识别的格无关最大互信息”，载于 IEEE SLT 会议论文集，希腊雅典，2018年12月，第 43–47 页。 [在线]。获取地址：https://doi.org/10 1109 2018 8639684 。/SLT。N. Kanda、Y. Fujita 和 K. Nagamatsu，“声学模型的格无关状态级最小贝叶斯风险训练”，载于 Interspeech 会议论文集，B. Yegnanarayana 编。印度海得拉巴：国际语音通信协会，2018年9月，第 2923–2927 页。 [在线]。获取地址：https://doi.org/10 21437 2018 79 。/Interspeech。[287] S. J. Young 和 P. C. Woodland，“状态绑定在连续语音识别中的应用”，载于 Eurospeech 会议论文集，德国柏林，1993年12月，第 2203–2206 页。[288] S. Wiesler、G. Heigold、M. Nußbaum-Thom、R. Schl¨uter 和 H. Ney，“语音决策树的判别式分裂准则”，载于 Interspeech 会议论文集，日本幕张，2010年9月，第 54–57 页，入围最佳学生论文奖候选名单。[289] H. Xu、S. Ding 和 S. Watanabe，“通过发音辅助子词建模改进端到端语音识别”，载于 IEEE ICASSP 会议论文集，英国布莱顿，2019年9月，第 7110–7114 页。[290] W. Zhou、M. Zeineldeen、Z. Zheng、R. Schl¨uter 和 H. Ney，“用于端到端语音识别的声学数据驱动子词建模”，载于 Interspeech 会议论文集，捷克布尔诺，2021年8月，第 2886–2890 页。[292] T. Raissi、E. Beck、R. Schl¨uter 和 H. Ney，“迈向一致的混合隐马尔可夫模型声学建模”，2021年4月，arXiv:2104.02387。 [在线]。获取地址：https://arxiv.org/abs/2104 02387 。M. Zeineldeen、A. Zeyer、W. Zhou、T. Ng、R. Schl¨uter 和 H. Ney，“编码器—解码器—注意力模型中基于字素与基于音素的标签单元的系统性比较”，2020年11月，arXiv:2005.09336。 [在线]。获取地址：https://arxiv.org/abs/2005.09336[293] C. L¨uscher、E. Beck、K. Irie、M. Kitza、W. Michel、A. Zeyer、R. Schl¨uter 和 H. Ney，“亚琛工业大学针对 LibriSpeech 的自动语音识别系统：混合模型与注意力模型”，载于 Interspeech 会议论文集，奥地利格拉茨，2019年9月，第 231–235 页。

[295] D. Park、Y. Zhang、Y. Jia、W. Han、C.-C. Chiu、B. Li、Y. Wu 和 Q. Le，“改进的自动语音识别噪声学生训练”，载于 Interspeech 会议论文集，中国上海，2020年10月，第 2817–2821 页。[296] D. S. Park、W. Chan、Y. Zhang、C.-C. Chiu、B. Zoph、E. D. Cubuk 和 Q. V. Le，“SpecAugment：一种简单的自动语音识别数据增强方法”，载于 Interspeech 会议论文集，奥地利格拉茨，2019年9月，第 2613–2617 页。[297] W. Zhou、W. Michel、K. Irie、M. Kitza、R. Schl¨uter 和 H. Ney，“亚琛工业大学 TED-LIUM 第二版自动语音识别系统：使用 SpecAugment 改进混合隐马尔可夫模型”，载于 IEEE ICASSP 会议论文集，西班牙巴塞罗那，2020年5月，第 7839–7843 页。[298] J. Cui、B. Kingsbury、B. Ramabhadran、A. Sethy、K. Audhkhasi、X. Cui、E. Kislal、L. Mangu、M. Nussbaum-Thom、M. Picheny、Z. T¨uske、P. Golik、R. Schl¨uter、H. Ney、M. J. F. Gales、K. M. Knill、A. Ragni、H. Wang 和 P. Woodland，“面向低资源语音识别和关键词检索的多语言表征”，载于 IEEE ASRU 会议论文集，亚利桑那州斯科茨代尔，2015年12月，第 259–266 页。[299] O. Adams、M. Wiesner、S. Watanabe 和 D. Yarowsky，“大规模多语言对抗式语音识别”，载于 NAACL 会议论文集，明尼苏达州明尼阿波利斯，2019年6月，第 96–108 页。[300] A. Kannan、A. Datta、T. N. Sainath、E. Weinstein、B. Ramabhadran、Y. Wu、A. Bapna、Z. Chen 和 S. Lee，“利用流式端到端模型实现大规模多语言语音识别”，载于 Interspeech 会议论文集，奥地利格拉茨，2019年9月，第 2130–2134 页。[301] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“Librispeech：基于公有领域有声读物的自动语音识别语料库”，载于 IEEE ICASSP 会议论文集，澳大利亚昆士兰，2015年4月，第 5206–5210 页。[302] J. Godfrey、E. Holliman 和 J. McDaniel，“SWITCHBOARD：用于研究与开发的电话语音语料库”，载于 IEEE ICASSP 会议论文集，第 1 卷，旧金山，加利福尼亚州，1992年3月，第 517–520 页，第1卷。[303] Y. Wang、A. Mohamed、D. Le、C. Liu、A. Xiao、J. Mahadeokar、H. Huang、A. Tjandra、X. Zhang、F. Zhang、C. Fuegen、G. Zweig 和 M. L. Seltzer，“用于混合式语音识别的基于Transformer的声学建模”，载于 IEEE ICASSP 会议论文集，西班牙巴塞罗那，2020年5月，第 6874–6878 页。[304] G. Synnaeve、Q. Xu、J. Kahn、E. Grave、T. Likhomanenko、V. Pratap、A. Sriram、V. Liptchinsky 和 R. Collobert，“端到端自动语音识别：从监督学习到基于现代架构的半监督学习”，载于 ICML，2020年7月，arXiv:1911.08460。[305] K. Kim、F. Wu、Y. Peng、J. Pan、P. Sridhar、K. J. Han 和 S. Watanabe，“E-branchformer：用于语音识别的增强型合并 Branchformer”，载于 IEEE SLT 会议论文集，多哈，卡塔尔，2023年1月，即将发表；arXiv:2210.00077。 [在线]。获取地址：https://arxiv.org/abs/2210.00077 。M. Kitza、P. Golik、R. Schluter 和 H. Ney，[306]“双向长短期记忆声学模型的累积自适应”，载于 Interspeech，奥地利格拉茨，2019 年 9 月，第 754–758 页。[307] A. Zeyer 、P. Bahar、K. Irie、R. Schl¨uter 和 H. Ney，“用于自动语音识别的Transformer与长短期记忆编码器-解码器模型比较”，载于 IEEE ASRU 会议论文集，新加坡圣淘沙，2019年12月，第 8–15 页。[308] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过掩码预测隐藏单元进行自监督语音表征学习”，IEEE/ACM 音频、语音与语言处理汇刊，第 19 卷，2021年，第 3451–3460 页。[309] E. G. Ng、C.-C. Chiu、Y. Zhang 和 W. Chan，“突破非自回归语音识别的性能极限”，载于 Interspeech 会议论文集，捷克布尔诺，2021年9月，第 3725–2729 页。[310] J. Kahn、M. Riviere、W. Zheng、E. Kharitonov、Q. Xu、P. E. Mazare、J. Karadayi、V. Liptchinsky、R. Collobert、C. Fuegen、T. Likhomanenko、G. Synnaeve、A. Joulin、A. Mohamed 和 E. Dupoux，“Libri-Light：监督信息有限或缺失条件下自动语音识别的基准测试”，载于 IEEE ICASSP 会议论文集，西班牙巴塞罗那，2020年5月，第 7669–7673 页。[311] C.-C. Chiu、T. N. Sainath、Y. Wu、R. Prabhavalkar、P. Nguyen、Z. Chen、A. Kannan、R. J. Weiss、K. Rao、E. Gonina、N. Jaitly、B. Li、J. Chorowski 和 M. Bacchiani，“基于序列到序列模型的先进语音识别”，载于 IEEE ICASSP 会议论文集，加拿大艾伯塔省卡尔加里，2018年4月，第 4774–4778 页。[312] K. Kim、K. Lee、D. Gowda、J. Park、S. Kim、S. Jin、Y.-Y. Lee、J. Yeo、D. Kim、S. Jung、J. Lee、M. Han 和 C. Kim，“利用大型语音语料库的设备端注意力流式语音识别”，载于 IEEE ASRU 会议论文集，新加坡圣淘沙，2019年12月，第 956–963 页。[313] R. Hsiao、D. Can、T. Ng、R. Travadi 和 A. Ghoshal，“基于 Listen, Attend and Spell 模型的在线自动语音识别”，IEEE信号处理快报，第 27 卷，第 1889–1893 页，2020年。[314] Y. Shi、Y. Wang、C. Wu、C.-F. Yeh、J. Chan、F. Zhang、D. Le 和 M. Seltzer，“Emformer：高效记忆Transformer声学低延迟流式语音识别模型”，载于《IEEE ICASSP 论文集》。加拿大安大略省多伦多：IEEE，2021年6月，第6783–6787页。[315] X. Chen、Y. Wu、Z. Wang、S. Liu 和 J. Li，“在大规模数据集上开发用于语音识别的实时流式 Transformer 转导器”，载于《IEEE ICASSP 论文集》。加拿大安大略省多伦多：IEEE，2021年6月，第5904–5908页。[316] T. N. Sainath、Y. He、B. Li、A. Narayanan、R. Pang、A. Bruguier、S.-y. Chang、W. Li、R. Alvarez、Z. Chen、C.-C. Chiu、D. Garcia、A. Gruenstein、K. Hu、M. Jin、A. Kannan、Q. Liang、I. McGraw、C. Peyser、R. Prabhavalkar、G. Pundak、D. Rybach、Y. Shangguan、Y. Sheth、T. Strohman、M. Visontai、Y. Wu、Y. Zhang 和 D. Zhao，“一种流式设备端端到端模型，其质量和延迟均超越服务器端传统模型”，载于《IEEE ICASSP 论文集》，西班牙巴塞罗那，2020年5月，第6059–6063页。[317] B. Li、A. Gulati、J. Yu、T. N. Sainath、C.-C. Chiu、A. Narayanan、S.-Y. Chang、R. Pang、Y. He、J. Qin、W. Han、Q. Liang、Y. Zhang、T. Strohman 和 Y. Wu，“一种更优、更快速的流式自动语音识别端到端模型”，载于《IEEE ICASSP 论文集》，加拿大安大略省多伦多，2021年6月，第5634–5638页。[318] T. N. Sainath、Y. He、A. Narayanan、R. Botros、R. Pang、D. Rybach、C. Allauzen、E. Variani、J. Qin、Q.-N. Le-The、S.-Y. Chang、B. Li、A. Gulati、J. Yu、C.-C. Chiu、D. Caseiro、W. Li、Q. Liang 和 P. Rondon，“一种高效的流式非循环设备端端到端模型及其罕见词建模改进”，载于《Interspeech 论文集》，捷克布尔诺，2021年9月，第1777–1781页。[319] N. P. Jouppi、C. Young、N. Patil、D. Patterson、G. Agrawal、R. Bajwa、S. Bates、S. Bhatia、N. Boden、A. Borchers 等，“数据中心内张量处理单元的性能分析”，载于《 $44^{th}$ 年度计算机体系结构国际研讨会论文集》，加拿大安大略省多伦多，2017年6月，第1–12页。[320] A. Bapna、Y.-A. Chung、N. Wu、, A. Gulati、Y. Jia、J. H. Clark、M. Johnson、J. Riesa、A. Conneau 和 Y. Zhang，“SLAM：通过语音文本联合预训练实现语音与语言建模的统一编码器”，2021年10月，arXiv:2110.10329。 [在线]。网址：https://arxiv.org/abs/2110 10329 。A. Bapna、C. Cherry、Y. Zhang、Y. Jia、[321]M. Johnson、Y. Cheng、S. Khanuja、J. Riesa 和 A. Conneau，“mSLAM：面向语音与文本的大规模多语言联合预训练”，2022年2月，arXiv:2202.01374。 [在线]。网址：https://arxiv.org/abs/2202 01374 。Y. Tang、H. Gong、N. Dong、C. Wag、W. Hsu、J. Gu、A. Baevski、X. Li、A. Mohamed、M. Auli 和 J. Pino，[322]“面向语音翻译与识别的统一语音文本预训练”，载于《ACL 论文集》，爱尔兰都柏林，2022年5月，第1488–1499页，arXiv:2204.05409。[ ] 2204 05409 在线。网址：https://arxiv.org/abs/。Y.-A. Chung、C. Zhu 和 M. Zeng，[323]“SPLAT：面向口语语言理解的语音语言联合预训练”，载于《NAACL 论文集》，2021年6月，第1897–1907页，arXiv:2010.02295。 [在线]。网址：https://arxiv.org/abs/2010 02295 。J. Ao、R. Wang、L. Zhou、C. Wang、S. Ren、Y. Wu、S. Liu、T. Ko、Q. Li、Y. Zhang、Z. Wei、Y. Qian、J. Li 和 F. Wei，[324]“SpeechT5：面向口语处理的统一模态编码器—解码器预训练”，载于《ACL 论文集》，爱尔兰都柏林，2022年5月，第5723–5738页，arXiv:2110.07205。 [在线]。网址：https://arxiv.org/ abs/2110.07205[325]S. Thomas、H. J. Kuo、B. Kingsbury 和 G. Saon，“探索减少构建口语理解系统所需语音训练数据的方法”，载于《IEEE ICASSP 论文集》，新加坡，2022年5月，第7932–7936页，arXiv:2203.00006。 [在线]。网址：https://arxiv.org/abs/2203 00006 。T. N. Sainath、R. Prabhavalkar、A. Bapna、Y. Zhang、Z. Huo、Z. Chen、B. Li、W. Wang 和 T. Strohman，[326]“JOIST：语音与文本联合流式模型，用于自动语音识别”，载于《IEEE SLT 论文集》，卡塔尔多哈，2023年1月，即将发表；arXiv:2210.07353。 [在线]。网址：https://arxiv.org/abs/2210 07353 。T. Hori、R. Astudillo、T. Hayashi、Y. Zhang、S. Watanabe 和 J. Le Roux，[327]“端到端语音识别的循环一致性训练”，载于《IEEE ICASSP 论文集》，英国布莱顿，2019年5月，第6271–6275页。[328] T. Ochiai、S. Watanabe、T. Hori 和 J. R. Hershey，“多通道端到端语音识别”，载于《ICML 论文集》。澳大利亚悉尼：PMLR，2017年8月，第2632–2641页。[329] 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”，载于《IEEE ASRU 论文集》。新加坡圣淘沙：IEEE，2019年12月，第237–244页。

[330] J. Li，《端到端自动语音识别的最新进展》，2021年11月，预印本 arXiv:2111.01690。 [在线]。网址：https://arxiv.org/abs/2111.01690




**Rohit Prabhavalkar** Rohit Prabhavalkar 于2013年获得美国俄亥俄州立大学计算机科学与工程博士学位。博士毕业后，Rohit 加入谷歌语音技术组，目前担任首席研究科学家。在谷歌，他的研究主要聚焦于开发能够在移动设备上高效运行的紧凑型声学模型，以及改进端到端自动语音识别系统。Rohit 合著了50多篇经过同行评审的论文，并获得过两项最佳论文奖（ASRU 2017；ICASSP 2018）。他目前担任IEEE语音与语言处理技术委员会委员（2018—2024年），并担任《IEEE/ACM 音频、语音与语言处理汇刊》副编辑。




**Takaaki Hori** 于1999年获得日本米泽山形大学系统与信息工程博士学位。1999年至2015年，他在日本电报电话公司（NTT）的网络空间实验室和通信科学实验室从事语音识别与口语处理研究。2015年至2021年，他担任美国三菱电机研究实验室（MERL）的高级首席研究科学家。他目前是一名机器学习研究员，任职于苹果。他的研究兴趣包括自动语音识别、口语语言理解和语言建模。他曾担任IEEE语音与语言处理技术委员会委员（2020—2022年）。




**Tara Sainath** 于2009年获得麻省理工学院电气工程与计算机科学博士学位。她的博士研究主要聚焦于噪声鲁棒语音识别的声学建模。获得博士学位后，她在 IBM T.J. Watson 研究中心的语音与语言算法组工作了5年，之后加入谷歌研究院。她曾担任2017年和2018年 ICLR 的程序委员会主席。此外，她还共同组织了多个专题会议和研讨会，包括 Interspeech 2010、ICML 2013、Interspeech 2016 和 ICML 2017。在此外，她还是IEEE语音与语言处理技术委员会（SLTC）的成员，并担任《IEEE/ACM 音频、语音与语言处理汇刊》的副编辑。




**拉尔夫·施尔¨特**拉尔夫·施尔¨特于2000年在亚琛工业大学获得计算机科学自然科学博士学位，并于2019年在该校完成计算机科学教授资格认证。1996年5月，拉尔夫·施尔¨特加入亚琛工业大学计算机科学系，目前担任讲师和学术主任，负责领导计算机科学6系的自动语音识别组——机器学习与人类语言技术讲席。2019年，拉尔夫还加入了 AppTek GmbH Aachen，担任高级研究员。他的研究兴趣涵盖序列分类，尤其是自动语音识别、决策理论、随机建模和信号分析的各个方面。Ralf 曾担任 Speech Communication（2013—2019年）的专题编辑。




**Shinji Watanabe** Shinji Watanabe 是美国马里兰州巴尔的摩约翰斯·霍普金斯大学的副研究教授。他于2006年在日本东京早稻田大学获得博士学位。他曾在京都的NTT通信科学实验室担任研究科学家，并在三菱电机研究实验室（MERL）担任高级首席研究科学家。他的研究兴趣包括自动语音识别、语音增强，以及用于语音和语言处理的机器学习。他曾担任 IEEE《音频、语音与语言处理汇刊》的副编辑，也是多个技术委员会的成员，其中包括 IEEE信号处理学会语音与语言技术委员会。