INTERSPEECH 2023 2023年8月20日至24日，爱尔兰都柏林

# FunASR：一个基础的端到端语音识别工具包

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Shiliang Zhang

阿里巴巴集团达摩院语音实验室，中国  
{zhifu.gzf, lzr25946, wangjiaming.wjm, haoneng.lhn, shixian.shi, mengzhe.cmz, wucong.lyb, ailsa.zly, neo.dzh, sly.zsl}@alibaba-inc.com

**摘要**

本文介绍了FunASR，这是一款开源语音识别工具包，旨在弥合学术研究与工业应用之间的差距。FunASR提供了基于大规模工业语料库训练的模型，并支持将其部署于各类应用中。该工具包的旗舰模型Paraformer是一种非自回归端到端语音识别模型，已在包含6万小时语音的人工标注普通话语音识别数据集上完成训练。为提升Paraformer的性能，我们在标准Paraformer基础上新增了时间戳预测以及热词定制功能。此外，为了便于模型部署，我们还开源了基于前馈顺序记忆网络（FSMN-VAD）的语音活动检测模型，以及基于可控时延Transformer（CT-Transformer）的文本后处理标点模型，两者均在工业语料库上进行了训练。这些功能模块为构建高精度长音频语音识别服务奠定了坚实的基础。与其他基于开放数据集训练的模型相比，Paraformer展现出了更优异的性能。  
**索引词：**FunASR、Paraformer、语音识别、FSMN-VAD、CT-Transformer

## 1. 引言

在过去的几年里，端到端（E2E）模型在自动语音识别（ASR）任务中的性能已经超过了传统的混合系统。目前流行的三种E2E方法为：连接时序分类（CTC）[1]，循环神经网络传输器（RNN-T）[2]，以及基于注意力的编码器-解码器（AED）[3, 4]在这些模型中，AED模型因其卓越的识别准确率，一直主导着面向ASR的seq2seq建模。[4-13]开源工具包，包括ESPNET[14], WeNet[15]，PaddleSpeech[16]和 K2[17]et al.，已经被开发出来以促进端到端语音识别的研究。这些开源工具在降低构建端到端语音识别系统的难度方面发挥了巨大作用。

在本研究中，我们介绍了FunASR，这是一款旨在弥合学术研究与工业应用之间差距的新型开源语音识别工具包。FunASR基于以往研究成果，提供了若干独特功能：

1. **Modelscope:**FunASR基于行业数据提供了全面的预训练模型。其旗舰模型Paraformer[18]，是一种非自回归端到端语音识别模型，在一个包含6万小时语音的手工标注普通话语音识别数据集上进行了训练。与Conformer相比[5]和RNN-T[2]由主流开源框架支持，Paraformer在效率更高的同时提供了可比拟的性能。
2. **训练与微调：**FunASR 是一个综合工具包，提供了一系列示例配方，可以从零开始训练端到端语音识别模型，包括用于像 AISHELL1 这样的数据集的 Transformer、Conformer 和 Paraformer 模型。[19, 20]，Wenetspeech[21]和LibriSpeech[22]此外，FunASR 提供了一个方便的微调脚本，使用户能够在少量领域数据上，快速对从 ModelScope 获得的预训练模型进行微调，从而获得高性能的识别模型。此功能对于那些可能无法获得足够数据和计算资源以从头训练模型的学术研究人员和开发者来说尤其有利。
3. **语音识别服务：**FunASR 使用户能够构建可以部署到实际应用中的语音识别服务。为了便于模型部署，我们发布了一个基于前馈序列记忆网络（FSMN-VAD）的语音活动检测模型。[23]以及基于可控时延Transformer（CT-Transformer）进行文本后处理标点的模型[24]两者均在工业语料库上进行了训练。为提升Paraformer的性能，我们在标准Paraformer主干上添加了时间戳预测和热词自定义功能。此外，FunASR还包含一个推理引擎，通过ONNX、libtorch和TensorRT支持CPU和GPU推理。这些功能模块简化了使用FunASR构建高精度长音频语音识别服务的流程。

总体而言，FunASR 是一个强大的语音识别工具包，提供了其他开源工具中没有的独特功能。我们相信，我们的贡献将有助于进一步推动语音识别领域的发展，使更多的研究人员和开发者能够将这些技术应用于实际场景。需要指出的是，由于篇幅限制，本文主要报告了在普通话语料库上的实验。事实上，FunASR 支持多种类型的语言，包括英语、法语、德语、西班牙语、俄语、日语、韩语等（更多详情可参见模型库）。

[Figure 1: Overview of FunASR design](0_0.png)

https://github.com/alibaba-damo-academy/FunASR

1593 10.21437/Interspeech.2023-1428

## 2. FunASR 概述

FunASR 的整体框架如图1所示。ModelScope 管理 FunASR 中使用的模型，并托管了关键模型，如 Paraformer、FSMN-VAD 和 CT-Transformer。

FunASR 的用户可以轻松使用其基于 Pytorch 的流水线进行实验，这些流水线分为学术流水线和工业流水线。学术流水线用 表示run.sh，使用户能够从零开始训练模型。run.sh脚本遵循ESPNET的recipe风格，并包括数据准备（stage 0）、特征提取（stage 1）、字典生成（stage 2）、模型训练（stage 3和4）、以及模型推理与评分（stage 5）等阶段。相比之下，工业流程提供了两个独立的脚本：infer.sh用于推理和finetune.sh用于微调。这些流程非常易于使用，用户只需指定模型名称和数据集。

FunASR 还提供了易于使用的运行时，用于在应用程序中部署模型。为了支持 CPU、GPU、Android 和 iOS 等各种硬件平台，我们提供了包括 Libtorch、ONNX 和 TensorRT 在内的不同运行时后端。此外，我们还利用了 AMP 量化。[25]以加速推理运行时并确保最佳性能。有了这些功能，FunASR 让在各种应用中部署和使用语音识别模型变得轻而易举。

## FunASR的主要模块

### 3.1. Paraformer

首先，让我们对Paraformer做一个简要概述。[18], 我们之前提出的模型，如图2(a)所示。Paraformer是一个单步非自回归（NAR）模型，引入了基于glancing语言模型的采样模块，以增强NAR解码器捕捉token间依赖性的能力。

Paraformer由两个核心模块组成：预测器和采样器。预测器模块用于生成声学嵌入，以捕获输入语音信号中的信息。在训练过程中，采样器模块通过将目标嵌入随机替换到声学嵌入中，从而生成语义嵌入。这种方法使模型能够捕捉不同标记之间的相互依赖关系，并提升模型的整体表现。然而，在推理过程中，采样器模块不激活，声学嵌入仅通过一次传递用于输出最终预测。这种方法保证了更快的推理速度和更低的延迟。

为了进一步提升Paraformer的性能，本文提出了包括时间戳预测和热词定制在内的改进。此外，所使用的损失函数在[18]已通过移除MWER损失进行了更新，因为发现MWER损失对性能提升贡献较小。现在在第一遍解码器中使用了额外的CE损失，以减少训练和推断之间的差异。下一小节将提供详细解释。

### 3.2. 时间戳预测器

准确的时间戳预测是ASR系统的关键功能。然而，传统的工业ASR系统通常需要一个额外的混合模型来执行强制对齐（FA）以进行时间戳预测（TP），这会导致计算和时间成本的增加。FunASR 提供了一种端到端的ASR模型，通过重新设计Paraformer预测器的结构，实现了准确的时间戳预测，如图2(b)所示。我们引入了转置卷积层和LSTM层来对编码器输出进行上采样，并通过对CIF的后处理来生成时间戳。[26]权重。我们将两个fireplace之间的帧视为token的持续时间，并根据标记出静音部分。此外，FunASR还发布了一个类似force-aligner的模型，名为TP-Aligner，其中包含更小尺寸的encoder和一个timestamp predictor。它以语音和相应转录为输入来生成时间戳。

|  |
| --- |
| 表1：时间戳预测评估。 |
| 数据 |
| ———— |
| AISHELL |
|  |
| 工业 |
| 数据 |
|  |

我们在AISHELL和60,000小时的工业数据上进行了实验，以评估时间戳预测的质量。用于衡量时间戳质量的评估指标是累积平均偏移量（AAS）。[27]我们使用了一个包含5,549条话语且带有人工标记时间戳的测试集，用于对比所提供模型的时间戳预测性能与使用Kaldi训练的FA系统。[28]结果显示，Paraformer-TP在AISHELL上的表现优于FA系统。在实际应用实验中，我们发现所提出的时间戳预测方法在时间戳精度方面与混合FA系统相当（差距小于10毫秒）。此外，单次传递的解决方案对于商业用途具有重要价值，因为它有助于减少计算和时间开销。

### 3.3. 热词自定义

Contextual Paraformer 提供了通过利用命名实体自定义热词的能力，这增强了激励机制并提升了召回率和准确率。在基本的 Paraformer 模型基础上增加了两个额外的模块——热词嵌入器和在解码器最后一层的多头注意力机制，如图2(c)所示。

我们将hotwords，表示为 ，作为输入到我们的hotword embedder。[29]热词嵌入器由一个嵌入层和LSTM层组成，以上下文热词作为输入，并利用LSTM的最后一个状态生成一个嵌入，记为。具体来说，热词首先被送入热词嵌入器，产生一系列隐藏状态。然后我们使用最后一个隐藏状态作为热词的嵌入，从而捕获输入序列的上下文信息。

为了捕捉热词嵌入 与FSMN记忆块最后一层输出 之间的关系，我们采用了多头注意力模块。然后，我们将 与上下文注意力 进行拼接。该操作在公式1中形式化表示：

$$
E\_c = MultiHeadAttention(E\_y'，W^Q\_y，E\_h W^K\_h，E\_h W^V\_h) \
E\_y' = MultiHeadAttention(E\_y, W^Q\_y, HW\_x, HW^K\_y, HW^V\_y) \\
O = ConV1d(E\_y'; E\_c)
$$

我们使用一维卷积层（ConV1d）来降低其维度，以匹配隐藏状态的维度，该隐藏状态作为后续层的输入。值得注意的是，除了这一修改之外，我们的Contextual Paraformer的其他流程与标准Paraformer相同。

**表2：本定制化任务中使用的测试集。**

|  |  |  |
| --- | --- | --- |
| 数据集 | Uts | 命名实体 |
| AI domain | 486 | 204 |
| Common domain | 1308 | 231 |

在训练过程中，热词会在每个训练批次中从目标中随机生成。至于推理，我们可以通过向模型提供命名实体列表来指定热词。

**表3：热词定制评估**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 热词 | CER | R | P | F1 |
| AISHELL 热词子集 | w/o | 10.01 | 16 | 100 | 27 |
|  | w/ | 4.55 | 74 | 100 | 85 |
| 工业AI领域 | w/o | 7.96 | 70 | 98 | 82 |
|  | w/ | 6.31 | 89 | 98 | 93 |
| 工业通用领域 | w/o | 9.47 | 67 | 100 | 80 |
|  | w/ | 8.75 | 80 | 98 | 88 |

为了评估Contextual Paraformer的热词定制效果，我们通过从AISHELL测试集采样235个包含实体词的音频片段（包括187个命名实体）构建了一个热词测试集。该数据集已上传到ModelScope，测试方案已开放给FunASR。此外，我们还将实验范围扩展到了工业任务的AI领域和通用领域，具体如表2所示。

表3展示了我们关于热词对Contextual Paraformer性能影响的实验结果。我们采用了CER和F1-score作为定制化任务的评估指标。我们的结果显示，在AISHELL-1命名实体子集上，F1-score取得了大约58%的显著提升。此外，在工业定制化任务中，我们平均实现了10%的提升。

### 3.4. 语音活动检测

语音活动检测（VAD）在语音识别系统中通过检测有效语音的开始和结束起着重要作用。FunASR 提供了一种基于FSMN结构的高效VAD模型。[23]为了提高模型的区分能力，我们使用单音素作为建模单元，鉴于相对丰富的语音信息。在推理过程中，VAD系统需要进行后处理以提升鲁棒性，包括阈值设置和滑动窗口等操作。

**表4：在连续话语上对VAD的评估。**

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集 | VAD | CER | 语音百分比 |
| 会议领域 | w/o | 2.42 | 1 |
|  | w/ | 2.31 | 0.92 |
| 视频领域 | w/o | 27.87 | 1 |
|  | w/ | 25.34 | 0.78 |

VAD 的评估详见表 4。测试集由两个领域的手工标注数据组成：2 小时的会议数据和 4 小时的视频数据。我们报告了字符错误率（CER）以及发送到 ASR 推理进行识别的话语百分比。结果表明，VAD 能够有效过滤无效语音，使识别系统能够聚焦于有效语音，从而显著提升 CER。

### 3.5. 文本后处理

文本后处理是生成可读ASR转录的重要步骤，涉及添加标点符号和去除语音不流畅之处。FunASR包括一个CT-Transformer模型，可以实时执行这两项任务，如下所述。[24]该模型的整体框架如图3所示。为了满足实时性要求，该模型允许部分输出以可控的时间延迟进行冻结。采用快速解码策略以在保持高精度的同时最大限度地减少延迟。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | #Parameters | AR/NAR | LM | AISHELL test\_ios | AISHELL-2 test\_il5 | WenetSpeech test\_meeting | RTF |
| ESPNET Conformer[14] | 46.25M | AR | w/ | 4.60 | 5.70 (1.50 †) | - | - |
| WeNet Conformer-U2++[30] | 47.30M | AR | w/ | 4.40 | 5.35 | 17.34 † | - |
| PaddleSpeechDeepSpeech2[16] | 58.4M | AR | w/ | 6.40 | - | - | - |
| K2 传感器[17] | 80M | AR | w/o | 5.05 | 5.56 | 14.44 † | - |
| Conformer | 46.25M | AR | w/ | 4.65 | 5.35 | 15.21 | 1.4300 |
| Paraformer | 46.3M | NAR | w/o | 5.21 | 5.83 | - | 0.2100 |
| Paraformer-large | 220M | NAR | w/o | 1.95 | 2.85 | 6.97 | 0.0251 |

表6：文本后处理的结果

|  |  |  |  |
| --- | --- | --- | --- |
| Model | 标点符号 | 不流畅 | 推理时间 |
|  | P | R | F1 |
| BLSTM | 60.2 | 45.8 | 53.9 |
| Full-Trans. | 62.1 | 55.9 | 58.8 |
| CT-Trans. | 62.7 | 55.3 | 58.8 |

1. 实验

4.1. 评估 ASR

在我们的实验中，我们在AISHELL、AISHELL-2和WenetSpeech数据集上评估了我们的模型性能，并在表5中详细展示了结果。与其他开源工具包相比，FunASR的表现与基线Conformer模型相当。为了与Paraformer进行公平比较，我们移除了Conformer的语言模型（LM）和联合CTC解码。结果显示，在识别准确率方面，Paraformer在没有LM的情况下略优于Conformer。我们还通过在GPU（V100）上使用批量大小为1的RTF评估了模型在推理速度方面的计算效率。Paraformer模型在推理中实现了12倍的加速，这在自回归模型中是一个显著优势。即使将Paraformer-large模型的参数数量从46M增加到220M，其RTF仍然优于参数为46M的AR模型。

FunASR 提供了一个预训练的 Paraformer-large 模型，该模型专门在行业常用的6万小时普通话语音识别语料库上训练。Paraformer-large 的性能非常出色，如表5所示。在 AISHELL 测试集、AISHELL-2 test\_ios 以及 WenetSpeech test.meeting 任务上，分别实现了 1.95%、2.85% 和 6.97% 的低字符错误率（CER）。这些结果说明采用大规模语音语料库对于提升 ASR 系统性能的重要性。

此外，FunASR为我们的预训练模型提供了定制能力，使其能够在特定领域的数据上进行微调。使用AISHELL和工业领域数据（200小时）对我们的Paraformer-large模型进行微调的结果如表7所示。在AISHELL开发集和测试集任务中，我们的模型分别实现了7.4%和8.7%的相对提升。此外，在物流领域的实验表明，对Paraformer-large模型进行微调能够显著提升识别效果和领域关键词召回率。短测试集的音频平均长度为5.7秒，总时长为2.41小时。长测试集的总时长为49.5小时，音频平均长度为16.02秒。总体来看，领域关键词召回率从76.7%提升到96.8%，CER从11.25%下降到10.16%。实验结果表明，预训练模型可以通过您的语料进行微调，从而在相关领域实现显著提升。

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Dataset | 预训练(CER%/召回率%) | 微调(CER%/召回率%) |
| AISHELL dev | 1.75 / - | 1.62 / - |  |
| test | 1.95 / - | 1.78 / - |  |
| 物流 | 简短 | 9.4 / 79.4 | 7.6 / 97.6 |
| domain long | 13.1 / 74.0 | 12.6 / 96.0 |  |

4.2. 运行时基准测试

本节评估了Paraformer-large在Intel(R) Xeon(R) CPU E5-8269CY @ 2.5GHz单线程核心上的CER和RTF两方面的运行时性能。我们使用两种运行时后端对Paraformer-large的性能进行了评估，如表8所示。为优化性能，FunASR采用了文中提出的自动混合精度量化（AMP）方法。[25, 31]结果表明，AMP量化在不显著降低识别准确率的情况下，将推理速度提升了40%。

|  |  |  |  |
| --- | --- | --- | --- |
| 量化 |  | Libtorch | Onnx |
|  |  | Float32 | Int8 |
| RTF |  | 0.1062 | 0.0597 |
| CER |  | 1.95 | 1.95 |

1. 结论

本文介绍了FunASR，这是一个旨在弥合语音识别领域学术研究与工业应用之间差距的系统。FunASR提供了在大规模工业语料库上训练的模型，并能够轻松地将这些模型部署到实际应用中。我们提供了多种工业级模型，包括Paraformer-large模型、FSMN-VAD和CT-Transformer模型等。通过开放这些模型，FunASR使研究人员能够轻松地在实际场景中部署它们。

## 参考文献

[1]A. Graves、S. Fernández、F. Gomez 和 J. Schmidhuber，“连接时序分类：使用循环神经网络对未分割的序列数据进行标注”，载于第23届国际机器学习大会论文集。ACM，2006年，页369–376。

[2]A. Graves、A-r. Mohamed 和 G. Hinton，“使用深度循环神经网络的语音识别”，载于2013年IEEE国际声学、语音与信号处理会议。IEEE, 2013, 第6645–6649页。

[3]W. Chan、N. Jaitly、Q. Le 和 O. Vinyals，“Listen, attend and spell: A neural network for large vocabulary conversational speech recognition”，发表于2016年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2016年，第4960–4964页。

[4]A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“Attention is all you need”，载于《神经信息处理系统进展》，2017年，页5998–6008。

[5]A. Gulati、J. Qin、C.-C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 等，“Conformer: 用于语音识别的卷积增强型transformer。” Interspeech 2020会议论文集，2020年，第5036–5040页。

[6]Z. Gao、S. Zhang、M. Lei 和 I. McLoughlin，“San-m: 配备记忆的自注意力机制用于端到端语音识别”，arXiv 预印本 arXiv:2006.01713，2020年。

[7]S. Zhang、Z. Gao、H. Luo、M. Lei、J. Gao、Z. Yan 和 L. Xie，“面向在线端到端语音识别的流式分块感知多头注意力”，arXiv 预印本 arXiv:2006.01712，2020。

[8]Z. Gao、S. Zhang、M. Lei 和 I. McLoughlin，“Universal ASR：使用单一编码器-解码器模型统一流式与非流式ASR”，arXiv 预印本 arXiv:2010.14909，2020。

[9]M. Radfar、R. Barnwal、R. V. Swaminathan、F.-J. Chang、G. P. Strimel、N. Susani 和 A. Mouchtaris，“CROWN-HT：用于流式语音识别的卷积增强型循环神经网络换能器”，见 Interspeech 2022 论文集，2022 年，第 4431–4435 页。

[10]I. Sklyar、A. Pivovar 和 C. Oenderfor，“Separator-Transducer-Segmenter：多说话者聚会语音的流式识别与分割”，发表于 Interspeech 2022 会议论文集，2022 年，第 4451–4455 页。

[11]C.-T. Do、M. Li 和 R. Doddiapla，“多假设RNN-Transducers用于神经转导器的无监督微调和自训练”，收录于 Interspeech 2022 论文集，2022年，页码4446–4450。

[12]J. Lee、L. Lee 和 S. Watanabe，“Memory-Efficient Training of RNN-Transducer with Sampled Softmax”，发表于 Interspeech 2022 会议论文集，2022 年，第 4411–4415 页。

[13]K. Zhao、H. Nguyen、A. Jain、N. Susani、A. Mouchtaris、L. Gupta 和 M. Zhao，“通过模块替换进行知识蒸馏以实现基于循环神经网络转导器的自动语音识别”，发表于 Interspeech 2022 会议，2022 年，第 4436–4440 页。

[14]S. Watanabe、T. Hori、S. Karita、T. Hayashi、J. Nishitoba、Y. Unno、N. E. Y. Soplin、J. Heymann、M. Wiesner、N. Chen等人，“Espnet：端到端语音处理工具包”，arXiv预印本 arXiv:1804.00015，2018年。

[15]Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, 和 X. Lei，“WeNet: 面向生产的流式及非流式端到端语音识别工具包”，arXiv 预印本 arXiv:2102.01547，2021年。

[16]H. Zhang、T. Yuan、J. Chen、X. Li、R. Zheng、Y. Huang、X. Chen、E. Gong、Z. Chen、X. Hu等，“Paddlespeech：一个易用的全能语音工具包”，arXiv预印本 arXiv:2205.12007, 2022。

[17]W. Kang, L. Guo, F. Kuang, L. Lin, M. Luo, Z. Yao, X. Yang, P. Zelasco 和 D. Povey，“用于转导器的快速并行解码”，arXiv 预印本 arXiv:2211.00485，2022 年。

[18]高志、张三、I. McLoughlin 和闫志，“Paraformer: 快速而准确的并行Transformer用于非自回归端到端语音识别，” arXiv预印本 arXiv:2206.08317，2022年。

[19]H. Bu、B. Du、X. Na、B. Wu 和 H. Zheng，“Aishell-1：一个开源的普通话语音语料库及语音识别基线”，收录于2017年第20届国际语音数据库与语音I/O系统及评估东方分会（O-COCOSDA）会议。IEEE，2017年，页码1–5。

[20]J. Du、X. Na、X. Liu 和 H. Bu，“Aishell-2：将普通话ASR研究转化为工业规模，”arXiv预印本 arXiv:1808.10583，2018。

[21]B. Zhang、H. Lv、P. Guo、Q. Shao、C. Yang、L. Xie、X. Xu、H. Bu、X. Chen、C. Zeng 等人，“WenetSpeech: 一个包含10000+小时、多领域的普通话语音识别语料库”，发表于ICASSP 2022——2022年IEEE国际声学、语音与信号处理会议（ICASSP），IEEE，2022年，第6182–6186页。

[22]V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“Librispeech：一个基于公有领域有声读物的ASR语料库”，载于2015年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2015年，第5206-5210页。

[23]S. Zhang、M. Lei、Z. Yan 和 L. Dai，“Deep-Sam 用于大词汇量连续语音识别”，载于2018年IEEE声学、语音与信号处理国际会议（ICASSP），IEEE，2018年，第5869–5873页。

[24]Q. Chen、M. Chen、B. Li 和 W. Wang，“可控时延Transformer用于实时标点预测与非流利检测”，发表于ICASSP 2020—2020年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE, 2020, 第8069–8073页。

[25]Z. Gao、Y. Yao、S. Zhang、J. Yang、M. Lei 和 I. McLoughlin，“智能设备用极低占用空间的端到端ASR系统”，见Interspeech 2021论文集，2021年，页4548-4552。

[26]L. Dong 和 B. Xu，“CTC：面向端到端语音识别的连续积分-发放模型”，发表于 ICASSP 2020-2020 IEEE 国际声学、语音与信号处理会议（ICASSP）。IEEE，2020，第6079-6083页。

[27]X. Shi、Y. Chen、S. Zhang 和 Z. Yan，“在非自回归端到端ASR模型中实现时间戳预测与识别”，arXiv预印本 arXiv:2301.12343，2023。

[28]D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz 等人，“Kaldi语音识别工具包”，发表于IEEE 2011自动语音识别与理解研讨会，编号CONF。IEEE信号处理学会，2011年。

[29]G. Pundak、T. N. Sainath、R. Prabhavalkar、A. Kannan 和 D. Zhao，“Deep context: 端到端上下文语音识别”，载于2018年IEEE口语语言技术研讨会（SLT），IEEE，2018，第418–425页。

[30]B. Zhang、D. Wu、Z. Feng、X. Song、Z. Yao、H. Lv、L. Xie、C. Yang、P. Fan 和 J. Niu，“Wenet 2.0: 更高效的端到端语音识别工具包”，arXiv 预印本 arXiv:2301.15455，2022。

[31]K. Zhu, W. Zhao, Z. Zheng, T. Guo, P. Zhao, J. Bai, J. Yang, X. Liu, L. Bian, 和 W. Lin，“Disc：面向机器学习工作负载的动态形状编译器”，载于《第1届机器学习与系统研讨会论文集》，2021年，第89-95页。