端到端流式语音识别研究综述

王澳回[1],张 珑[1],宋文宇[2],孟 杰[1]
1.天津师范大学 计算机与信息工程学院,天津 300387
2.广州华立科技职业学院 计算机信息工程学院,广州 511325

摘 要:语音识别是实现人机交互的一种重要途径,是自然语言处理的基础环节,随着人工智能技术的发展,人机交互等大量应用场景存在着流式语音识别的需求。流式语音识别的定义是一边输入语音一边输出结果,它能够大大减少人机交互过程中语音识别的处理时间。目前在学术研究领域,端到端语音识别已经取得了丰硕的研究成果,而流式语音识别在学术研究以及工业应用中还存在着一些挑战与困难,因此,最近两年,端到端流式语音识别逐渐成为语音领域的一个研究热点与重点。从端到端流式识别模型与性能优化等方面对近些年所展开的研究进行全面的调查与分析,具体包括以下内容:(1)详细分析和归纳了端到端流式语音识别的各种方法与模型,包括直接实现流式识别的CTC与RNN-T模型,以及对注意力机制进行改进以实现流式识别的单调注意力机制等方法;(2)介绍了端到端流式语音识别模型提高识别准确率与减少延迟的方法,在提高准确率方面,主要有最小词错率训练、知识蒸馏等方法,在降低延迟方面,主要有对齐、正则化等方法;(3)介绍了流式语音识别一些常用的中英文开源数据集以及流式识别模型的性能评价标准;(4)讨论了端到端流式语音识别模型的未来发展与展望。

关键词:人机交互;语音识别;端到端;流式;延迟
文献标志码:A **中图分类号:**TN912.34 doi:10.3778/j.issn.1002-8331.2206-0306

Review of End-to-End Streaming Speech Recognition

WANG Aohui[1], ZHANG Long[1], SONG Wenyu[2], MENG Jie[1]
1.College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China
2.College of Computer Information Engineering, Guangzhou Huali Vocational College of Science and Technology, Guangzhou 511325, China

Abstract: Speech recognition is an important way to realize human-computer interaction and the basic link of natural language processing. With the development of artificial intelligence technology, streaming speech recognition is required in a large number of application scenarios such as human-computer interaction. Streaming speech recognition is defined as input speech and output result. It can greatly reduce the processing time of speech recognition in human-computer interaction. At present, end-to-end speech recognition has achieved fruitful research achievements in the academic research field, while streaming speech recognition still has some challenges and difficulties in academic research and industrial applications. Therefore, in the last two years, end-to-end speech recognition has gradually become a research hotspot and focus in the field of speech. From the aspects of end-to-end streaming recognition model and performance optimization, the research in recent years is comprehensively investigated and analyzed, including the following contents: (1) Various methods and models of end-to-end streaming speech recognition are analyzed and summarized in detail, including CTC and RNN-T models which directly realize streaming speech recognition, and monotone attention mechanism which improves attention mechanism to realize streaming speech recognition. (2) The methods to improve the recognition accuracy and reduce the delay of the end-to-end streaming speech recognition model are introduced. In terms of improving the accuracy, there are mainly methods such as minimum word error rate training and knowledge distillation, and in terms of reducing the delay, there are mainly methods such as alignment and regularization. (3) Some common Chinese and English open source data

基金项目:国家自然科学基金面上项目(61771173);天津市自然科学基金重点项目(20JCZDJC00400)。
作者简介:王澳回(2000—),男,硕士研究生,CCF学生会员,研究方向为语音识别;张珑(1978—),通信作者,男,教授,CCF高级会员,研究方向为语音识别、群体智能,E-mail:zhanglong@tjnu.edu.cn;宋文宇(1983—),男,博士,工程师;孟杰(2000—),女,硕士研究生,研究方向为语音信号处理。
收稿日期:2022-06-17 **修回日期:**2022-09-02 **文章编号:**1002-8331(2023)02-0022-12

sets and performance evaluation criteria of streaming speech recognition models are introduced. (4) The future development and prospect of the end-to-end streaming speech recognition model are discussed.

Key words: human-computer interaction; speech recognition; end to end; streaming; delay

语音识别模型从最初的基于GMM-HMM[1]的模型，发展到基于DNN-HMM[2-4]深度神经网络模型，再到现在的端到端[5-8]语音识别模型，已经历经三个阶段。通过这三个阶段的发展，模型结构越加简单，语音识别的准确率几乎趋于饱和状态，然而，大部分模型都是针对非流式语音识别而言的，在测试模型性能的时候很少会去考虑模型识别延迟的问题。近几年来，语音识别模型进入端到端的时代，不再依赖传统语音识别系统中已经使用了几十年的建模组件，使用单个网络便可将输入的语音序列直接转换成输出的标签序列，使得模型的尺寸更小，因此，大量研究人员开始从深度神经网络模型转向研究端到端语音识别模型，另外，大量的研究证明，端到端模型已经在学术研究领域[7]以及工业生产领域[9-10]超越了基于DNN-HMM的深度神经网络模型。未来几年，端到端模型将是语音识别领域研究的重点。常见的端到端模型有CTC[11]、RNN-T[12]、attention-based encoder-decoder[13-14]、LAS[8]等模型，前两种能够直接实现流式识别，而后两种模型由于注意力机制需要获取完整的声学序列而不能够直接进行流式识别。流式语音识别又称为实时语音识别，它指的是用户在说话的时候模型便已经开始进行识别，与之相对的非流式识别则是用户说完了一句话或一段话之后模型开始识别。随着科技的不断发展，各种穿戴式、便携式的智能设备，以及大量的应用软件已经完全融入大众生活，常用的输入法、在线会议、直播、实时翻译等一系列的应用存在着流式语音识别的需求。端到端流式识别模型不需要额外的语言模型，更容易部署在设备端，另外，智能客服等多种需要流式识别的人机交互场景也在不断产生，所以端到端流式语音识别模型将会是未来几年的研究热点，而且也具有广阔的应用前景。因此，本文主要从模型结构、性能优化、常用的中英文开源数据集以及模型性能评价标准等方面分析总结了目前端到端流式语音识别模型的研究状况，进而提出了未来的发展与展望。

2021年国外有两篇相关的语音识别领域的综述，文献[15]主要总结了近十年语音识别模型结构与性能的发展，并从研究与应用两个方面预测了语音识别未来十年的发展趋势。文献[16]详细概述了端到端语音识别模型的发展及其在实际工业生产中的应用情况，同时从行业角度出发，重点介绍了端到端语音识别模型如何去解决未来的应用部署中的一些挑战与困难。以上两篇文章都是从大的领域、更高视野出发，总结概述端到端语音识别的发展，而这篇文章，则是聚焦到端到端流式语音识别这个领域，去分析总结其发展现状。

1 端到端流式语音识别模型

1.1 可直接实现流式识别的端到端模型

在端到端流式语音识别模型中，能够直接进行流式识别的模型主要有connectionist temporal classification (CTC)[11]、recurrent neural network transducer(RNN-T)[12]、recurrent neural aligner(RNA)[17]等模型。文献[11]提出connectionist temporal classification(CTC)损失函数，用来对模型中的循环神经网络产生的转录进行评分，使得模型能够完成音频帧与标签的自动对齐。从端到端语音识别模型的发展来看，CTC最先被应用到端到端语音识别模型[5-6,18-23]，它能够直接将输入的语音序列转换成输出的标签序列，其结构如图1[16]所示，输入的语音序列 x_t 通过编码器进行编码输出特征表示 h_t^{enc}，再经过一个线性分类器得到每个时刻输出类别的概率 $P(y_t|x_t)$。

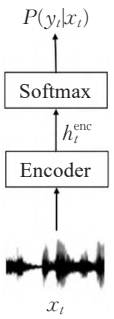

图1 CTC结构

Fig.1 Structure of CTC

通过在编码器中使用单向的循环神经网络(unidirectional RNN)，CTC模型能够实现流式语音识别。文献[12]提出了recurrent neural network transducer(RNN-T)模型，该模型为流式语音识别提供了一种自然的方式，因为它的输出取决于之前的输出标签序列和当前步及之前的输入语音序列，即 $P(y_u|x_{1:t}, y_{1:u-1})$，通过这种方式，消除了CTC的条件独立假设，由于其具备自然的流式性质，在该领域应用中受到了广泛的使用[9,24-31]。

RNN-T模型的结构如图2[16]所示，它包含一个编码器网络、一个预测网络和一个联合网络，编码器将输入的语音序列 x_t 转换成高级特征表示 h_t^{enc}，预测网络基于RNN-T之前的输出标签 $y_{1:u-1}$，生成高级表示 h_u^{pre}，联合网络是一个前馈网络，将 h_t 与 h_u 作为输入，输出 $z_{t,u}$。

针对CTC所存在条件独立性假设的问题，文献[17]提出了一种新的模型：recurrent neural aligner(RNA)，类似于CTC模型，该模型定义了目标标签序列上的概率分布，包括对应于输入中每个时间步长的空白标签，

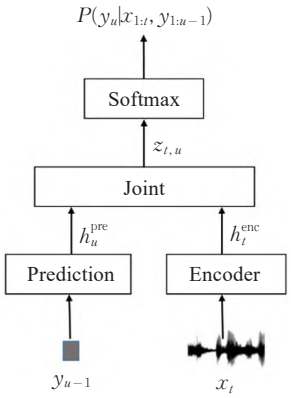

图2　RNN-T结构

Fig.2　Structure of RNN-T

通过边缘化所有可能的空白标签位置来计算标签序列的概率。但该模型并不做标签预测的条件独立性假设，此外，它在输入的每个时间步预测一个输出标签，而不是通过RNN-T预测多个标签，从而简化了波束搜索解码，使得训练更加有效，在执行流式语音识别任务时，它成功地应用于多种口语识别任务[32]。

1.2　改进后可实现流式识别的端到端模型

在端到端语音识别模型中，基于注意力[33-36]的模型由于其自身特点不能够直接实现流式识别，而这些模型已经被证明在机器翻译[37-38]、语音识别[34,39]等领域的许多问题中非常有效，在该结构中，首先，编码器对整个输入序列进行编码，产生相对应的隐藏状态序列，其次，解码器根据编码器所产生的状态序列来进行预测，最终产生输出序列。目前，基于注意力的端到端模型已在相关的语音识别[34,40]任务中取得了重大进展，在识别准确率方面，实现了非流式语音识别模型的最好性能[39]。然而，基于注意力模型并不能够直接应用于流式语音识别问题，一方面，这些模型通常需要获取完整的声学序列作为输入，使得编码与解码不能够同步进行；另一方面，对于语音来说，它们没有固定的长度，模型的计算复杂度随着输入序列的增加而二次增加。为了能够将注意力机制应用于流式语音识别任务中，大量的研究人员针对以上问题开展研究，通过对全局注意（local attention）机制做出改进，针对在时刻 t 将哪一部分的输入序列信息进行编码，同时对于已编码的信息，将哪一部分进行解码的问题，提出了基于单调注意力机制（monotonic attention mechanism）[41-45]、基于块（chunk-wise）[46-51]、基于信息累积（accumulation of information）[52-55]以及触发注意（triggered attention）[56-58]等方法。

1.2.1　基于单调注意力机制的方法

文献[42]提出了一种局部单调注意（local monotonic attention）机制，它具有局部性和单调性，局部性帮助模型的注意模块专注于解码器想要转录的输入序列的某一个部分，单调性严格地从输入序列的开始到结束左右生成对齐。该机制迫使模型在每个解码步骤预测中心位置，并仅在中心位置周围计算软注意权重。然而，仅仅基于有限的信息，很难准确预测下一个中心位置。与软注意相比较，硬单调性约束限制了模型的表达能力，文献[43]提出了单调组块注意（monotonic chunk-wise attention，MoChA）机制来缩小软、硬注意之间性能的差距，它基于预测的选择概率自适应地将编码的状态序列分割成小的组块，如图3[43]所示，块边界由虚线表示，允许模型在硬单调注意机制选择参与的小组块上执行软注意，但是它的训练过程非常复杂困难，以至于最终难以实现。

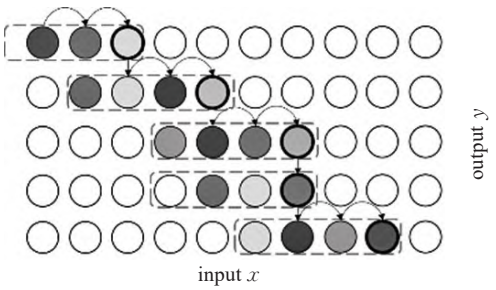

图3　单调组块注意

Fig.3　Monotonic chunk-wise attention

文献[44]提出了单调多头注意（monotonic multi-head attention，MMA），该机制结合了多层多头注意和单调注意的优点，同时提出了两种变体，即Hard MMA（MMA-H）和Infinite Lookback MMA（MMA-IL），前者在设计时考虑到了注意力持续时间必须有限的流式系统，而后者强调识别系统的质量。文献[45]对于一些应用局部单调注意机制的模型的变体进行了修改，同时也对这些模型进行了全面的比较，最后通过采用固定大小的窗口实现了一种简单有效的启发式执行局部注意的方法。

1.2.2　基于块的方法

文献[46]提出了Neural Transducer，它根据部分观察到的输入序列和部分生成的序列来计算下一步的分布，使用编码器来处理输入，将处理后的结果作为Transducer的输入，在每个时间步长，根据编码器处理好的输入块，Transducer决定可以产生零到多个输出标签，由此实现流式解码，然而，由于该模型受到循环神经网络时间相关特性的束缚，它仅仅优化对应于组块序列的近似最佳对齐路径。文献[47]使用自注意模块替代了RNN-T结构中的RNN模块，提出了一种自注意transducer（self-attention transducer，SAT），它能够利用自注意块来模拟序列内部的长期依赖性，同时引入了块流（block-flow）机制，通过应用滑动窗口来限制自注意的范围，并且堆叠多个自注意块来模拟长期依赖性，但从整体而言，虽然块流机制能够帮助SAT实现流式解码，但仍然引起了识别准确率的下降。因此，文献[49]提出了一种同步transformer（synchronous transformer，Sync-

Transformer)模型，能够同步进行编码与解码，其结构与推理过程如图4[49]所示。Sync-Transformer将transformer与SAT深入组合，为了消除self-attention机制对于未来帧的依赖，则强制编码器中的每个节点仅仅关注左侧上下文并完全忽略右侧上下文。一旦编码器产生了固定长度的状态序列块，解码器则立即开始预测标签。

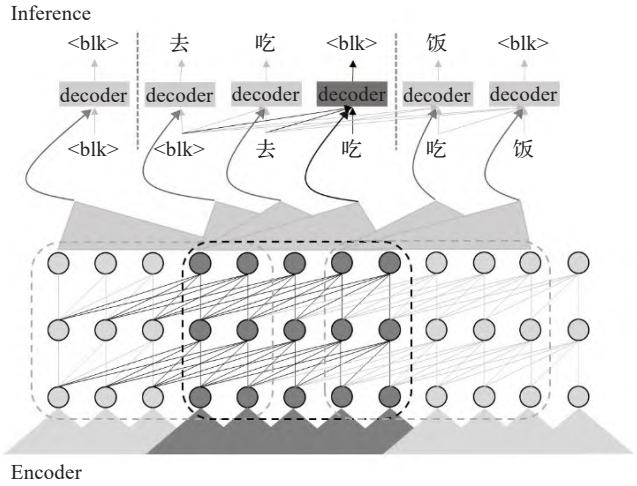

图4　Synchronous Transformer的结构与推理过程

Fig.4　Structure and reasoning process of Synchronous Transformer

1.2.3　基于信息堆叠的方法

文献[53]提出了自适应时间(adaptive computation time，ACT)算法，该算法支持RNN以学习在接受输入和产生输出之间需要采取多少计算步骤，为后续自适应计算步的研究打下了基础。文献[54]提出了一种新颖的自适应计算步算法(adaptive computation steps，ACS)，该算法使端到端语音识别模型能够动态地决定应该处理多少帧来预测语言输出，一方面，对准器在思考间隔内计算每个编码器时间步长停止的概率，并且像基于软注意模型一样来总结上下文向量，另一方面，该模型不断检查停止概率的累积，如果总和达到阈值之后立即做出输出的决定。文献[55]提出了解码器端自适应计算步算法(decoder-end adaptive computation steps，DACS)来解决标准transformer不能够直接用于流式识别的问题，该算法通过在从编码器状态获得的置信度达到某个阈值之后触发输出来传送transformer ASR的解码，通过引入最大前瞻(look-ahead)性步骤来限制DACS层可以查看每个输出步骤的时间步数，以防止过快地达到语音结束，但DACS对transformer解码器采用异步多头注意机制，破坏了在线解码的稳定性。受到spiking neural networks中的integrate-and-fire模型的启发，文献[66]提出了用于序列转换的新型软单调对比机制continuous integrate-and-fire(CIF)，能够支持各种在线识别任务以及声学边界定位。在每个编码器步中，接受当前编码器步的向量表示和缩放向量中包含的信息量的相应权重，向前累积权重并积分向量信息，直到累积的权重达到阈值，此时声学边界被定位，且当前的编码器步的声学信息由两个相邻标签共享，CIF将信息分为两个部分：一部分用于完成当前标签的集成；另一部分用于下一个标签的集成，模拟处理在编码器步期间的某个时间点触发时，将集成的声学信息触发到解码器以预测当前标签，如图5[56]所示，每条虚线代表一次触发，直到整个声学序列完成编码。文献[57]提出了存储器自注意传感器(memory-self-attention transducer，MSAT)，其结构如图6[57]所示，MSA模块将历史信息添加到受限制的自我注意单元中，通过参与存储器状态有效地模拟长时间的上下文，并使用RNN损失来对MSA模块进行训练，实现了该结构在流式任务中的应用。

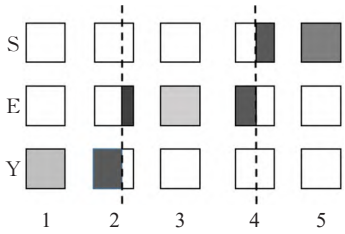

图5　CIF编码过程

Fig.5　Encoding process of CIF

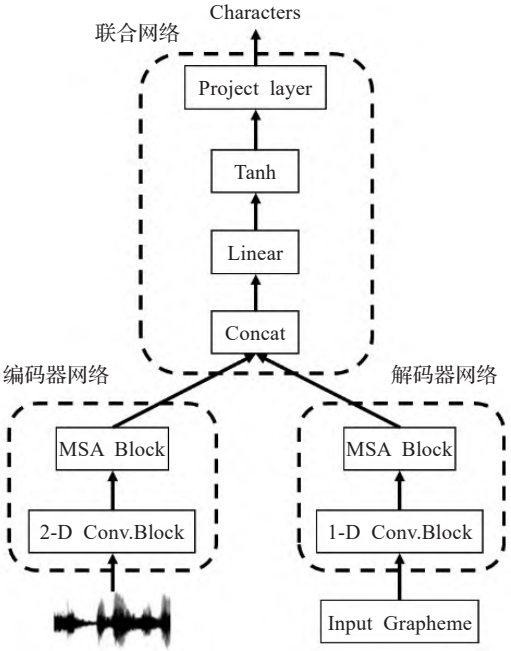

图6　MAST结构

Fig.6　Structure of MAST

1.2.4　其他方法

以上提出来的方法能够实现流式语音识别，但也存在问题。基于单调注意力机制的方法由于使用软硬注意机制导致训练过程非常困难；基于块的方法往往由于忽略组块之间的关系而导致性能下降；而基于信息堆叠的方法打破了Transformer在训练中的并行性，通常需要更长的训练时间[58]。文献[59]提出了触发注意(triggered

attention,TA)[59-61],其结构如图7[59]所示,TA解码器由一个触发模型和一个基于注意的解码器神经网络组成,编码器神经网络由触发网络和注意机制共享。注意权重只能看到触发事件之前的编码器帧及一些向前的帧。在训练期间,CTC输出序列的强制对齐用于导出触发的时间,在解码期间,考虑CTC训练的触发模型的不确定性以分别生成替代的触发序列和输出序列,推理以帧同步解码方式进行。此外,一些研究人员用Transformer替换了RNN-T结构中的RNN,构建了Transformer Transducer(TT)[62-69]结构,大量的研究[62-69]证明了该结构也具有较好的流式识别能力。

2　端到端流式语音识别模型的优化方法与策略

端到端流式语音识别模型是当前语音识别领域的研究热点与重点,对于非流式模型而言,需要占用尽可能小的内存去实现更高的识别准确率,然而,对于流式识别模型,既需要考虑模型的识别准确率又需要考虑识别的延迟大小。这两个方面共同决定了流式语音识别模型的性能。以下将从延迟与准确率两个方面来探索流式语音识别模型的优化问题。

2.1　如何降低流式语音识别模型的延迟

当识别一句话时,一般有两种语音延迟[70]:第一种是第一标签产生延迟(first token emission delay),通过分析用户实际说话开始时间与语音识别系统实际产生出第一个标签的时间可以获取到该种延迟的时间;第二种是用户感知延迟(user perceived latency),当用户停止说话时开始计时,直到模型发出最后一个非空标签,一般将这段时间称为用户感知延迟。

近期研究[70]表明影响流式语音识别模型用户感知延迟的主要因素有模型结构、训练标准、解码超参数以及端点指示器,而模型的大小与模型计算速度并不总是严重影响用户感知延迟。目前,研究人员主要从训练策略、对齐与正则化[71]训练等角度出发来探索如何降低模型的延迟,文献[72]提出一种自适应的前瞻(adaptive look-ahead)方法来权衡延迟和词错率,其中的上下文窗口大小并不固定,可以动态地修改,引入scout network(SN)和recognition network(RN)两个神经组件,其中,scout network负责检测语音中一个单词的开始和结束边界,recognition network通过向前看预测边界进行帧同步单通道解码,虽然这个方法在权衡延迟与准确率方面取得了很好的效果,但SN没有解决随着左上下文长度的平方增长的繁重的自我注意计算。文献[73]基于MoChA提出了最小延迟训练策略(minimum latency training strategies),利用从混合模型中提取的外部硬对齐作为监督,迫使模型学习准确的对齐方式,在解码器端提出了延迟约束训练(DeCoT)和最小延迟训练(MinLT)两种方法,有效地减少了模型的延迟。文献[74]则从模型结构与端点指示器出发,提出了一个双通道的RNN-T+LAS模型,其中LAS对RNN-T的假设进行重评分,同时通过预测查询结束(end-of-query)符号,将EOQ端点指示器集成到端到端模型中,用来帮助关闭麦克风,这种方法实现了端到端模型在质量与延迟的权衡方面对传统混合模型的首次超越。文献[75]提出了一种

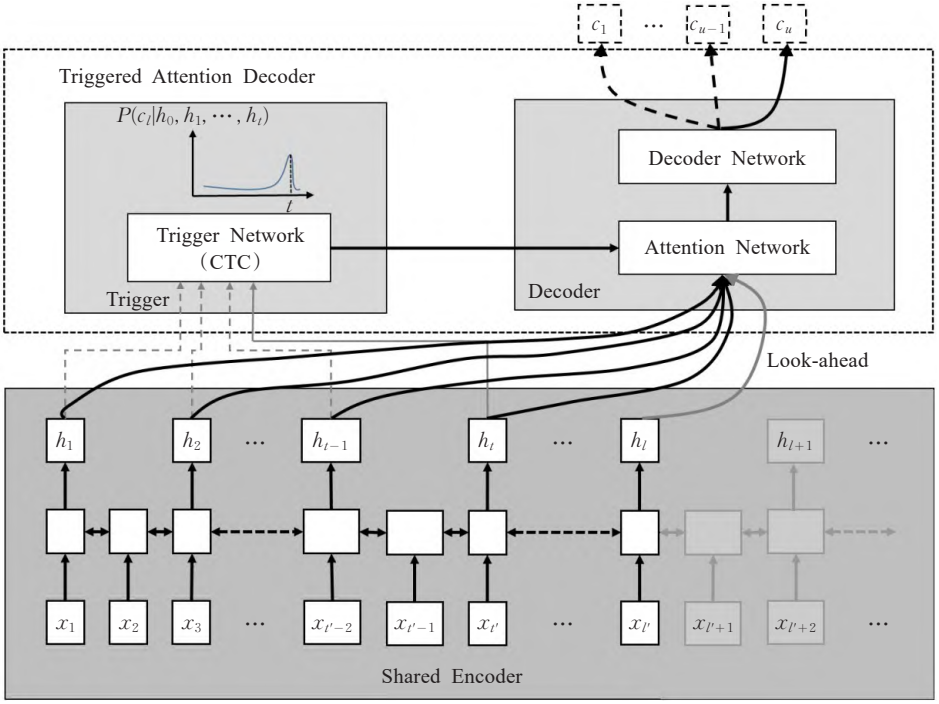

图7　Triggered attention系统结构

Fig.7　System structure of Triggered attention

新的延迟约束方法：自对准，该方法不需要外部对准模型，而是通过利用自训练模型的维特比强制对齐来寻找较低延迟对齐方向。文献[76]从延迟正则化训练的角度出发，基于Transducer的流式模型提出了一种新的序列级产生正则化方法FastEmit，在训练transducer模型时能够直接对每序列概率应用延迟正则化，而不需要任何语音-单词对齐信息，同时，相较于其他正则化方法，FastEmit方法需要调整的超参数最少。通过在大量端到端模型上展开实验，表明该方法能够实现很好的词错率与延迟的权衡。通过以上研究可知目前已有限制对齐、正则化等多种方法可以相对解决流式语音识别模型的延迟问题，大多数的方法虽然降低了模型的延迟，但同时也导致了识别质量的下降，这将是未来仍需不断探索的一个研究方向。

2.2 如何提高流式语音识别模型的准确率

提高语音识别模型的准确率一直是个热门话题，从1988年第一个基于隐马尔科夫模型(HMM)的语音识别系统Sphinx[77]诞生开始，到现在语音识别模型步入端到端的时代，研究人员不断做出探索希望语音识别模型的准确率能够得到进一步提升，从传统混合模型[78]到深度神经网络模型[2]再到现在的端到端模型[40]，模型结构改变的同时，语音识别模型准确率也得到大幅度提升。与非流式模型一样，提升流式模型准确率的方式有改变模型基本结构、预训练、扩大数据域、最小词错率训练(MWER)[79-83]、知识蒸馏[84-89]等方式，其中，改变模型结构已在第1章进行阐述。文献[73]以MoChA作为流式语音识别模型，在编码器端，采用了多任务学习并使用帧交叉熵目标进行预训练，提升了模型的识别准确率。文献[83]提出了一种新颖且有效的基于RNN-T模型的MWER训练算法，对 N 个最佳列表中每个假设的所有可能对比的得分求和，并使用它们来计算参考和假设之间的预期编辑距离，当为endpointer(EP)添加end-of-sentence(EOS)，所提出的MWER训练还可以显著减少高删除错误。文献[84]研究了基于知识蒸馏的模型压缩方法来训练CTC声学模型，评估了CTC模型的帧级知识蒸馏方法和序列级知识蒸馏方法，通过在WSJ数据集上展开实验，提高了模型的识别准确率。文献[90]实现了从非流式双向RNN-T模型到流式单向RNN-T模型的知识蒸馏，实验结果表明，通过所提出的知识蒸馏训练的单向RNN-T比用标准方法训练的单向模型具有更好的准确性。文献[85]研究了非流式到流式Transformer-Transducer模型的知识蒸馏，在实验中比较了两种不同的方法：隐藏向量的L2距离最小化和头部L2距离的最小化，实验结果表明，基于隐藏向量相似性的知识蒸馏优于基于多头相似性的知识蒸馏。

3 数据集与评估标准

3.1 数据集

在语音识别领域，ASR模型性能的优劣不仅仅与模型的架构有关，同时也依赖于量大且质量高的数据集。随着互联网与一些终端设备的不断发展，每天都会产生大量的数据信息，通过对电话录音、新闻、智能家居、科学研究等领域相关语音信息的收集，各大科研机构、数据公司相继发布了一系列的语音数据集。为语音识别领域的科研发展提供了基本的实验条件。目前，一些科研机构、数据公司已经开源了他们的数据集，以供学术界免费使用进行科学研究，通过OpenSLR平台，能够获取来自世界各地的开源语音数据资源，然而，由于法律以及商业等一方面的原因，大量的数据集需要购买才能够获得相关的使用权限。本节将主要介绍一些中文普通话以及英语等常见的一些数据集。

中文语音识别开源数据集如表1所示，2015年，清华大学信息技术研究院语音语言技术中心发布了第一个开源中文语音数据库THCHS30[91]，以帮助研究人员搭建起第一个语音识别系统。但是该数据集的语音总时长仅仅只有35 h，对于模型的训练还不够充分，2017年，北京希尔贝壳科技有限公司发布了AISHELL-1[92]语料库，成为了当时最大的开源汉语语音识别语料库，冲浪科技也发布了ST-CMDS语音数据集[93]，2018年，北京希尔贝壳科技有限公司发布了AISHELL-2[94]语料库，上海原语公开了Primewords Set1数据集，2019年，数据堂(北京)科技有限公司开源了中文普通话语音数据集DTZH1505[93]，记录了6 408位来自中国八大方言地域、33个省份的说话人的自然语言语音，时长达1 505 h，语料内容涵盖社交聊天、人机交互、智能客服以及车载命令等[93]，这是目前最大最全面的中文开源语音数据集。

表1 部分常用汉语普通话开源数据集

Table 1 Part of common Mandarin open source data set

数据集	开源时间	时长/h	人数	语料内容
THCHS30[91]	2015年	35	50	以新闻为主
AISHELL-1[92]	2017年	178	400	金融、新闻、科技、数字序列等11个领域
ST-CMDS[93]	2017年	500	855	以网上语音聊天、智能语音控制为主
AISHELL-2[94]	2018年	1 000	1 991	唤醒词、语音控制词、工业生产等12个领域
Primewords Set1	2018年	100	296	—
DTZH1505[93]	2019年	1 505	6 408	社交聊天、人机交互等领域

在语音识别领域,最早开源的是一些国外的语音数据集,如表2所示,正是由于这些科研机构、企业开源了大量的优质数据集,在此基础上,语音识别模型的性能能够得到一次又一次的提升。1993年,美国的一些科研机构发布了语音数据集TIMIT[95],该数据集旨在为获取声学语音知识以及开发和评估自动语音识别系统提供语音数据,由于该数据集较小同时标记信息比较完整,研究人员能够快速完成实验并展现出模型的性能。此后,美国等多个科研机构开源了多个大型语音数据集,例如TED-LIUM[96]、LibriSpeech[97]、Common Voice[98]、MLS[99]、The People's Speech[100]、GigaSpeech[101],这些数据集中的数据通过智能设备、音频录制、自动合成等多种方式进行获取,此外,一些数据集也采集了一些无标签数据用于无监督学习。

3.2 评价指标

对于端到端流式识别模型来说,主要通过模型的准确率与识别的延迟两个方面来评价其性能的优劣,在准确率方面,通过计算出语句的词错率(word error rate,WER)或者字错率(character error rate,CER)来评价模型,常用词错率来计算,把 T 作为一句话中的总单词数,S 作为识别结果中替换单词数,D 作为识别结果中删除的正确话语中的单词数[102],I 作为没有在正确话语中而出现在识别结果中的插入单词数,那么词错率(WER)则定义为:

$$WER=\frac{S+D+I}{T} \tag{1}$$

WER的值越低,则说明模型的识别准确率越高,性能越好。在延迟方面,实时因子(real time factor,RTF)则是流式语音识别过程中的评价标准,它的值小于1的时候,称模型是实时识别的,此外也可以计算出语句级或词语级的延迟数值(latency)。把 M 作为一段音频的时长,把 N 作为识别出这段音频的时长,则实时因子(RTF)则定义为:

$$RTF=\frac{N}{M} \tag{2}$$

RTF的值越小,则说明延迟越小,模型的性能越好。

4 流式语音识别模型的未来发展方向与应用

虽然端到端语音识别模型已经超越了传统混合模型的性能,实现了输入语音序列直接产生对应的标签序列,极大程度简化了模型的训练过程,但端到端流式语音识别仍是一个需要重点关注的任务,在其性能准确率与识别延迟的权衡问题上仍然值得研究人员去深入研究与探索。本章从七个方面提出一些问题,这些问题值得今后进一步去思考研究。

(1)"词错率-延迟"如何权衡。

一般来说,减小语音识别的延迟常常需要以降低识别精确度为代价。对于一个流式语音识别模型,可以通过大量的实验绘制出词错率-延迟曲线,随着延迟的降低,其词错率在随之增加,词错率-延迟的权衡问题,其折中点在何处?在可以接受的识别质量的情况下,其能做到的最小延迟是多少?这仍需要结合实际的应用需求来进一步地探索。

(2)流式与非流式模型的统一结构

在模型的结构方面,常见的模型为流式识别模型或者非流式识别模型,它们都是流式或非流式单一结构。基于全注意力的端到端模型能实现最优性能,因此,在处理非流式任务时,研究人员一般选择基于全注意力的端到端模型,以实现更高的准确率,但是,在处理流式任务时,则会对模型结构进行改变,选择CTC模型、RNN-T模型以及改进的注意力模型以牺牲准确率的代价来减小识别的延迟。训练一个模型能够实现流式识别与非流式识别两种需求,同时大幅减少模型开发、训练以及部署的成本,因此,流式与非流式模型的统一结构将会是未来语音识别领域的一个研究重点与热点问题。文献[103]提出了一个框架U2来将流式识别与非流式识别相统一,不仅降低了流式模型与非流式模型之间的精度差距,同时大幅度减少了成本。

(3)自监督预训练模型

相较于传统的语音识别模型,端到端语音识别模型更需要大规模的数据。由于中文普通话、英语等语言受到广泛的使用,获取这类语言大规模数据集并不是一件困难的事,但当面临中文方言或者一些比较小众语言时,想要获取数据集便十分困难,获取其大规模的数据集更是难上加难。因此,可以通过自监督学习来预训练端到端流式语音识别模型,在预训练的过程中不需要带有标签的数据,有效解决低资源的问题。

表2 部分常用外语开源数据集

Table 2 Part of foreign languages open source data set

数据集	发布时间	时长/h	人数	语料内容
TIMIT[95]	1993年	5	630	阅读材料等
TED-LIUM[96]	2012年	118	—	音频讲座等
LibriSpeech[97]	2015年	1 000	—	文本和语音的有声读物等
Common Voice[98]	2017年	2 500	50 000	多领域的日常生活语音
MLS[99]	2020年	50 500	—	有声读物
The People's Speech[100]	2020年	31 400	—	网络音频等

(4)轻体量的个性化语言模型

传统的语音识别模型由独立的声学、发音与语言模型组成,而端到端语音识别模型则将这三种独立的模型统一成一个神经网络,对于体量较大的传统语音识别模型来说,其识别精确度优于端到端模型的主要原因是在于其具有非常大的语言模型。因此,为了提升端到端流式语音识别模型的识别准确率,可以在模型的解码阶段引入一个轻体量的个性化语言模型,这样做既不会大幅增强模型的推理时间,同时又能够实现热词增强和个性化解码。

(5)端到端流式语音识别模型后处理

在流式语音识别过程中,模型能够通过部分上下文即可快速输出识别结果,但在该过程中由于获取的下文内容受到限制可能也会导致识别结果出现一些错误。纠错模型和双通道重评分机制是语音识别后处理的两个重要策略,但大部分纠错模型由于采用自回归结构导致其存在较大的延迟,并不适用于端到端流式语音识别模型后处理。然而基于多输入的快速纠错模型FastCorrect2[104]的提出,使得在端到端流式语音识别模型后处理过程中引入快速纠错模型成为可能,通过快速纠错模型或者双通道重评分机制,可以对语音识别的结果进行检测,快速纠正其中的错误,在保持低延迟的情况下,能够进一步地提升端到端流式语音识别模型的性能。

(6)基于设备端部署小尺寸流式语音识别模型

近些年来,用户数据泄漏、隐私受到侵犯、遭遇诈骗等热点问题频发,用户个人隐私问题越来越受到重视,同时,智能家居、智能手机以及各种可穿戴设备进入人们的生活当中。近期,一些科研人员开始研究基于设备端的流式语音识别模型部署问题[105-110]。语音识别模型一般部署在服务器端,将音频以流的方式传输到服务器端,在服务器上进行识别,最终将结果传输到终端设备上,而随着端到端模型的发展,它不需要额外的语言模型,以便将模型部署在设备端,直接在设备上完成识别工作,这有助于保护用户的隐私,同时能够通过减少数据传输时间进一步减少设备的识别延迟[70],增加模型识别的稳定性。因此,未来基于设备端部署小尺寸的流式语音识别模型则将成为工业界应用的趋势。

(7)流式语音识别模型的工业应用

目前,端到端流式语音识别模型是学术研究与工业应用的一个热点问题,随着人工智能技术的不断发展,出现了智能客服[111]、语音售票机等大量需要流式语音识别的人机交互场景,通过将端到端流式语音识别模型应用到这些场景,能够大幅提升语音识别的效率,节省人力,提高服务的效率。因此,未来几年,将会出现更多的流式识别的语音场景,流式语音识别模型也将会更广泛地应用到工业产品中。

自2014年以来,端到端语音识别模型成为了第三代语音识别模型,在语音领域掀起了研究狂潮,同时,端到端流式语音识别也成为语音识别领域的一个热点与重点问题,受到学术界广泛关注,大量科研单位开展了深入的研究并取得了丰硕的研究成果。本文从流式模型实现方式、优化策略、开源数据集与评价标准、未来发展等方面进行研究、总结与分析,最后也讨论了未来几年流式识别模型的发展方向,希望能够为该领域的一些研究人员提供一些帮助。

参考文献:

[1] BILMES J A.What HMMs can do[J].IEICE Transactions on Information and Systems,2006,89(3):869-891.

[2] LI J, YU D, HUANG J T, et al.Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM[C]//2012 IEEE Spoken Language Technology Workshop(SLT),2012:131-136.

[3] MIAO Y,METZE F.Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training[C]//Proceedings of INTERSPEECH,2013:2237-2241.

[4] SHAHIN M,AHMED B,MCKECHNIE J,et al.A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech[C]//Fifteenth Annual Conference of the International Speech Communication Association,2014:1583-1587.

[5] HANNUN A,CASE C,CASPER J,et al.Deep speech:scaling up end-to-end speech recognition[J].arXiv:1412.5567,2014.

[6] GRAVES A,JAITLY N.Towards end-to-end speech recognition with recurrent neural networks[C]//International Conference on Machine Learning,2014:1764-1772.

[7] WATANABE S,HORI T,KIM S,et al.Hybrid CTC/attention architecture for end-to-end speech recognition[J].IEEE Journal of Selected Topics in Signal Processing,2017,11(8):1240-1253.

[8] CHAN W,JAITLY N,LE Q V,et al.Listen,attend and spell[J].arXiv:1508.01211,2015.

[9] LI J,ZHAO R,MENG Z,et al.Developing RNN-T models surpassing high-performance hybrid models with customization capability[J].arXiv:2007.15188,2020.

[10] SAINATH T N,HE Y,LI B,et al.A streaming on-device end-to-end model surpassing server-side conventional model quality and latency[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2020:6059-6063.

[11] GRAVES A,FERNÁNDEZ S,GOMEZ F,et al.Connectionist temporal classification:labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning,2006:

369-376.
[12] GRAVES A.Sequence transduction with recurrent neural networks[J].arXiv:1211.3711,2012.
[13] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//Advances in Neural Information Processing Systems,2017:6000-6010.
[14] BAHDANAU D,CHOROWSKI J,SERDYUK D,et al. End-to-end attention-based large vocabulary speech recognition[C]//2016 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2016: 4945-4949.
[15] HANNUN A.The history of speech recognition to the year 2030[J].arXiv:2108.00084,2021.
[16] LI J.Recent advances in end-to-end automatic speech recognition[J].arXiv:2111.01690,2021.
[17] SAK H,SHANNON M,RAO K,et al.Recurrent neural aligner: an encoder-decoder neural network model for sequence to sequence mapping[C]//Proceedings of INTERSPEECH, 2017:1298-1302.
[18] MIAO Y,GOWAYYED M,METZE F.EESEN:end-to-end speech recognition using deep RNN models and WFST-based decoding[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU), 2015:167-174.
[19] SOLTAU H,LIAO H,SAK H.Neural speech recognizer: acoustic-to-word LSTM model for large vocabulary speech recognition[J].arXiv:1610.09975,2016.
[20] ZWEIG G,YU C,DROPPO J,et al.Advances in all-neural speech recognition[C]//2017 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),2017:4805-4809.
[21] ZEYER A,BECK E,SCHLÜTER R,et al.CTC in the context of generalized full-sum HMM training[C]//Proceedings of INTERSPEECH,2017:944-948.
[22] LI J,YE G,DAS A,et al.Advancing acoustic-to-word CTC model[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018: 5794-5798.
[23] AUDHKHASI K,KINGSBURY B,RAMABHADRAN B, et al.Building competitive direct acoustics-to-word models for English conversational speech recognition[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018:4759-4763.
[24] SAON G,TÜSKE Z,BOLANOS D,et al.Advancing RNN transducer technology for speech recognition[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:5654-5658.
[25] PRABHAVALKAR R,RAO K,SAINATH T N,et al.A comparison of sequence-to-sequence models for speech recognition[C]//Proceedings of INTERSPEECH,2017:939-943.
[26] SAINATH T N,HE Y,LI B,et al.A streaming on-device end-to-end model surpassing server-side conventional model quality and latency[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),2020:6059-6063.
[27] HE Y,SAINATH T N,PRABHAVALKAR R,et al.Streaming end-to-end speech recognition for mobile devices[C]// 2019 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2019:6381-6385.
[28] BATTENBERG E,CHEN J,CHILD R,et al.Exploring neural transducers for end-to-end speech recognition[C]// 2017 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU),2017:206-213.
[29] LI J,ZHAO R,HU H,et al.Improving RNN transducer modeling for end-to-end speech recognition[C]//2019 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU),2019:114-121.
[30] ZHANG X,ZHANG F,LIU C,et al.Benchmarking LF-MMI, CTC and RNN-T criteria for streaming ASR[C]//2021 IEEE Spoken Language Technology Workshop(SLT),2021:46-51.
[31] PUNJABI S,ARSIKERE H,RAEESY Z,et al.Joint ASR and language identification using RNN-T:an efficient approach to dynamic language switching[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:7218-7222.
[32] DONG L,ZHOU S,CHEN W,et al.Extending recurrent neural aligner for streaming end-to-end speech recognition in mandarin[J].arXiv:1806.06342,2018.
[33] VINYALS O,TOSHEV A,BENGIO S,et al.Show and tell: a neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015:3156-3164.
[34] CHOROWSKI J K,BAHDANAU D,SERDYUK D,et al. Attention-based models for speech recognition[C]//Advances in Neural Information Processing Systems,2015:577-585.
[35] KIM S,HORI T,WATANABE S.Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]// 2017 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2017:4835-4839.
[36] DONG L,XU S,XU B.Speech-transformer:a no-recurrence sequence-to-sequence model for speech recognition[C]// 2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018:5884-5888.
[37] BAHDANAU D,CHO K,BENGIO Y.Neural machine translation by jointly learning to align and translate[J]. arXiv:1409.0473,2014.
[38] LUONG M T,MANNING C D.Stanford neural machine translation systems for spoken language domains[C]//Proceedings of the 12th International Workshop on Spoken Language Translation:Evaluation Campaign,2015.

[39] CHIU C C,SAINATH T N,WU Y,et al.State-of-the-art speech recognition with sequence-to-sequence models[C]// 2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018:4774-4778.

[40] CHAN W,JAITLY N,LE Q,et al.Listen,attend and spell: a neural network for large vocabulary conversational speech recognition[C]//2016 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP), 2016:4960-4964.

[41] RAFFEL C,LUONG M T,LIU P J,et al.Online and linear-time attention by enforcing monotonic alignments[C]// International Conference on Machine Learning,2017: 2837-2846.

[42] TJANDRA A,SAKTI S,NAKAMURA S.Local monotonic attention mechanism for end-to-end speech and language processing[J].arXiv:1705.08091,2017.

[43] CHIU C C,RAFFEL C.Monotonic chunkwise attention[J]. arXiv:1712.05382,2017.

[44] MA X,PINO J,CROSS J,et al.Monotonic multihead attention[J].arXiv:1909.12406,2019.

[45] MERBOLDT A,ZEYER A,SCHLÜTER R,et al.An analysis of local monotonic attention variants[C]//Proceedings of INTERSPEECH,2019:1398-1402.

[46] JAITLY N,SUSSILLO D,LE Q V,et al.A neural transducer[J].arXiv:1511.04868,2015.

[47] TIAN Z,YI J,TAO J,et al.Self-attention transducers for end-to-end speech recognition[J].arXiv:1909.13037,2019.

[48] TSUNOO E,KASHIWAGI Y,WATANABE S.Streaming Transformer ASR with blockwise synchronous beam search[C]//2021 IEEE Spoken Language Technology Workshop(SLT),2021:22-29.

[49] TIAN Z,YI J,BAI Y,et al.Synchronous transformers for end-to-end speech recognition[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),2020:7884-7888.

[50] SAINATH T N,CHIU C C,PRABHAVALKAR R,et al. Improving the performance of online neural transducer models[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018: 5864-5868.

[51] JAITLY N,LE Q V,VINYALS O,et al.An online sequence-to-sequence model using partial conditioning[C]//Advances in Neural Information Processing Systems,2016:1-11.

[52] DI GANGI M A,NEGRI M,TURCHI M.Adapting transformer to end-to-end spoken language translation[C]// Proceedings of INTERSPEECH,2019:1133-1137.

[53] GRAVES A.Adaptive computation time for recurrent neural networks[J].arXiv:1603.08983,2016.

[54] LI M,LIU M,MASANORI H.End-to-end speech recognition with adaptive computation steps[C]//2019 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2019:6246-6250.

[55] LI M,ZORILĂ C,DODDIPATLA R.Transformer-based online speech recognition with decoder-end adaptive computation steps[C]//2021 IEEE Spoken Language Technology Workshop(SLT),2021:1-7.

[56] DONG L,XU B.Cif:continuous integrate-and-fire for end-to-end speech recognition[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2020:6079-6083.

[57] LUO J,WANG J,CHENG N,et al.Unidirectional memory-self-attention transducer for online speech recognition[C]// 2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:910-914.

[58] WANG F,XU B.Shifted chunk encoder for transformer based streaming end-to-end ASR[J].arXiv:2203.15206,2022.

[59] MORITZ N,HORI T,LE ROUX J.Triggered attention for end-to-end speech recognition[C]//2019 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2019:5666-5670.

[60] ZHAO H,HIGUCHI Y,OGAWA T,et al.An investigation of enhancing CTC model for triggered attention-based streaming ASR[J].arXiv:2110.10402,2021.

[61] MORIYA T,ASHIHARA T,ANDO A,et al.Hybrid RNN-T/ attention-based streaming ASR with triggered chunkwise attention and dual internal language model integration[C]//2022 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2022:8282-8286.

[62] YEH C F,MAHADEOKAR J,KALGAONKAR K,et al. Transformer-transducer:end-to-end speech recognition with self-attention[J].arXiv:1910.12977,2019.

[63] ZHANG Q,LU H,SAK H,et al.Transformer transducer: a streamable speech recognition model with transformer encoders and rnn-t loss[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),2020:7829-7833.

[64] DALMIA S,LIU Y,RONANKI S,et al.Transformer-transducers for code-switched speech recognition[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:5859-5863.

[65] XIE Y,MACOSKEY J,RADFAR M,et al.Compute cost amortized transformer for streaming ASR[J].arXiv:2207. 02393,2022.

[66] SUN E,LI J,MENG Z,et al.Improving multilingual transformer transducer models by reducing language confusions[C]//Proceedings of INTERSPEECH,2021:3470-3474.

[67] CHEN X,WU Y,WANG Z,et al.Developing real-time streaming transformer transducer for speech recognition on large-scale dataset[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing

(ICASSP),2021:5904-5908.
[68] SHI Y,WU C,WANG D,et al.Streaming transformer transducer based speech recognition using non-causal convolution[C]//2022 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2022:8277-8281.
[69] XIA W,LU H,WANG Q,et al.Turn-to-diarize:online speaker diarization constrained by transformer transducer speaker turn detection[C]//2022 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2022:8077-8081.
[70] SHANGGUAN Y,PRABHAVALKAR R,SU H,et al.Dissecting user-perceived latency of on-device E2E speech recognition[J].arXiv:2104.02207,2021.
[71] BA J L,KIROS J R,HINTON G E.Layer normalization[J].arXiv:1607.06450,2016.
[72] WANG C,WU Y,LIU S,et al.Low latency end-to-end streaming speech recognition with a scout network[J].arXiv:2003.10369,2020.
[73] INAGUMA H,GAUR Y,LU L,et al.Minimum latency training strategies for streaming sequence-to-sequence ASR[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2020:6064-6068.
[74] SAINATH T N,HE Y,LI B,et al.A streaming on-device end-to-end model surpassing server-side conventional model quality and latency[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2020:6059-6063.
[75] KIM J,LU H,TRIPATHI A,et al.Reducing streaming ASR model delay with self alignment[J].arXiv:2105.05005,2021.
[76] YU J,CHIU C C,LI B,et al.Fastemit:low-latency streaming asr with sequence-level emission regularization[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:6004-6008.
[77] LEE K F,HON H W,REDDY R.An overview of the SPHINX speech recognition system[J].IEEE Transactions on Acoustics,Speech,and Signal Processing,1990,38(1):35-45.
[78] GALES M J F.Maximum likelihood linear transformations for HMM-based speech recognition[J].Computer Speech & Language,1998,12(2):75-98.
[79] SHANNON M.Optimizing expected word error rate via sampling for speech recognition[J].arXiv:1706.02776,2017.
[80] PRABHAVALKAR R,SAINATH T N,WU Y,et al.Minimum word error rate training for attention-based sequence-to-sequence models[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018:4839-4843.
[81] CUI J,WENG C,WANG G,et al.Improving attention-based end-to-end ASR systems with sequence-based loss functions[C]//2018 IEEE Spoken Language Technology Workshop(SLT),2018:353-360.
[82] WENG C,YU C,CUI J,et al.Minimum bayes risk training of RNN-transducer for end-to-end speech recognition[J].arXiv:1911.12487,2019.
[83] GUO J,TIWARI G,DROPPO J,et al.Efficient minimum word error rate training of RNN-transducer for end-to-end speech recognition[J].arXiv:2007.13802,2020.
[84] TAKASHIMA R,LI S,KAWAI H.An investigation of a knowledge distillation method for CTC acoustic models[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018:5809-5813.
[85] KOJIMA A.Knowledge distillation for streaming transformer-transducer[C]//Proceedings of INTERSPEECH,2021:2841-2845.
[86] INAGUMA H,KAWAHARA T.Alignment knowledge distillation for online streaming attention-based speech recognition[J].arXiv:2103.00422,2021.
[87] PANCHAPAGESAN S,PARK D S,CHIU C C,et al.Efficient knowledge distillation for rnn-transducer models[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:5639-5643.
[88] DOUTRE T,HAN W,MA M,et al.Improving streaming automatic speech recognition with non-streaming model distillation on unsupervised data[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:6558-6562.
[89] KIM Y,RUSH A M.Sequence-level knowledge distillation[J].arXiv:1606.07947,2016.
[90] KURATA G,SAON G.Knowledge distillation from offline to streaming RNN transducer for end-to-end speech recognition[C]//Proceedings of INTERSPEECH,2020:2117-2121.
[91] WANG D,ZHANG X.Thchs-30:a free Chinese speech corpus[J].arXiv:1512.01882,2015.
[92] BU H,DU J,NA X,et al.Aishell-1:an open-source Mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment(O-COCOSDA),2017:1-5.
[93] 王东,王丽媛,王大亮,等.DTZH1505:大规模开源中文普通话语音库[J].计算机工程与应用,2022,58(11):295-301.
WANG D,WANG L Y,WANG D L,et al.DTZH1505:large scale open source Mandarin speech corpus[J].Computer Engineering and Applications,2022,58(11):295-301.
[94] DU J,NA X,LIU X,et al.Aishell-2:transforming Mandarin asr research into industrial scale[J].arXiv:1808.10583,2018.

[95] GAROFOLO J S,LAMEL L F,FISHER W M,et al.DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM[R].1993.

[96] ROUSSEAU A,DELÉGLISE P,ESTEVE Y.TED-LIUM:an automatic speech recognition dedicated corpus[C]//Proceedings of LREC,2012:125-129.

[97] PANAYOTOV V,CHEN G,POVEY D,et al.Librispeech:an asrcorpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2015:5206-5210.

[98] ARDILA R,BRANSON M,DAVIS K,et al.Common voice:a massively-multilingual speech corpus[J].arXiv:1912.06670,2019.

[99] PRATAP V,XU Q,SRIRAM A,et al.Mls:a large-scale multilingual dataset for speech research[J].arXiv:2012.03411,2020.

[100] GALVEZ D,DIAMOS G,CIRO J,et al.The People's speech:a large-scale diverse English speech recognition dataset for commercial usage[J].arXiv:2111.09344,2021.

[101] CHEN G,CHAI S,WANG G,et al.Gigaspeech:an evolving,multi-domain asr corpus with 10,000 hours of transcribed audio[J].arXiv:2106.06909,2021.

[102] MCCOWAN I A,MOORE D,DINES J,et al.On the use of information retrieval measures for speech recognition evaluation[R].LIDIAP,2004:1-13.

[103] ZHANG B,WU D,YAO Z,et al.Unified streaming and non-streaming two-pass end-to-end model for speech recognition[J].arXiv:2012.05481,2020.

[104] LENG Y,TAN X,WANG R,et al.Fastcorrect 2:fast error correction on multiple candidates for automatic speech recognition[J].arXiv:2109.14420,2021.

[105] HE Y,SAINATH T N,PRABHAVALKAR R,et al.Streaming end-to-end speech recognition for mobile devices[C]//2019 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2019:6381-6385.

[106] KIM K,LEE K,GOWDA D,et al.Attention based on-device streaming speech recognition with large speech corpus[C]//2019 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU),2019:956-963.

[107] GARG A,VADISETTI G P,GOWDA D,et al.Streaming on-device end-to-end ASR system for privacy-sensitive voice-typing[C]//Proceedings of INTERSPEECH,2020:3371-3375.

[108] SAINATH T N,HE Y,LI B,et al.A streaming on-device end-to-end model surpassing server-side conventional model quality and latency[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2020:6059-6063.

[109] OH Y R,PARK K.On-device streaming transformer-based end-to-end speech recognition[J].Proceedings of INTERSPEECH,2021:967-968.

[110] ZHANG Y,SUN S,MA L.Tiny transducer:a highly-efficient speech recognition model on edge devices[C]//2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2021:6024-6028.

[111] 颜永红,张鹏远,徐及,等.智能语音能力平台关键技术及其在智能客服行业应用[EB/OL].(2020-05-10)[2022-09-12].https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=SNAD&filename=SNAD000001823879.
YAN Y H,ZHANG P Y,XU J,et al.Key technologies of intelligent voice capability platform and its application in intelligent customer service industry[EB/OL].(2020-05-10)[2022-09-12].https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=SNAD&filename=SNAD000001823879.