融合外部语言知识的流式越南语语音识别

王俊强[1,2]，余正涛*[1,2]，董凌[1,2]，高盛祥[1,2]，王文君[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
814220330@qq.com, ztyu@hotmail.com, 46761956@qq.com
gaoshengxiang.yn@foxmail.com, 175360805@qq.com

摘要

越南语为低资源语言，训练语料难以获取；流式端到端模型在训练过程中难以学习到外部大量文本中的语言知识，这些问题在一定程度上都限制了流式越南语语音识别模型的性能。因此，本文以越南语音节作为语言模型和流式越南语语音识别模型的建模单元，提出了一种将预训练越南语语言模型在训练阶段融合到流式语音识别模型的方法。在训练阶段，通过最小化预训练越南语语言模型和解码器的输出计算一个新的损失函数L_{AED-LM}，帮助流式越南语语音识别模型学习一些越南语语言知识从而优化其模型参数；在解码阶段，使用Shallow Fusion或者WFST技术再次融合预训练语言模型进一步提升模型识别率。实验结果表明，在VIVOS数据集上，相比基线模型，在训练阶段融合语言模型可以将流式越南语语音识别模型的词错率提升2.45%；在解码阶段使用Shallow Fusion或WFST再次融合语言模型，还可以将模型词错率分别提升1.35%和4.75%。

关键词： 流式语音识别；越南语；语言模型；预训练；端到端模型

Streaming Vietnamese Speech Recognition Based on Fusing External Vietnamese Language Knowledge

Junqiang Wang[1,2], Zhengtao Yu*[1,2], Ling Dong[1,2], Shengxiang Gao[1,2], Wenjun Wang[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
814220330@qq.com, ztyu@hotmail.com, 46761956@qq.com
gaoshengxiang.yn@foxmail.com, 175360805@qq.com

Abstract

Vietnamese is known as a low-resource language with few available corpora, and for end-to-end speech streaming recognition, it's difficult to fuse external knowledge from large-scale text corpora, which limits the performance. Given that, we proposed a method that fuses a pre-trained Vietnamese transformer language model into the streaming Vietnamese speech recognition model at the training stage both using the token level of Syllable. At the training stage, a novel loss function called L_{AED-LM} was introduced to optimize the parameters, learning the language knowledge by minimizing the difference between the output of pre-trained Vietnamese transformer language model and decoder. At the inference stage, we applied the Shallow Fusion or WFST technology to enhance the performance further. Experiments on the Vivos dataset show that, compared with the baseline model, the WER of streaming Vietnamese speech recognition can be improved by 2.45% by fusing the pre-trained language model during training; At the inference stage, Shallow Fusion or WFST improved the WER by 1.35% and 4.75% respectively.

Keywords: streaming speech recognition , Vietnamese , language model , pre-trained , end-to-end model

*余正涛(通信作者):ztyu@hotmail.com
国家自然科学基金（61732005, U21B2027, 61972186）；云南高新技术产业发展项目（201606）；云南省重大科技专项计划（202103AA080015, 202002AD080001-5）；云南省基础研究计划（202001AS070014）；云南省学术和技术带头人后备人才（202105AC160018）

1 引言

越南作为“一带一路”沿线的重要国家，中越沟通合作交流越来越频繁。越南语语音识别可以提高中越双方沟通交流效率，开展越南语语音识别研究对促进中越贸易、政策沟通以及企业合作具有重要意义。

近几年来，端到端模型在语音识别任务中受到了广泛关注。端到端模型将传统语音识别系统的声学模型、发音词典和语言模型融合成一个模型，极大地减少了语音识别模型的训练流程和复杂性。目前，主流的端到端模型有Connectionist temporal classification (CTC)(Graves et al., 2006)、Recurrent Neural Network Transducer (RNN-T)(Rao et al., 2017)、Attention Based Encoder-Decoder(AED) (Chorowski et al., 2014; Chorowski et al., 2015; Chan et al., 2015)和Hybrid CTC/Attention(Kim et al., 2017; Hori et al., 2017)等模型。虽然这些端到端模型在多资源语种上取得了很好的效果，但是在训练过程中端到端模型难以利用外部大量文本中的语言知识(Gulcehre et al., 2015)。因此，一些研究者针对此问题，提出了一些在训练阶段将语言模型融合到语音识别模型的方法(Deep Fusion(Gulcehre et al., 2015)、Cold Fusion(Sriram et al., 2018)和Component Fusion(Shan et al., 2019))。实验结果表明，在训练阶段，将预训练语言模型融合到语音识别模型可以有效地帮助语音识别模型学习到一些语言知识，并弥补端到端模型在训练过程中难以利用外部大量文本语言知识的缺陷，同时提升语音识别模型的识别准确率。但是Deep Fusion、Cold Fusion和Component Fusion方法都需要语音识别模型增加额外的参数来融合语言模型，因此导致语音识别模型参数量增加的问题。并且这三种方法都采用RNN作为语言模型，在训练过程中不能像Transformer(Vaswani et al., 2017)模型一样并行训练，在一定程度上增加了语音识别模型的训练时间。

在越南语标注语音语料缺失的情况下，越南语语音识别模型的性能难以提升。相比获取越南语语音语料，获取越南语文本语料要容易得多，但目前的越南语语音识别模型并没有利用外部大量越南语文本中的语言知识来提升语音识别模型的识别率。同时，国内外针对流式端到端越南语语音识别模型的研究还很有限，大部分流式越南语语音识别模型仅在解码阶段使用了Shallow Fusion(Chorowski and Jaitly, 2016)方法融合语言模型，并没有在训练阶段融合语言模型的方法研究。

因此，本文针对以上问题，提出了一种将预训练Transformer越南语语言模型在训练阶段融合到流式越南语语音识别模型的方法。在训练阶段，仅通过预训练越南语语言模型和解码器的输出计算一个新的损失函数L_{AED-LM}，不会额外增加模型参数，还可以帮助流式越南语语音识别模型在训练过程中学习到一些越南语语言知识从而优化其模型参数；在解码阶段，使用传统的Shallow Fusion或者WFST(Wang et al., 2021)技术再次融合语言模型来纠正流式越南语语音识别模型的识别结果进一步提升模型性能。

本文的贡献如下：

(1)在训练阶段，将预训练越南语语言模型融合到流式越南语语音识别模型中，提升了流式越南语语音识别模型的识别率。

(2)在解码阶段，使用Shallow Fusion和WFST方法再次融合越南语语言模型进一步提升了流式越南语语音识别模型的识别率。

(3)本文在开源越南语数据集VIVOS上进行实验，在解码阶段不融合语言模型的情况下，相比基线模型，将流式越南语语音识别模型的词错率从31.03%降到了28.58%。在解码阶段，使用Shallow Fusion融合方法融合Transformer语言模型，词错率能提升到27.23%；使用WFST方法融合3元语言模型，词错率能提升到23.83%。

2 相关工作

近年来，虽然端到端语音识别受到了广泛关注，但目前针对越南语语音识别研究还比较少。Nguyen等人(2018)构建了500小时的越南语数据集并使用TDNN和BLSTM神经网络构建声学模型，在解码阶段融合了4元语言模型。为了提升模型性能，它将4元语言模型替换为RNN语言模型，在3小时测试集数据上进行测试，词错率达到6.9%。Nguyen和Huy(2019)使用CTC损失函数将TDNN和BLSTM模型结合一起联合训练越南语语音识别模型，在FPT测试数据集上，词错率达到14.41%。刘佳文(2020)提出了一种基于Transformer模型的越南语语音识模型，在VIVOS数据集上，字符错率达到40.4%。ESPNET(2021)基于不同的Transducer(Graves, 2012)模型在VIVOS数据集上做了不同实验，RNN-T词错率达到36.6%，Conformer(Gulati et al., 2020)/RNN-T词错率达到26%。为了提升模型识别率，这些模型都在解码阶段融合了语言模型，但在解码阶段融合语言模型只能影响模型的识别结果，并不能利用语言模型来优化语音识别模型的参数。因此，本文在流式越南语语音识别模型的训练阶段和解码阶段都融合了语言模型。在训练阶段融合语言模型可以帮助流式语音识别模型学习一些越南语语言知识优化其模型参数；在解码阶段融合语言模型可以帮助流式越南语语音识别模型纠正识别错误进一步提升其模型的识别率。

3 融合外部语言知识的流式越南语语音识别

为了解决流式越南语语音识别模型难以学习到大量外部语言知识的问题，本文使用Hybrid CTC/Attention模型架构作为流式越南语语音识别模型的基线模型，在此基础上使用越南语单语文本语料预训练Transformer-xl(Dai et al., 2019)语言模型，并将其在训练阶段融合到流式越南语语音识别模型中，以提升模型识别效果，具体方法如下所述。

3.1 模型架构

语音识别Hybrid CTC/Attention模型架构由三个部分组成：共享编码器、CTC解码器和Attention-Based解码器。共享编码器由多层Transformer编码器构成；CTC解码器由一个线性层、log softmax层构成；Attention-Based解码器由多层Transformer解码器构成。在Hybrid CTC/Attention模型架构的基础上，本文将预训练越南语Transformer-xl语言模型与Hybrid CTC/Attention模型中的Transformer解码器进行了融合，如图1所示。

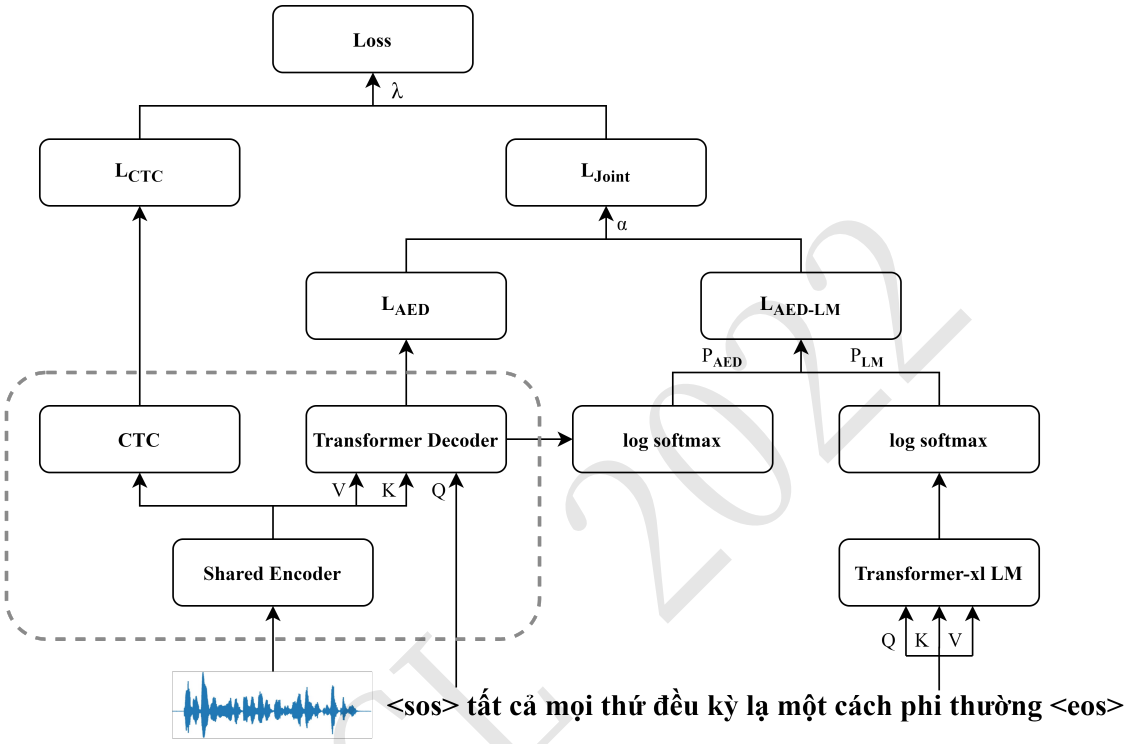


图 1. Hybrid CTC/attention模型融合越南语语言模型架构图

音频特征X=(x_t|t= 1,2,...,T)经过共享编码器编码生成高维音频特征向量H=(h_t|t= 1,2,...,T)，然后将音频特征向量H和标签送入CTC解码器和Transformer解码器进行解码。利用Transformer解码器和预训练Transformer-xl语言模型的输出经过log softmax函数计算出两个输出概率P_{AED}和P_{LM}。使用P_{AED}概率计算Transformer解码器的损失函数L_{AED}；使用P_{AED}和P_{LM}两个输出概率计算越南语Transformer-xl语言模型和Transformer解码器的融合损失函数L_{AED-LM}。再通过语言模型融合超参数$\alpha(0\leq\alpha\leq1)$联合P_{AED}和L_{AED-LM}生成L_{Joint}损失函数，最终经过CTC权重超参数λ联合L_{CTC}和L_{Joint}损失函数训练流式越南语语音识别模型。

3.2 越南语语言模型

在构建语言模型时，虽然越南语是一种以单音节为主的语言(Haudricourt, 2010; Alves, 2006; Hwa-Froelich et al., 2002; Thompson, 1991)，但是少部分词会包含多个音节。在包含多音节词的句子中，上下文音节之间的依赖长度较长会导致模型出现长期依赖丢失的问题，并且在模型编码越南语音节时，句子长度过长会使音节丢失在句子中的位置信息。因此，本文使用Transformer-xl作为越南语语言模型，可以解决越南语音节长期依赖和位置编码丢失的问题，从而使越南语语言模型更好地表征越南语语言知识。在融合过程中能让语音识别模型从越南语语言模型更好地学习到越南语语言知识，提升语音识别模型的识别率。

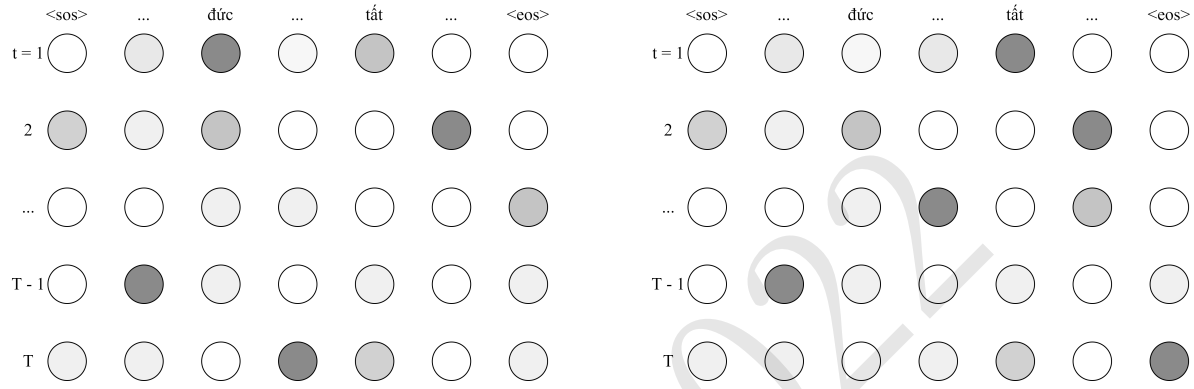
本文使用33万句越南语文本作为训练语料，在训练阶段，Transformer-xl使用片段递归的方法将当前隐藏层状态作为Q，将之前的隐藏层状态与当前隐藏层状态拼接后作为K和V，最终使用Q、K、V来计算attention的输出。这种方法使得越南语语言模型具有更强的长期依赖能力，从而更好地表征越南语语

言知识。但由于语音识别任务句子之间没有上下文关系，因此在解码阶段，本文并没有使用片段递归的方法来保存之前句子的隐藏层状态，而是直接将目标句子作为Transformer-xl的输入经过解码后得到预测每一个音节的概率分布，并在语音识别模型中使用此概率来融合越南语语言知识。

3.3 越南语语言模型融合方法

在训练阶段将语言模型和流式越南语语音识别模型融合，本文使用KL散度来计算Transformer解码器和越南语Transformer-xl语言模型输出之间的融合损失函数。其目的是为了让Transformer解码器的输出概率分布向越南语语言模型的输出概率分布靠近，从而帮助语音识别模型从越南语语言模型中学习到越南语语言知识。具体融合方法如下所述。

假设，目标序列长度为L，0表示开始符<sos>，L表示结束符<eos>，P_{AED}=P($Y_{1\sim L}$|H,$Y_{0\sim L-1}$)表示Transformer解码器在给定共享编码器输出特征向量H和输入目标序列$Y_{0\sim L-1}$的条件下，预测出目标序列$Y_{1\sim L}$的输出概率分布；P_{LM}=P($Y_{1\sim L}$|$Y_{0\sim L-1}$)表示越南语语言模型在给定输入目标序列$Y_{0\sim L-1}$的情况下输出目标序列$Y_{1\sim L}$的输出概率分布。本文将越南语语言模型输出的P_{LM}作为真实分布，Transformer解码器输出的P_{AED}作为理论数据分布，如图2所示。



(a) Transformer解码器输出(P_{AED})　　(b) 越南语语言模型输出(P_{LM})

图 2. 越南语语言模型融合方法。每一个节点代表Transformer解码器和越南语语言模型在给定输入后输出对应越南语音节的概率值（颜色深浅代表概率值大小）。将(a)作为理论数据分布，(b)作为真实分布。

然后使用KL散度来计算Transformer解码器与越南语Transformer-xl语言模型的融合损失函数，如公式(1)所示：

$$D_{KL}(P_{LM}||P_{AED}) = \sum_{i=0}^{L} P_{LM} \log \frac{P_{LM}}{P_{AED}} \tag{1}$$

最后令L_{AED-LM}损失函数等于$D_{KL}(P_{LM}||P_{AED})$并使其最小化。

在融合过后，本文引入了一个语言模型融合超参数$\alpha(0\leq\alpha\leq1)$，用来调节Transformer解码器L_{AED}损失函数和L_{AED-LM}损失函数。令联合损失函数为L_{Joint}，其计算过程如下：

$$L_{Joint} = (1-\alpha)L_{AED} + \alpha L_{AED-LM} \tag{2}$$

最终训练的联合损失函数与传统Hybrid CTC/Attention模型损失函数类似，不同的是将Attention损失函数修改为Transformer解码器融合了越南语语言模型的损失函数。如公式(3)所示，其中$\lambda(0\leq\lambda\leq1)$参数和传统Hybrid CTC/attention模型一致。

$$Loss = \lambda L_{CTC} + (1-\lambda)L_{Joint} \tag{3}$$

流式越南语Hybrid CTC/Attention模型最后使用Loss损失函数来训练。这种训练方式可以让融合损失函数L_{AED-LM}在训练阶段同时优化CTC解码器和Transformer解码器的参数，帮助CTC和Transformer解码器联合学习到越南语语言知识从而提升流式越南语Hybrid CTC/Attention模型的识别率。

3.4 解码

传统Hybrid CTC/Attention模型使用自回归的方式解码，解码速度慢，所以本文采用了two-pass(Sainath et al., 2019)模型架构中的二次评分模式实现非自回归方式解码加快解码速度。首先，

使用CTC解码器生成N个最好的序列，然后再将这N个序列和编码器的输出送入到Transformer解码器以Teacher-Forcing的方式解码，最终选取评分最高的序列输出。

由于在训练阶段融合语言模型后，可以在解码阶段选择性地使用传统语言模型再次融方法进一步提升模型识别率，因此，本文在解码阶段分别使用了Shallow Fusion和WFST两种方法再次融合语言模型进一步提升模型识别率。

4 实验

4.1 数据集

本文使用开源越南语数据集VIVOS进行实验。VIVOS数据集的训练数据集由46个说话人在安静环境下录制了15个小时的音频，其中包含11660个句子；测试集由19个说话人在相同环境下录制了45分钟的音频，其中包含了760个句子。所有音频都是以16kHz的采样率采样，并且以小端WAV格式存储。

本文爬取了33万句越南语文本语料来训练语言模型。在语料的预处理过程中，去除了标点符号、数字。由于部分网站使用不同的越南语编码格式，在构建语料库时，需要将越南语文本语料统一编码为utf-8编码格式，确保在识别过程中不会因为编码问题而导致同一个音节识别错误和同一个音节在词典中出现多次的问题。

4.2 参数设置

本文使用80维log梅尔滤波器组(FBANK)在窗口大小为25ms帧移为10ms的条件下对VIVOS数据集的音频文件进行特征提取；使用SpecAugment(Park et al., 2019)技术对提取的音频特征进行数据增强；在数据前处理阶段，使用卷积核大小为3*3，步长为2的两个卷积层对音频数据进行做下采样处理；使用12层带有4头注意力的Transformer作为编码器，令编码器的输出维度为256；为了使编码器支持流式编码，使用固定chunk和动态chunk(Zhang et al., 2020)两种方式进行编码，分别对比了不同编码方式对流式越南语语音模型的影响；使用6层带有4头注意力的Transformer联合CTC作为解码器；为了防止过拟合，在解码器和编码器的每一层都设置了dropout，并且将dropout比率设置为0.1；使用Adam优化器，学习率设置为0.002，学习率预热设置为25000步；使用标签平滑技术来计算损失函数，标签平滑率设置为0.1。最后，选取30个最好的模型进行参数平均得到最终模型，提升模型的泛化能力。在解码阶段，使用前缀束搜索算法对CTC解码器的输出进行搜索，束搜索宽度大小设置为16，束搜索产生束宽度大小个first-pass结果，然后再将这些结果送入Transformer解码器中进行二次评分。在二次评分阶段使用CTC权重超参数λ来控制CTC解码器的输出权重和Transformer解码器输出权重。

本文使用Transformer-xl作为语言模型，它由12层带有4头注意力的Transformer编码器构成并引入了相对位置编码和片段递归机制。本文使用自己构建的33万句越南语单语文本语料训练越南语语言模型并采用Adam优化器优化模型参数并设置学习率为0.00025，学习率预热为20000步。

模型词表中共9078个词，其中包含4个特殊标签，<blank>表示CTC的空标签，<unk>表示未登录词，<sos>和<eos>表示句子的开始和结束。最终CTC解码器和Transformer解码器的输出维度为9078。本文所有实验在一张NVIDIA Tesla T4上完成。

4.3 实验结果及分析

4.3.1 不同chunk方法训练对流式越南语语音识别模型性能的影响

本文使用Hybrid CTC/Attention模型架构作为越南语语音识别的基线模型。为了使模型能够流式输出，本文修改了Hybrid CTC/Attention模型的编码方式，使用不同chunk大小和动态chunk的编码方式训练模型，并对比了不同编码方式对模型性能的影响。

在对比不同chunk方法对模型的影响时，本文将CTC权重超参数λ固定设置为0.3，语言模型融合权重超参数α固定设置为0，然后将chunk大小分别设置为8/16/动态chunk进行对比实验，实验结果见表1。

表 1. 不同chunk方法对模型性能的影响

chunk大小	解码chunk大小	词错率(WER%)
8	8	36.69
16	16	34.48
动态chunk	16	31.03

根据实验结果数据显示，当chunk大小设置为8/16时，流式越南语语音识别模型的词错率分别为36.39%和34.48%。当chunk大小设置为动态chunk时，流式越南语语音识别模型识别性能达到最佳31.03%。

对于越南语语音识别任务而言，动态chunk方式训练的模型效果明显优于以固定chunk大小训练的模型识别性能。主要是因为越南语自身是一种以单音节为主的语言，但有一些词包含多个音节，因此使用动态chunk的方式编码更符合越南语由不同音节个数构成词的特点从而使得模型识别性能更佳。

在接下来的实验中，我们均采用动态chunk编码方式训练的Hybrid CTC/Attention语音识别模型作为流式越南语语音识别模型的基线模型。

4.3.2 融合语言模型对流式越南语语音识别模型性能的影响

为了验证本文提出的方法对流式越南语语音识别模型性能有提升，本文将流式越南语语音识别模型的CTC超参数λ和语言模型融合超参数α分别设置为不同值，对比在训练阶段融合语言模型前后和不同超参数对流式越南语语音识别模型性能的影响，实验结果见表2和表3。

表 2. 当CTC权重为0.3时，融合语言模型权重α对流式越南语语音识别模型的影响

语言模型融合权重α	词错率(WER%)
0(baseline)	**31.03**
0.3	**28.58**
0.5	33.22
0.7	29.15

表 3. 当CTC权重为0.5时，融合语言模型权重α对流式越南语语音识别模型的影响

语言模型融合权重α	词错率(WER%)
0(baseline)	**30.30**
0.3	29.41
0.5	29.54
0.7	29.60

实验结果数据显示，当CTC权重参数设置为0.3时，在不融合语言模型(融合语言模型权重参数α为0)的情况下，流式越南语语音识别模型词错率为31.03%(baseline)。当以0.3的权重融合语言模型时，性能有明显提升，词错率达到了28.58%。但当语言模型融合权重设置为0.5时，性能相比基线模型有一定下降。当语言模型融合权重设置为0.7时，性能相比基线模型又有一定提升，达到29.15%。当CTC权重参数设置为0.5时，在不融合语言模型的情况下，流式越南语语音识别模型词错率为30.30%(baseline)。当语言模型融合权重参数分别设置为0.3/0.5/0.7时，流式越南语语音识别模型的识别性能相比基线模型都有所提升，但语言模型融合权重参数对流式越南语语音识别模型的识别词错率影响不怎么明显，词错率保持在29%左右。

当CTC权重参数为0.3，语言模型融合权重参数为0.5时，性能相比基线模型有一定下降。主要是因为当语言模型融合权重设置为0.5时，解码器和语言模型的输出比重相同，语音识别模型不能抉择解码器和越南语语言模型输出的重要性，从而导致模型混乱，识别性能下降。但是当语言模型融合权重设置为其他值时，性能相比基线模型都有一定提升。这说明了流式越南语语音识别模型可以从越南语语言模型中学习到越南语语言知识从而优化其模型参数，达到识别性能提升的效果。

4.3.3 识别结果示例分析

本文将流式越南语语音识别模型的CTC权重参数设置为0.3，语言模型融合权重参数分别设置为0/0.3进行了对比实验，并通过对测试集识别结果示例的分析来说明融合越南语语言模型可以提升模型的识别效果。实验结果见4。

表 4. 识别结果示例分析

识别结果α	词错率(WER%)
tất cả mọi thứ đều kỳ lạ một cách phi thường (原标签)	-
tất cả mọi thứ đều kỳ **lạ** một cách phi thường (融合语言模型识别出的标签)	**0**
đức cả mọi thứ đều kỳ **là** một cách phi thường (未融合语言模型识别出的标签)	18.18

实验结果表明，融合了语言模型的流式越南语语音识别模型识别结果完全正确，而未融合语言模型的流式越南语语音识别模型识别结果词错率为18.18%。

未融合语言模型的流式越南语语音识别模型识别错了两个音节đức和là，主要原因是đức和tất、là和lạ音节的发音非常相似，提取出来的语音特征也非常接近，从而导致语音识别模型不能辨别。而融合了语言模型的流式越南语语音识别模型可以从语言模型中学习到tất cả和kỳ lạ可以组

成一个词，而đức cả和kỳ là不能组成词，从而tất cả和kỳ lạ的输出概率高于đức cả和kỳ là，因此语音识别模型选择tất cả和kỳ lạ输出。

实验结果表明，在训练阶段融合语言模型确实可以优化流式越南语语音识别模型参数从而纠正一些将越南语音节识别错误的情况。

4.3.4 二次融合语言模型对流式越南语语音识别模型性能的影响

当流式越南语语音识别模型的CTC权重参数设置为0.3，语言模型融合权重参数设置为0.3时性能最佳，因此在解码阶段二次融合语言模型也使用此参数配置。为了验证二次融合语言模型对流式越南语语音识别模型识别率的影响。本文在解码阶段使用Shallow Fusion和WFST方法分别对Transformer-xl语言模型和3元语言模型进行融合。实验结果如表5和表6所示。

表 5. 使用Shallow Fusion融合方法对流式越南语语音识别模型性能的影响

模型	词错率(WER%)
Hybrid CTC/Attention(baseline)	**31.03**
Hybrid CTC/Attention +训练阶段融合语言模型	28.58
Hybrid CTC/Attention + Shallow Fusion	29.83
Hybrid CTC/Attention + 训练阶段融合语言模型+ Shallow Fusion	**27.23**

表 6. 使用WFST方法对流式越南语语音识别模型性能的影响

模型	词错率(WER%)
Hybrid CTC/Attention(baseline)	**31.03**
Hybrid CTC/Attention +训练阶段融合语言模型	28.58
Hybrid CTC/Attention + WFST	24.32
Hybrid CTC/Attention + 训练阶段融合语言模型+ WFST	**23.83**

实验数据结果显示，在训练阶段融合语言模型后，在解码阶段使用Shallow Fusion方法再次融合Transformer-xl语言模型还可以将模型的识别率提升1.35%；在训练阶段融合语言模型后，在解码阶段使用WFST融合3元语言模型，性能达到最佳23.83%。

虽然使用Shallow Fusion或WFST方法进行解码，模型识别率会有所差距，但实验数据结果显示，在训练阶段融合语言模型后，在解码阶段再次融合语言模型确实可以进一步提升流式越南语语音识别模型的识别率。同时，在训练阶段和解码阶段都融合语言模型，模型的识别率要明显高于在解码阶段单独融合语言模型的识别率。

4.3.5 和其他模型的性能比较

本次实验对比了本文使用的流式模型和ESPNET使用RNN-T、Conformer/RNN-T模型在VIVOS测试数据集上的结果。实验结果如表7所示。

表 7. 和其他模型识别效果对比

模型	词错率(WER%)
RNN-T	36.6
Conformer/RNN-T	26.0
Hybrid CTC/Attention + 训练阶段融合语言模型+ WFST	**23.83**

实验结果数据显示，本文使用的流式模型词错率达到23.83%，RNN-T和Conformer/RNN-T模型的词错率分别为36.6%和26.0%。

本文在训练阶段融合语言模型后，再使用WFST在解码阶段融合3元语言模型的识别率达到最佳。其主要原因是本文同时在训练阶段和解码阶段都融合了语言模型。在训练阶段融合语言模型可以优化模型的参数；在解码阶段融合语言模型可以纠正语音识别模型识别结果。而ESPNET仅在解码阶段融合了语言模型，只影响了语音识别模型的识别结果，并不能优化模型的参数。

5 总结

由于越南语标注语音语料难以获取，流式越南语语音识别模型难以在训练阶段利用外部文本语言知识的问题，本文提出了一种在训练阶段将预训练越南语Transformer-xl语言模型融入到流式越南

语Hybrid CTC/Attention模型的方法。实验表明，这种融合方法可以提升流式越南语语音识别模型的识别率并弥模型在训练过程中难以学习外部语言知识的缺陷。另外，在解码阶段再使用一些传统语言模型融合方法还可以进一步提升语音识别模型的识别率。

参考文献

Mark Alves. 2006. Linguistic research on the origins of the vietnamese language: An overview. *Journal of Vietnamese Studies*, 1(1-2):104–130.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Jan Chorowski and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *ArXiv*, abs/1211.3711.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, and Yuekai Zhang. 2021. Recent developments on espnet toolkit boosted by conformer. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878.

André-Georges Haudricourt. 2010. The origin of the peculiarities of the vietnamese alphabet. *Mon-Khmer Studies*, 39:89–104.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. *arXiv preprint arXiv:1706.02737*.

Deborah Hwa-Froelich, Barbara W Hodson, and Harold T Edwards. 2002. Characteristics of vietnamese phonology.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

Quoc Bao Nguyen, Van Tuan Mai, Quang Trung Le, Ba Quyen Dam, and Van Hai Do. 2018. Development of a vietnamese large vocabulary continuous speech recognition system under noisy conditions. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, pages 222–226.

Huy Nguyen. 2019. An end-to-end model for vietnamese speech recognition. pages 1–6, 03.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199. IEEE.

Tara N Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, et al. 2019. Two-pass end-to-end speech recognition. *arXiv preprint arXiv:1908.10992*.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Component fusion: Learning replaceable language model component for end-to-end speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5361–5635. IEEE.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. Cold fusion: Training seq2seq models together with language models. pages 387–391, 09.

Laurence C Thompson. 1991. A vietnamese reference grammar (revised edition).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhichao Wang, Wenwen Yang, Pan Zhou, and Wei Chen. 2021. Wnars: Wfst based non-autoregressive streaming end-to-end speech recognition. *arXiv preprint arXiv:2104.03587*.

Binbin Zhang, Di Wu, Zhuoyuan Yao, Xiong Wang, Fan Yu, Chao Yang, Liyong Guo, Yaguang Hu, Lei Xie, and Xin Lei. 2020. Unified streaming and non-streaming two-pass end-to-end model for speech recognition. *arXiv preprint arXiv:2012.05481*.

刘佳文, 屈丹, 杨绪魁, 张昊, and 唐君. 2020. 基于transformer的越南语连续语音识别. 信息工程大学学报, 21(2):129–133,152, 4.