基于字节对编码的端到端藏语语音识别研究

蔡郁青[1,2]，王超[2,3]，仁增多杰[*1,2]，朱宇雷[1,2]，张瑾[1,2]，尼玛扎西[*1,2]
1.西藏大学，信息科学技术学院，拉萨，850000
2.藏语智能信息处理及应用国家重点实验室，拉萨，850000
3.青海师范大学，计算机学院，西宁，810016
yuqingcai@stu.utibet.edu.cn,wangchao980328@163.com,rzdj@utibet.edu.cn,
zhuyulei0808@163.com,jinzhang@stu.utibet.edu.cn,niqiongda@163.com

摘要

针对藏语端到端语音识别研究中存在的建模单元不统一和识别效果不理想的问题，本文提出了一种BPE-Conformer-CTC/Attention端到端藏语语音识别方法。首先，该方法采用了字节对编码算法进行语音建模，通过反复合并出现频率最高的字符对，将文本分割成易于管理、有意义的单元，平衡建模单元的粒度，从而解决藏语语音识别中建模单元不统一的问题。其次，使用了Conformer编码器，有效地融合了音频序列的全局和局部依赖关系，从而增强了模型的表征能力。最后，通过CTC/Attention联合解码策略，加速了对齐和解码过程，进而提高了识别效果的准确性和效率。在开源数据集XBMU-AMDO31和TIBMD@MUC上的实验结果表明，所提出的BPE-Conformer-CTC/Attention模型分别取得了9.0%和4.6%的词错误率，相较于基线模型Transformer-CTC/Attention，词错误率分别相对降低了14.2%和30.3%。该研究方法为藏语端到端语音识别任务提供了一种有效的解决方案。

关键词： 藏语语音识别；端到端；字节对编码；安多方言

End-to-End Tibetan Speech Recognition Study Based on Byte Pair Coding

Yuqing Cai[1,2],Chao Wang[2,3],Renzeng Duojie[*1,2],Yulei Zhu[1,2],Jin Zhang[1,2],Nyima Tashi[*1,2]
1.Tibet University,School of Information Science and Technology,Lhasa,850000
2.The State Key Laboratory of Tibetan Intelligent Information Processing and Application,Lhasa,850000
3.Qinghai Normal University,The College of Computer,Xining,810016
yuqingcai@stu.utibet.edu.cn,wangchao980328@163.com,rzdj@utibet.edu.cn,
zhuyulei0808@163.com,jinzhang@stu.utibet.edu.cn,niqiongda@163.com

Abstract

Aiming at the problems of inconsistent modeling units and unsatisfactory recognition results in Tibetan end-to-end speech recognition research, this paper proposes a BPE-Conformer-CTC/Attention end-to-end Tibetan speech recognition method. Firstly, the method adopts the byte-pair encoding algorithm for speech modeling, which solves the problem of inconsistent modeling units in Tibetan speech recognition by repeatedly merging the most frequent character pairs, segmenting the text into manageable and meaningful units, and balancing the granularity of modeling units. Secondly, the Conformer encoder is utilized to effectively fuse the global and local dependencies of audio

*尼玛扎西（通信作者）：niqiongda@163.com，仁增多杰（通信作者）：rzdj@utibet.edu.cn
基金项目：新一代人工智能国家科技重大专项(2022ZD0116100)，西藏大学研究生“高水平人才培养计划”项目(2022-GSP-S096)

sequences, thus enhancing the model's representational capability. Finally, the alignment and decoding process is accelerated by a joint CTC/Attention decoding strategy, which in turn improves the accuracy and efficiency of the recognition results. Experimental results on the open-source datasets XBMU-AMDO31 and TIBMD@MUC show that the proposed BPE-Conformer-CTC/Attention model achieves a word error rate of 9.0% and 4.6%, respectively, and compared to the baseline Transformer-CTC/Attention model, the word error rate is reduced by 14.2% and 30.3%, respectively. This proposed approach provides an effective solution for the Tibetan end-to-end speech recognition task.

Keywords: Tibetan speech recognition , End to end , Byte pair encoding , Amdo dialect

1 引言

近年来，端到端的自动语音识别(ASR)方法越来越受到关注 (Amodei et al., 2015; Chorowski et al., 2015; Huang et al., 2020)。与传统语音识别方法相比，端到端ASR系统将发音词典、声学模型和语言模型等模块集成到单个框架中进行联合优化，大幅降低了训练和推理的复杂度。

藏语端到端语音识别起步较晚。(黄晓辉 and 李京, 2018)提出结合RNN和CTC的方法建立藏语声学模型并进行端到端训练，以提高准确性和性能。(王松, 2019)采用LSTM-CTC端到端方法进行藏语拉萨话识别。随着Transformer(Vaswani et al., 2017)在机器翻译任务中的广泛应用，(Dong et al., 2018)首次将其引入语音识别领域，提出了Speech-Transformer模型。(Yang et al., 2020)提出了基于Transformer的藏语语音识别方法，使藏语识别错误率降至30%以下。(高耀荣 and 边巴旺堆, 2023)构建了Transformer-CTC/Attention模型，为后续藏语语音识别研究奠定了基础。

尽管端到端藏语语音识别取得了一定进展，但仍存在一些问题。相较于主流语言，藏语端到端识别效果较差。处理具有复杂构字特点的藏语时，建模单元存在不统一和重复实验的问题。同时，未知词汇的处理问题也一直制约着藏语识别性能的提升。

针对上述问题，本文以安多方言为研究对象，结合CTC(Graves, 2006)自动对齐和Attention上下文建模的优点，探索了基于CTC/Attention联合解码的Transformer语音识别模型。为解决Transformer捕捉局部信息能力有限的问题，提出使用Conformer作为编码器，构建了Conformer-CTC/Attention模型。在建模单元选择上，采用BPE算法进行语音建模，以解决构件、字丁和音节等单元的争议，同时通过BPE的灵活组合一定程度上能解决未知词汇问题。

2 建模单元的选择

对于语音识别模型来说，建模单元的选择是影响模型识别性能的重要因素之一。不同的建模单元所需的训练数据规模和模型训练难度存在差异，最终会对模型的识别精度产生影响。在藏语语音识别研究领域，由于研究者使用的训练语料规模不同，目前采用何种粒度的建模单元尚未统一。与此同时，语音识别过程中未知词汇的问题也一直没有得到有效解决。

2.1 藏语文字构成

藏语作为汉藏语系的一种语言，包含卫藏、康巴和安多三大方言。尽管不同方言使用同一个藏语文字系统，但研究藏字的构字规则对于覆盖所有方言而言意义重大。藏语是一种拼音文字，由30个辅音和4个元音构成。一个藏字通常对应一个音节，由字丁横向叠加而成，而字丁又由构件纵向叠加而成。因此，藏语的建模单元可分为构件、字丁和音节三种，具体示例如图1所示。

2.2 基于字节对编码的语音建模

在端到端语音识别系统中，建模单元粒度的选择是一个需要权衡的问题。采用较小粒度的建模单元(如构件)虽然能够覆盖所有词汇、降低标注成本，但也会导致输出序列过长、模型规

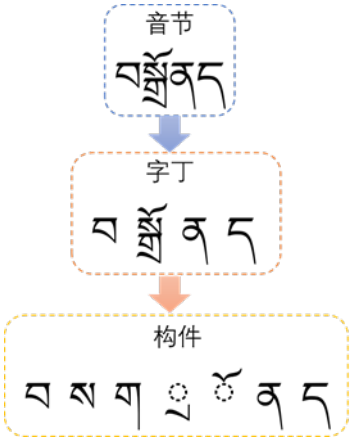

图 1. 藏字构成示例

模庞大、参数量剧增，从而加大了模型优化的难度，进而影响识别精度。相反，选择较大粒度的建模单元(如字丁和音节)则可以有效控制输出序列长度和模型大小，但存在词典覆盖范围有限、标注成本高昂的弊端。因此，选择适当粒度的建模单元对于平衡上述利弊至关重要，不仅能够有效控制输出序列长度、降低建模复杂度，还可以在一定程度上解决未登录词问题，提高识别性能。

针对这一问题，本文提出了使用字节对编码(Byte Pair Encoding，BPE)(Shibata et al., 1999)算法进行藏语语音建模。BPE是一种数据压缩算法，可以通过一个有限的词表来解决所有词的分词问题，同时能够较好地平衡未登录词的问题。BPE算法首先将每个藏字拆分成单个构件，然后依次用另一个字符替换频率最高的一对字符，直至达到预设的终止条件。根据训练文本，我们可以获得藏文句中各音节的BPE拆分结果，如图2所示，其中“_”表示该子词为词前缀。

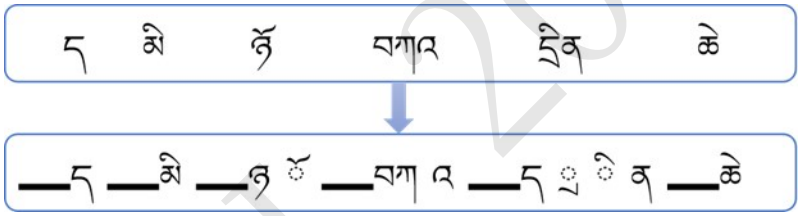

图 2. 藏文句BPE拆分示例

3 模型构建

针对藏语端到端语音识别任务，本文提出了一种BPE-Conformer-CTC/Attention混合架构。如图3所示，该架构由三个主要部分组成：Conformer编码器、CTC解码分支和Attention解码分支。首先，音频特征被Conformer编码器编码为高级语音特征序列，然后该序列并行传递给后续的CTC和Attention两个解码分支进行处理。在训练过程中，两个解码分支的损失函数按照一定权重线性组合，实现多任务联合学习。而在解码阶段，两个分支的输出概率则通过加权束搜索算法进行融合，得到最终的识别结果。

3.1 编码器层

本文采用Conformer(Gulati et al., 2020)作为编码器，其结构如图3所示。Conformer编码器主要由四个子模块组成，包括第一个前馈神经网络模型、多头自注意力模块、卷积模块和第二个前馈神经网络模块。不同于Transformer，Conformer中的多头注意力模块使用了相对位置嵌入编码，并采用了Swish激活函数。另一方面，卷积模块以逐点卷积(Pointwise Convolution)和门控线性单元(GLU)为起点，然后进行一维深度卷积和批次归一化(Batch norm)，这一设计有助于有效地训练深度模型。具体地，对于第k个Conformer模块的输入序列x_k，其输出y_k 的计算流程如下所示：

$$x_{FFN1} = x_k + 1/2FFN(x_k) \tag{1}$$

$$x_{MHSA} = x_{FFN1} + MHSA(x_{FFN1}) \tag{2}$$

$$x_{Conv} = x_{MHSA} + Cona(x_{MHSA}) \tag{3}$$

$$y_k = LayerNorm(x_{Conv} + 1/2FFN(x_{Conv})) \tag{4}$$

其中FFN表示前馈神经网络模块，$MHSA$表示多头自注意力模块，$Conv$表示卷积模块，每个模块间都使用残差连接。

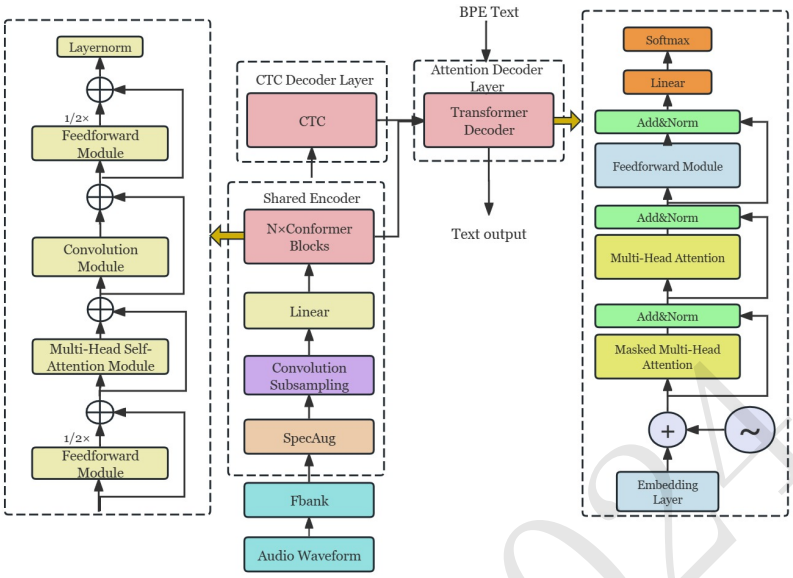

图 3. BPE-Conformer-CTC/Attention架构图

3.2 解码器层

3.2.1 Transformer解码器

Transformer解码器通过自注意力机制直接建模输出序列中元素之间的依赖关系，而不需要递归计算，并且Transformer架构中的多头注意力机制和位置编码有助于捕获输入和输出序列中的长依赖关系。因此，本文将Transformer解码器部分直接用作端到端语音识别模型解码器的一部分。具体应用的多头注意力机制如图4所示。

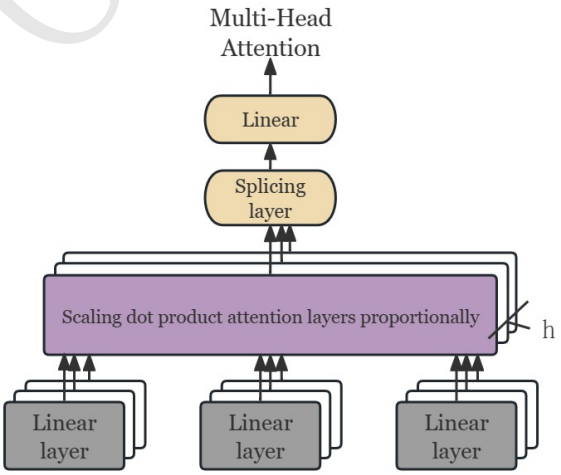

图 4. 缩放点积注意力流程图

注意力机制是基于缩放的点积注意力方程，其定义如下：

$$Attention(Q,K,V) = Softmaxa((QK^T)/\sqrt{(d_k)})V \tag{5}$$

3.2.2 连接时序分类

CTC通过对目标序列和原始输入序列的对齐分布进行建模，消除了在单个声学特征建模单元和文本序列建模单元之间手动对齐的繁琐操作，从而成功实现了端到端声学模型的训练。该方法允许输出中存在重复的标签，并在这些重复标签之间插入一个特殊的空白符号，例如“_”，以形成与输入语音帧长度相匹配的CTC路径。这种灵活性使得模型能够更好地适应不同语音输入的变化和复杂性。CTC的实现基于softmax分类层，该分类层包含了所有可能类别的节点以及用于表示空白的节点。

3.2.3 混合CTC/Attention模型

在处理大规模数据时，传统的Attention机制对齐方式可能会导致训练时间增加，这源于它缺乏明确的时间前后顺序依赖性。相比之下，CTC中采用的前向-后向算法则能够按照时间顺序依次高效地对齐输入序列和输出序列。为了解决Attention对齐效率较低的问题，同时提高整体训练效率，本文设计了一种混合CTC/Attention模型。在该模型中，CTC被引入作为辅助任务，用于加速对齐和解码过程．模型的整体结构示意图如图5所示。

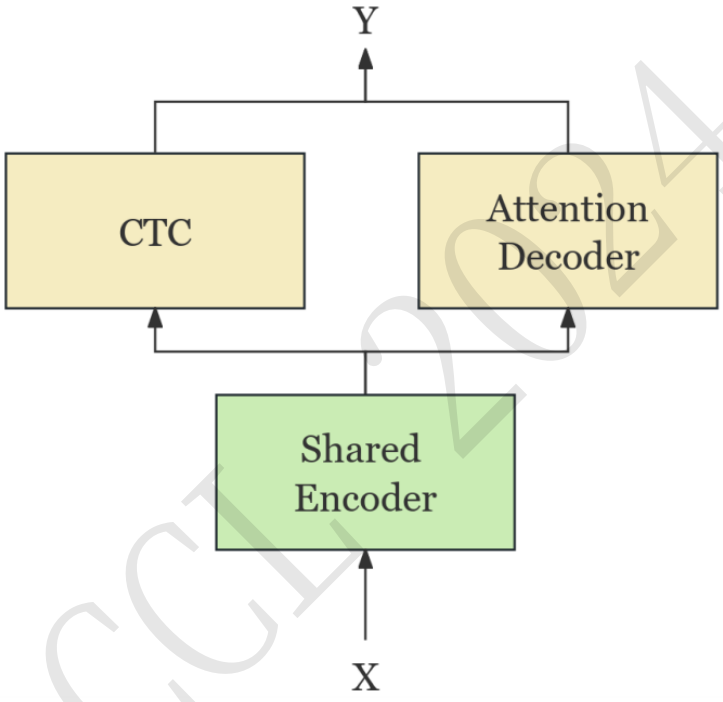

图 5. 混合CTC/Attention结构简图

3.3 损失函数

为联合优化模型，在帧级别上采用CTC损失，在标签级别上采用CE损失进行训练。本文模型损失函数的计算是将CTC损失(Lee and Watanabe, 2021)和CE损失加权相加，具体公式如下。

$$L(x,y) = \lambda CTC(x,y) + (1-\lambda)CE(x,y) \tag{6}$$

其中x是模型语音特征输入，y是相应的模型输出，以及λ是平衡CE损失和CTC损失的比重。

4 实验结果及分析

4.1 实验数据及评价指标

本文实验数据来自OpenSLR网站上开源的藏语安多数据集XBMU-AMDO31(Li et al., 2022)和藏语多方言TIBMD@MUC数据集(Zhao et al., 2020)。数据集XBMU-AMDO31由66位

母语发言人在安静的室内环境下通过手机录制，含有大约31小时的录音，22630个句子，总字数为2754。数据集TIBMD@MUC由87位母语发言人在室内环境下通过录音机、笔记本电脑和手机录制，包括卫藏方言52.53小时、康巴方言5.87小时和安多方言25.93小时，10390个句子，总字数为3006。两个数据集音频文件的采样频率均为16KHz，量化精度为16位。实验时，TIBMD@MUC数据集中仅采用25.93小时的安多方言数据，并对两个数据集的原始音频进行0.9倍和1.1倍的速度扰动将数据集扩充至3倍(Ko et al., 2015)。

采取词错误率(word error rate, WER)作为评价指标，其中S代表替换的音节数量，D代表删除的音节数量，I代表插入的音节数量，N代表总的音节数量，WER的值越小越好。具体公式如下：

$$WER = (S + D + I)/N \tag{7}$$

4.2 实验环境与参数

本文的深度学习框架基于Pytorch开发，编程语言为Python。具体工作站配置为：

环境详情	
操作系统	Ubuntu18.04
Linux	4.15.0
显卡数量	4
显卡版本	Tesla P100
内存	12G
Python版本	3.8
Torch版本	1.13.1
Cuda版本	11.2

表 1. 平台配置

实验中，使用80维Fbank(filter banks）作为输入特征，其中帧长为25 ms，帧移为10 ms。模型中的多头注意力头数均为4个，Encoder块有12个，Decoder块有6个，每个块的参数独立。训练中采用Swish作为激活函数，Dropout系数为0.1。使用Adam优化器，采用Warmup Step为30000的Warmup lr学习策略训练，设定学习率为0.0005，训练50轮。

4.3 实验分析

本文首先在XBMU-AMDO31和TIBMD@MUC数据集上验证了所提出的使用字节对编码算法进行语音建模的BPE-Conformer-CTC/Attention模型效果，并与基线模型Transformer-CTC/Attention进行对比，如表2所示。

模型(CTC权重为0.3，beam size为6)	建模单元	XBMU-AMDO31		TIBMD@MUC	
		验证集	测试集	验证集	测试集
Transformer-CTC/Attention	构件	22.2	20.6	13.2	12.3
	字丁	16.3	15.2	9.8	9.7
	音节	11.2	10.5	6.8	6.6
Conformer–CTC/Attention	构件	9.7	9.5	5.5	5.1
	字丁	9.8	9.3	4.9	4.6
	音节	9.5	9.2	4.9	4.6
BPE-Conformer-CTC/Attention	BPE500	**9.4**	**9.0**	**4.7**	**4.6**

表 2. 不同编解码器的WER(%)实验结果

实验结果表明，在TIBMD@MUC测试集上的BPE-Conformer-CTC/Attention模型效果与Conformer-CTC/Attention模型效果相当，相比于Transformer-CTC/Attention相对降低了30.3%。而在XBMU-AMDO31的测试集上，提出的BPE-Conformer-CTC/Attention模型相比于Conformer-CTC/Attention模型的词错误率相对降低了2.2%，相比于Transformer-CTC/Attention相对降低了14.2%。

通过对比两个数据集，TIBMD@MUC文本长度较短，而XBMU-AMDO31文本长度较长且内容更加丰富，使XBMU-AMDO31数据集复杂度较高，从而突显了本文所提出的模型具有较强的鲁棒性。基于此，本文以下实验均在XBMU-AMDO31数据集上进行。

为了探讨BPE词典规模大小对藏语语音识别的影响，本文选择在XBMU-AMDO31数据集上设置不同的词典规模进行实验。实验结果如表3所示。

BPE规模大小	模型	XBMU-AMDO31 验证集	XBMU-AMDO31 测试集
BPE100	ours	9.9	9.6
BPE300		9.4	9.1
BPE400		9.5	9.1
BPE500		**9.4**	**9.0**
BPE600		9.7	9.2
BPE800		9.5	9.2

表 3. BPE规模对实验结果WER(%)的影响

选择适当大小的BPE词典对语音识别模型的性能至关重要。若选择过小的词典，会导致一些词汇被合并成单个子词，从而丢失语言的细微差别和语义信息，进而降低模型对文本的理解能力，尤其是处理语义丰富的文本时。而选择过大的词典则会增加识别困难度，使模型复杂度和过拟合风险增加。实验结果表明，当BPE词典大小为500时，能在词汇覆盖、模型性能之间取得平衡，提供良好的文本处理和理解能力。此时该模型的WER为9.0%。

为了更好的体现基于藏语文字构成规则与基于字节对编码算法进行语音建模时的差异，本文在对解码时搜索束大小beam_size进行讨论的同时，分别对比了以构件、字丁、音节和BPE500为语音建模单元的模型效果。对比结果如表4所示。

模型(CTC权重为0.3)	建模单元	XBMU-AMDO31 验证集	XBMU-AMDO31 测试集
beam_size= 2	构件	9.9	9.4
	字丁	9.8	9.7
	音节	9.8	9.4
	BPE500	**9.4**	**9.2**
beam_size= 4	构件	10.0	9.5
	字丁	9.8	9.6
	音节	9.6	9.1
	BPE500	**9.5**	**9.1**
beam_size= 6	构件	9.7	9.5
	字丁	9.8	9.3
	音节	9.5	9.2
	BPE500	**9.4**	**9.0**
beam_size= 10	构件	10.0	9.6
	字丁	9.9	9.5
	音节	9.3	9.2
	BPE500	**9.5**	**9.0**

表 4. 探究beam-size大小以及建模单元的选择对实验结果WER(%)的影响

实验结果表明，纵向对比中，当beam_size的大小选为6时，四种建模单元的效果相较于其它beam_size大小的选择效果更好。横向对比中，相较于其它三种建模单元，BPE500作为建模单元效果最好。原因可能为以下几点：

（1）BPE通过反复合并出现频率最高的字符对，实现了对常见藏文字更好的捕捉。

（2）BPE能够减少稀疏性，特别是对于藏语中可能存在的稀有字符组合，能够有效提高模型的泛化能力。

（3）BPE能够更好地捕捉藏语的复杂语法结构，动态生成词汇，有助于提升模型在语音识别任务中的效果。

为了进一步验证混合CTC/Attention解码的有效性，本文通过改变CTC在联合优化损失中的占比进行实验。其中λ代表CTC所占的比重。CTC占比导致的模型性能如表5所示。

模型	CTC占比	XBMU-AMDO31 验证集	XBMU-AMDO31 测试集
BPE-Conformer-Attention	λ=0	11.6	11.1
BPE-Conformer-CTC	λ=1	39.8	36.9
BPE-Conformer-CTC/Attention	λ=0.2	9.6	9.0
	λ=0.3	**9.4**	**9.0**
	λ=0.5	9.5	9.3
	λ=0.7	10.0	9.5

表 5. CTC占比对实验结果WER(%)的影响

实验结果表明，当CTC和Attention机制联合解码时，实验效果都优于CTC或Attention机制单独解码，将CTC损失函数占比设置为0.3时藏语语音识别的效果最佳，此时WER为9.0%，验证了混合CTC/Attention解码的有效性。

5 总结和展望

本文构建了BPE-Conformer-CTC/Attention端到端藏语语音识别模型，在公开的藏语语音数据集上探究了不同的建模单元以及联合参数对藏语语音识别的影响，验证了该模型对藏语语音识别的有效性。在未来工作中，将继续深入研究和完善本文中提出的语音识别方法，进而拓展到藏语多方言的语音识别研究领域，以不断提升藏语语音识别系统的性能和实用性。

参考文献

Amodei, Dario and Ananthanarayanan, Sundaram and Anubhai, Rishita and Bai, Jingliang and Zhu, Zhenyao. 2015. *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.* Computer Science.

Chorowski, Jan and Bahdanau, Dzmitry and Serdyuk, Dmitriy and Cho, Kyunghyun and Bengio, Yoshua. 2015. *Attention-Based Models for Speech Recognition.* Computer ence, 10(4):429–439.

Dong, Linhao and Xu, Shuang and Xu, Bo. 2018. *Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition.* ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Graves, A. 2006. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks.* Proc. Int. Conf. on Machine Learning, 2006.

Gulati, Anmol and Qin, James and Chiu, Chung Cheng and Parmar, Niki and Zhang, Yu and Yu, Jiahui and Han, Wei and Wang, Shibo and Zhang, Zhengdong and Wu, Yonghui. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition.*

高耀荣，边巴旺堆. 2023. 基于端到端深度学习的藏语语音识别研究. 现代计算机, 29(17):25–30.

Huang, Mingkun and Zhang, Jun and Cai, Meng and Zhang, Yang and Ma, Zejun. 2020. *Improving RNN transducer with normalized jointer network.*

黄晓辉，李京. 2018. 基于循环神经网络的藏语语音识别声学模型. 中文信息学报, 32(5):7.

Ko, Tom and Peddinti, Vijayaditya and Povey, Daniel and Khudanpur, Sanjeev. 2015. *Audio augmentation for speech recognition.* Interspeech 2015.

Lee, Jaesong and Watanabe, Shinji. 2021. *Intermediate Loss Regularization for CTC-based Speech Recognition.*

Li, Senyan and Li, Guanyu and Ning, Jiewen. 2022. *XBMU-AMDO31: An open source of Amdo Tibetan speech database and speech recognition baseline system.* National Conference on Man-Machine Speech Communication, NCMMSC2022.

Shibata, Yusuke and Kida, Takuya and Fukamachi, Shuichi and Takeda, Masayuki and Shinohara, Takeshi. 1999. *Byte Pair Encoding: A Text Compression Scheme That Accelerates Pattern Matching.*

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Lukasz and Polosukhin, Illia. 2017. *Attention Is All You Need.* arXiv.

王松. 2019. 基于*LSTM-CTC*的藏语拉萨话语音识别系统. Master's thesis, 西北民族大学.

Yang, Xiaodong and Wang, Weizhe and Yang, Hongwu and Jiang, Jiaolong. 2020. *Simple Data Augmented Transformer End-To-End Tibetan Speech Recognition.* 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP).

Yue Zhao, Xiaona Xu, Jianjian Yue, Wei Song, Xiali Li, Licheng Wu, Qiang Ji. 2020. *An open speech resource for Tibetan multi-dialect and multitask recognition.* International Journal of Computational Science and Engineering.