基于离散化自监督表征增强的老挝语非自回归语音合成方法

冯子健[1,2]，王琳钦[1,2]，高盛祥*[1,2]，余正涛[1,2]，董凌[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
1456644199@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.com,
ztyu@hotmail.com,46761956@qq.com,

摘要

老挝语的语音合成对中老两国合作与交流意义重大，但老挝语语音发音复杂，存在声调、音节及音素等发音特性，现有语音合成方法在老挝语上效果不尽人意。基于注意力机制建模的自回归模型难以拟合复杂的老挝语语音，模型泛化能力差，容易出现漏字、跳字等灾难性错误，合成音频缺乏自然性和流畅性。本文提出基于离散化自监督表征增强的老挝语非自回归语音合成方法，结合老挝语的语言语音特点，使用老挝语音素粒度的标注时长信息构建非自回归架构声学模型，通过自监督学习的预训练语音模型来提取语音内容和声调信息的离散化表征，融入到声学模型中增强模型的语音生成能力，增强合成音频的流畅性和自然性。实验证明，本文方法合成音频达到了4.03的MOS评分，基于离散化自监督表征增强的非自回归建模方法，能更好的在声调、音素时长、音高等细粒度层面刻画老挝语的语音特性。

关键词： 语音合成；老挝语；非自回归；预训练语音模型

A Discretized Self-Supervised Representation Enhancement based Non-Autoregressive Speech Synthesis Method for Lao Language

Zijian Feng[1,2],Linqin Wang[1,2],Shengxiang Gao*[1,2],Zhengtao Yu[1,2], Ling Dong[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
1456644199@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.com,
ztyu@hotmail.com,46761956@qq.com,

Abstract

*高盛祥（通信作者）：gaoshengxiang.yn@foxmail.com
基金项目：国家自然科学基金（61972186,U21B2027）；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）；云南省基础研究计划（202001AS070014）；云南省科技人才与平台计划（202105AC160018）

Speech synthesis of Lao language is significant to the cooperation and communication between China and Lao, but Lao speech pronunciation is complex, and there are pronunciation characteristics such as tones, syllables and phonemes, and the existing speech synthesis methods do not work well on Lao language. The autoregressive model based on attention mechanism modeling is difficult to fit complex Lao speech, and the model has poor generalization ability and is prone to catastrophic errors such as word omission and word skipping, and the synthesized audio lacks naturalness and fluency. In this paper, we propose a non-autoregressive speech synthesis method for Lao based on discrete self-supervised representation enhancement, combining the linguistic phonetic features of Lao, using the annotated temporal information of Lao phoneme granularity to construct a non-autoregressive architectural acoustic model, extracting discrete representations of speech content and intonation information through a pre-trained speech model with self-supervised learning, and incorporating them into the acoustic model to enhance the model's speech generation capability of the model and the fluency and naturalness of the synthesized audio.The experiments demonstrate that the synthesized audio of this paper achieves a MOS score of 4.03, and the non-autoregressive modeling approach based on discrete self-supervised representation enhancement can better portray the speech characteristics of Lao language at the fine-grained level of intonation, phoneme duration, pitch, etc.

Keywords: TTS , Laotion , Non-Autoregressive , Pre-trained Speech Model

1 引言

老挝语是东南亚地区重要的语言之一，是老挝人民的官方语言，也在泰国、柬埔寨、越南等国家被广泛使用，研究老挝语语音合成对中老两国合作与交流意义重大。同时，研究如何融入语音特征提高老挝语语音合成性能，对后续研究非通用语种语音合成具有重要帮助。

老挝语与中英语种区别较大的地方是老挝语是声调语言，音调会直接添加到字符的上方，音调的改变会改变词语本身的意思，使得老挝语的语音合成需要在音节及音调上准确建模，由于其独特的发音方式，通用的语音合成方法在老挝语上难以保持语音合成效果，因此研究工作还存在很大困难。现有老挝语的语音合成缺乏自然度和流畅度，并且合成速度较慢。最近，(Anh et al., 2022)首次实现了基于神经网络的老挝语语音合成，然而该方法只是在基准模型上复现了老挝语语音合成。针对老挝语的语音合成任务还值得探索。

在语音合成任务中，主要目标有两个方面：(1)高质量：为了提高合成语音的自然度，模型应该捕捉到自然语言中的细节部分。(2)快速：在实际应用场景下，高速生成语音是至关重要的。一个完整的语音合成模型包括文本前端(Lai et al., 2021)，声学模型和声码器(Oord et al., 2016)(Kong et al., 2020)。本文从声学模型出发，结合老挝语语言语音特性研究如何构建更加高质量且快速的声学模型。声学模型从文本前端提供的信息中生成梅尔频谱图，然后使用单独训练的声码器根据梅尔频谱图来

合成语音，基于神经网络的语音合成系统显著的提高了合成音频的质量和自然度，并出现了很多面向特定领域的实际应用系统，Shen等人提出的Tacotron2模型架构(Shen et al., 2018)，相比于之前基于统计参数和级联系统语音合成方法，极大程度上改进了合成音频的质量。Ren等人提出基于Transformer(Vaswani et al., 2017)的非自回归架构FastSpeech(Ren et al., 2019)，移除了传统注意力机制对齐文本-语音的方法，选用了基于预测时长对齐的方式，使模型在解码的时候可以并行计算，极大地的提高了解码速度，同时解决了以往语音合成模型漏字和跳字等鲁棒性的问题。在此之后，Ren等人继续提出FastSpeech2(Ren et al., 2020)，训练和推理速度比自回归架构声学模型在速度上有极大提升。也有工作是在Tacotron2的基础上发展为非自回归模型(Elias et al., 2021)。这些工作在中、英等大语种上的语音合成取得了较好的成果，而针对非通用语种例如老挝语的语音合成工作还不足。为了解决老挝语语音合成音频缺乏自然度等问题，本文根据老挝语特性构建数据集，并利用微调的预训练语音模型提取老挝语语音特征，在传统FastSpeech2的语音合成声学模型架构体系上融合语音的离散化自监督表征来提升老挝语语音合成模型的性能。本文的贡献如下：

（1）实现了非自回归老挝语语音合成任务，解决了老挝语语音合成任务中，计算效率低，生成速度慢，声学模型泛化能力差的问题。

（2）提出了预训练语音模型融合机制策略，微调预训练语音模型，实现了将语音特征融入到声学模型中，改进了通用语音合成方法在老挝语上表现差的问题。

（3）在1小时左右的音频文本训练数据上的老挝语语音合成任务达到了4.03的MOS值。

2 相关工作

(1)自回归语音合成(Autoregressive Speech Synthesis)是一种基于序列模型的语音合成方法。在自回归语音合成中，语音信号被视为一个序列，每个样本都依赖于前面的样本。在传统的自回归语音合成方法中，通常采用的是循环神经网络（Recurrent Neural Network，RNN)(Grossberg, 2013)或者卷积神经网络(Convolutional Neural Network，CNN)(Gu et al., 2018)来建模语音信号的序列关系。这些模型能够学习到语音信号的时序特征，从而实现从文本到语音的转换。自回归语音合成方法具有一些优势，已被证明能生成连贯的语音信号，同时可以生成高质量的音频样本。

然而，由于其自回归的特性，有着昂贵的计算成本，自回归语音合成方法在生成速度方面相对较慢，因此在一些实时场景中不太适用，且在老挝语语音合成数据集较少的情况下，自回归语音合成模型可能会出现跳字、漏字等现象。

(2)非自回归语音合成（Non-autoregressive Speech Synthesis）是一种与自回归语音合成相对的语音合成方法。与自回归语音合成不同的是，非自回归语音合成模型不需要依赖于前面的样本来生成当前的样本，因此具有更高的生成速度和更低的延迟。Ren等人提出的FastSpeech2在FastSpeech的基础上添加了音调(Pitch)、音高（Energy）、时长（Duration）的外部语音信息，提高了合成语音的流畅度与自然度。另外，FastPitch(Łańcucki, 2021)基于FastSpeech对频率轮廓进行调节，提高了合成语音的整体质量。非自回归模型能以令人满意的速度生成语音音频，适用于实时语音合成等场景。

老挝语语音合成发音规律复杂，通用非自回归语音合成方法可能无法学习到足够的语音变化，从而导致过拟合或泛化能力不足，导致难于取得较好的效果。

第二十二届中国计算语言学大会论文集，第90页-第101页，哈尔滨，中国，2023年8月3日至5日。

3 方法

本文基于FastSpeech2的模型结构，根据老挝语语言语音特点构建数据集，在训练时融合语音特征，受Fang等人的启发(Fang et al., 2022)，本文将Wav2vec2.0提取的语音特征编码为隐状态，该隐状态序列与适应层的输出进行混合训练，对解码器输出的分布添加额外的JSD（Jensen-Shannon Divergence）损失来增强语音特征对生成语音的影响并提升训练效率。

3.1 数据构建

由于老挝语极其缺少语音文本数据对，很难利用MFA等工具(McAuliffe et al., 2017)对老挝语数据集进行切割，因此，为了做到音素对齐，本文在构建数据集阶段，对采集到的语音数据进行预处理，以使其适合进一步的分析和建模，包括对语音信号进行滤波、预加重、分割、去除噪声等预处理操作。其次使用Praat等语音分析工具(Styler, 2013)，从预处理后的语音信号中提取音高、音量、语速、音素持续时间等语音特征，对采集到的语音数据进行标注，对语音进行音素级别的标注和语音类型的分类，按照发音单元进行分割后，将预处理后的语音数据、提取的语音特征、标注信息和分割信息等整合在一起，构建数据集。老挝语音节的发音方式与其书写系统密切相关，老挝语以音节为最小发音单位，以音素为最小标注单位。并且老挝语没有官方的拉丁音译系统。本文首先老挝语辅音进行了详细的音素标注。如图1所示：

字符	ມ	ນ	ຍ	ງ	ບ	ດ	ປ	ຕ	ກ	ຜ, ພ
发音	[m]	[n]	[ɲ]	[ŋ]	[b]	[d]	[p]	[t]	[k]	[p^h]
字符	ຖ,ທ	ຂ, ຄ	ຝ, ຟ	ສ, ຊ	ຫ, ຮ	ຈ	ວ	ຢ	ຣ, ລ	ອ
发音	[t^h]	[k^h]	[f]	[s]	[h]	[tɕ]	[w]	[j]	[l]	[ʔ]

Figure 1: 老挝语辅音表

在老挝语中，音节是最小的发音单位，要组成音节则需要辅音与元音进行组合，图2中列出了具体的标注方式。

字符	ກະ	ກັ	ກິ	ກຶ	ກຸ	ເກະ	ເກັ	ເແກະ	ເແກັ	ໂກະ	ກໍ
发音	[ka]	[ka]	[ki]	[kɯ]	[ku]	[ke]	[ke]	[kɛ]	[kɛ]	[ko]	[ko]
字符	ເກາະ	ກັອ	ເກິ	ກາ	ກີ	ກຶ	ກຸ	ເກ	ເແກ	ໂກ	ກໍ
发音	[kɔ]	[kɔ]	[kɤ]	[ka:]	[ki:]	[kɯ:]	[ku:]	[ke:]	[kɛ:]	[ko:]	[kɔ:]
字符	ກອ	ເກີ	ເກັຍະ	ກັງ	ເກືອ	ກົວະ	ກົອ	ໄກ	ໃກ	ກັຍ	ເກຍ
发音	[kɔ:]	[kɤ:]	[kiə]	[kiə]	[kɯə]	[kuə]	[kuə]	[kai]	[kai]	[kai]	[ki:ə]
字符	ກງ	ເກືອ	ກົວ	ກອ	ກາຍ	ກ໋າ					
发音	[ki:ə]	[kɯ:ə]	[ku:ə]	[ku:ə]	[ka:i]	[kam]					

Figure 2: 老挝语音节组合实例

3.2 模型结构

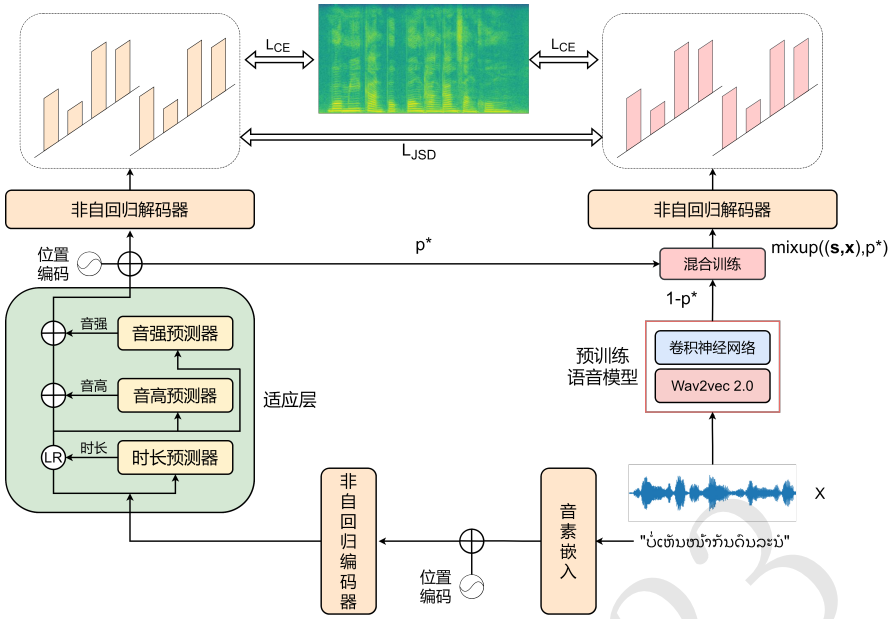

Figure 3: 基于离散化自监督表征增强的老挝语非自回归语音合成方法

模型的整体结构如图3所示，模型的输入为平行的文本-音频数据对，图中s为输入文本序列所对应的音素序列，x为文本序列对应的音频序列。采用非回归形式的编码器+解码器的架构，其中编码器、解码器分别由N个transformer层组成(N=4)，在编码层与解码层之间引入变换适应层(Variance Adaptor)用来作音素之间停顿的预测以及音调、音强的预测，使模型更好地建模音频特征。适应层预测的输出与提取的语音特征做混合训练，由于文本与语音结构不同，模态差异较大，语音是连续的时序信号，而文本是离散的符号序列，通过mixup方法，模型可以在不同模态之间建立联系和相互影响，混合文本-音频的数据可以帮助模型学习到文本到语音的相关性和一致性信息，从而提高模型对输入的理解和表达能力(Fang et al., 2022)。对解码器输出的分布添加额外的JSD损失，引入JSD损失用于度量生成的样本分布与真实样本分布之间的差异，它可以帮助模型学习生成更逼真的样本，来增强语音特征对生成语音的影响并提升训练效率(Gulrajani et al., 2017)。该损失函数为：

$$\mathcal{L}_{\text{JSD}}(\mathbf{s},\mathbf{x},\mathbf{y},p^*) = \sum_{i=1}^{|\mathbf{y}|} \text{JSD}\{p_\theta(\mathbf{y}_i \mid \mathbf{y}_{<i}, \mathbf{h1}(\mathbf{s})) \parallel p_\theta(\mathbf{y}_i \mid \mathbf{y}_{<i}, \mathbf{h2}(\text{mixup}((\mathbf{s},\mathbf{x}),p^*)))\} \quad (1)$$

其中h1(s)为文本通过音素编码器和适应层输出的上下文表示，h2(mixup((s,x),p^*))为通过预训练语音模型提取的语音特征向量经过声学编码器输出的向量与适应层输出的混合表征。

加上两次交叉熵损失，最终的损失函数如下：

$$\mathcal{L} = \lambda\mathcal{L}_{\text{JSD}}(\mathbf{s},\mathbf{x},\mathbf{y},p^*) + \mathcal{L}_{\text{CE}}(\mathbf{s},\mathbf{y}) + \mathcal{L}_{\text{CE}}(\text{mixup}((\mathbf{s},\mathbf{x}),p^*),\mathbf{y}) \quad (2)$$

其中 λ 是控制JSD损失的权重系数。

3.3 微调预训练语音模型

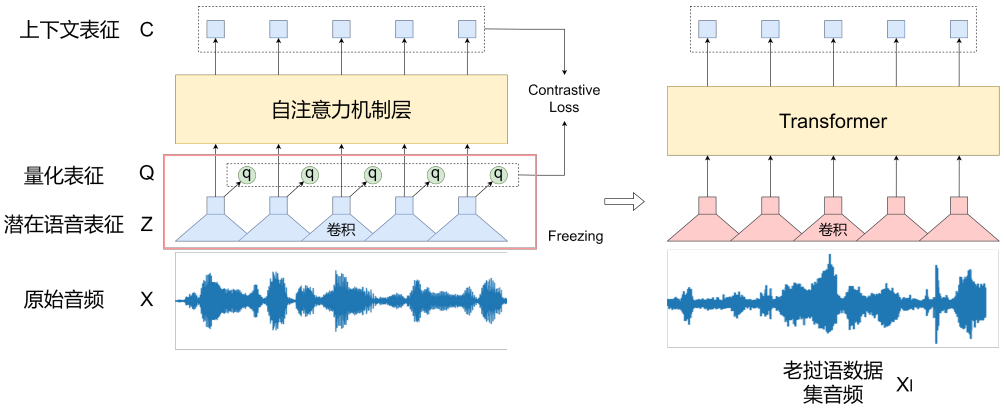

Figure 4: 微调wav2vec2.0

老挝语与中英等大语种差异较大，在通用方法下只对文本到音频进行建模使得模型难以充分训练，本文提出通过融入语音特征来提高模型的性能。

受wav2vec2.0启发，本文微调了预训练的多语言语音模型，利用无监督的语音预训练模型，迁移到语音合成任务中。具体而言，本文用预训练的wav2vec2.0模型结构中添加了一个自注意力机制层来增加语音量化表征上下文表征向量的相似度，根据数据集调整了学习率以及迭代次数。使用测试集来评估微调模型的性能。用评估结果来进一步调整模型架构和训练超参数。实验结果表明本文微调过的预训练语音模型所提取出来的语音特征可以提高语音合成的质量与流畅度。

在微调过程中，由于本文的数据集具有精确的音素持续时长标注，本文对特征编码器的输出采用了与SpecCutout类似的屏蔽策略(Kriman et al., 2020)：随机选择一些起始时间步长，对这些时间步长的数个后续时间步长的语音信号，将语音信号的频谱图进行分割，得到一些小块的频谱子图。在这些频谱子图中随机选择一些子图，然后将这些子图内的所有频谱值全部屏蔽（即用0来替换这些子图内的所有频谱值）。将所有被屏蔽的频谱子图拼接起来，得到一张被局部屏蔽的频谱图来代替原本的频谱图。本文使用与预训练时相同的屏蔽时间步长嵌入。图4中的对比损失为：

$$L = -\log \frac{\exp\left(\mathrm{sim}\left(c_t, q_t\right)/\kappa\right)}{\sum_{\hat{q}\sim Q_t} \exp\left(\mathrm{sim}\left(c_t, \hat{q}\right)/\kappa\right)} \tag{3}$$

其中

$$\mathrm{sim}(a,b) = a^T b/\|a\|\|b\| \tag{4}$$

4 实验

4.1 实验设置

本文在内部时长约一小时的老挝语数据集上进行了实验，音频采样率为22.05kHz，中间输出为特征维度大小为80的梅尔频谱图。音频由母语为老挝语的单人在专业录音

室录制，其训练集和验证集的大小比为4:1。训练时batchsize为32，使用$\beta_1 = 0.9,\beta_2 = 0.999,\varepsilon = 10^{-6}$的Adam优化器，学习率为$10^{-3}$。

4.2 评价指标

主观评价部分采用两种评价指标对模型综合能力进行评价。指标一：采用平均意见分（Mean Opinion Score Score，MOS）来评价合成语音的自然度和流畅度，听者根据自身感受对合成音频进行打分，MOS值评分共分为1-5五个等级，1分差，2分一般，3分正常，4分良好，5分最优，最后根据所有听者给出的意见分计算平均意见分。指标二：采用ABtest方式选择出听者感受更好的音频，ABtest方案会设置两组不同的语音合成系统的合成音频，听者盲听两组的音频并选择出较优的一方，最后计算听者的选择占比。以下评价模型优劣均为老挝语为母语的听众完成。

客观评价部分采用三种评价指标，指标一：实时因子RTF(Real Time Factor)，该值是评估非自回归模型推理速度的客观指标。指标二：MCD(Mel cepstral distortion)值，它表示的是转换后语音的MFCC 特征与标准输出语音的MFCC 特征的差距，用来验证本文的合成语音在语音特征上的保留程度。指标三：SSIM (Structural Similarity),它表示的是两个图像之间的相似程度，本文用来计算合成音频与真实音频之间的梅尔谱图相似度。

4.3 实验结果与分析

4.3.1 主观评价及客观评价

为了证明本文方法能够在保证快速生成语音的同时提高语音的自然度与流畅度，本文设置了与4个基准模型的对比实验，其中Tacotron2(Elias et al., 2021)是自回归语音合成模型，Tacotron2+GA是在Tacotron2模型的基础上加入了guide attention(Tachibana et al., 2018)，占比权重$\alpha = 1$。FastSpeech2(Ren et al., 2020)是本文的基准模型。FastPitch(Łańcucki, 2021)是另一个基于FastSpeech的非自回归模型。所有基准模型以及本文提出的模型都使用预先训练的HiFi-GAN(Kong et al., 2020)作为声码器，各基准模型的实验设置与提出该基准模型的原论文一致。各评价指标得分如表1所示。

Table 1: 客观评价指标与主观评价指标MOS评分。

方法	MCD	SSIM	RTF	MOS
Ground Truth	-	-	-	4.52±0.07
Tacotron2	7.76	0.45	2.31×10^{-1}	3.71±0.08
Tacotron2+GA	7.82	0.48	2.55×10^{-1}	3.86±0.07
FastSpeech2	7.78	0.43	$\mathbf{1.98} \times 10^{-2}$	3.85±0.08
FastPitch	8.38	0.46	2.16×10^{-2}	3.82±0.08
Our Model	**7.61**	**0.53**	2.07×10^{-2}	**4.03±0.07**

通过对表1的MOS评分进行分析，本文提出的模型能在只有一小时的老挝语音频文本训练数据下训练出完整的语音合成模型，并且使FastSpeech2在主观听觉上的表现超过了自回归的Tacotron2，证明本文通融合老挝语语音特征训练模型能够在一定程度上提升老挝语语音合成模型在韵律上的表现，并且取得了相比于基准模型更高的评分，相较于基准模型FastSpeech2,增加了0.18的MOS评分。

第二十二届中国计算语言学大会论文集，第90页-第101页，哈尔滨，中国，2023年8月3日至5日。

本节同时对实验结果进行客观评估，包括MCD(Kominek et al., 2008)值以及SSIM值(Wang et al., 2004)以及RTF值。其中MCD值的范围在0-10之间，数字越小说明两个音频之间的差距越小。SSIM的范围在0-1之间，数字越大说明两个图片的相似度越高。RTF值的计算方式为处理音频的时长/音频时长，在统一的设备配置下，RTF值越小说明实时率越高。通过表1进行分析，可以看出本文提出的模型在MCD指标上相比基准模型FastSpeech2降低了0.17。且SSIM值提高了0.1。

4.3.2 推理时长表现

为了验证本文提出模型是否具有快速的语音生成速度，本文对各基准模型在不同长度文本上进行了实验，推理时长实验结果如图5所示。

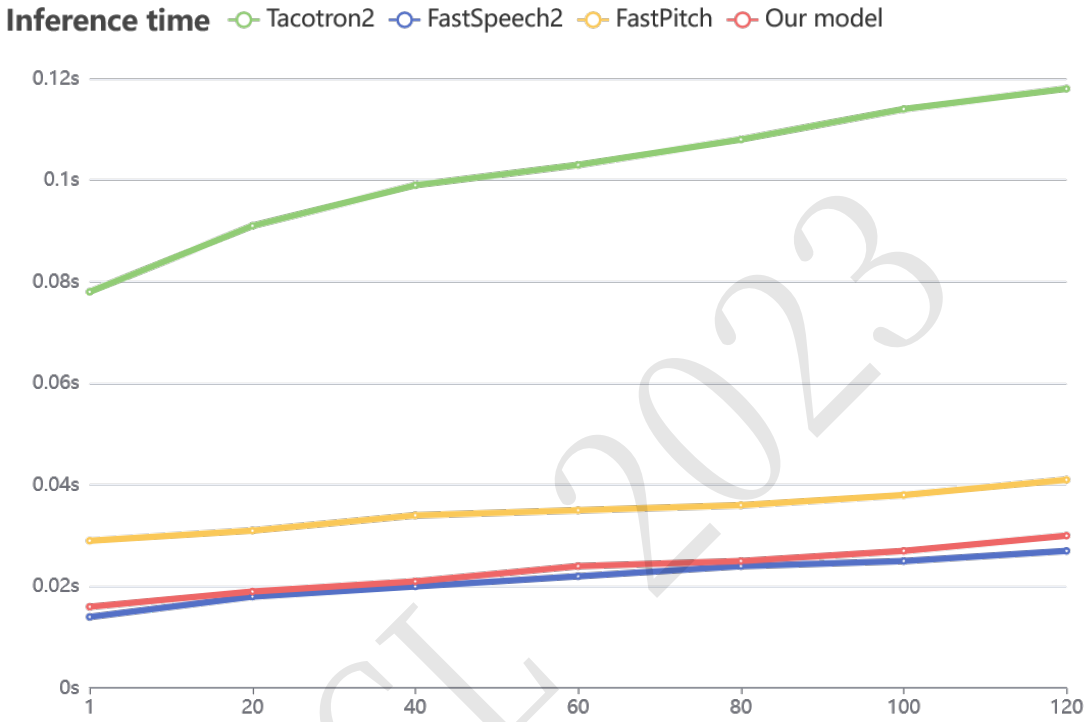

Figure 5: 模型推理时长表现

通过图5可以看出。由于自回归架构中每个样本都依赖于前面的样本，Tacotron2的推理时长相较于其他基准模型是花销更大的。而能够并行计算的FastPitch、FastSpeech2以及本文方法都具有较快的推理速度。本文方法在推理速度上的表现远远超出Tacotron2，相较于FastSpeech2也几乎没有增加开销。

4.3.3 消融实验

为了探索融合语音特征的最佳方式以及探索微调后的语音特征是否对模型性能提升具有决定性作用，本文进行了消融实验。分别尝试了对Tacotron2+GA模型直接融合提取的语音特征，对FastSpeech2将适应层的输出与特征向量直接进行拼接、加入注意力机制(Cross Attention)以及最终的模型。实验结果如表2所示。

Table 2: 消融实验：不同融合语音特征方式对声学模型性能的影响

方法	MCD	SSIM	RTF	MOS
Tacotron2+GA without finetuned wav2vec2.0	7.82	0.48	2.55×10^{-1}	3.86±0.07
Tacotron2+GA with finetuned wav2vec2.0	7.88	0.47	2.71×10^{-1}	3.87±0.06
FastSpeech2 without finetuned wav2vec2.0	7.78	0.43	$\mathbf{1.98} \times 10^{-2}$	3.85±0.08
FastSpeech2 with finetuned wav2vec2.0	7.98	0.46	2.01×10^{-2}	3.88±0.08
通过Cross Attention方式融合语音特征	7.64	0.51	2.51×10^{-2}	3.92±0.07
Our Model	**7.61**	**0.53**	2.07×10^{-2}	**4.03±0.08**

通过对表2进行分析，本文提出的融合语音特征方法在MOS评分上取得了最好的效果。FastSpeech2不融合语音特征的基线模型MOS评分为3.85，在直接融合了语音特征之后提升了0.03的MOS评分，达到了3.88，在一定程度上提高了语音的自然度和流畅度。而Tacotron2+GA基线模型的MOS评分达到了3.86，在直接融合语音特征之后提升了0.01的MOS评分，原因是Tacotron2模型受限于自回归模型本身泛化能力差，容易出现漏字、跳字等灾难性错误的问题。可以得出结论，本文主实验中的性能提升主要来源于本文提出的方法上的改进，而非在老挝语数据集对wav2vec2.0上进行微调。本文提出的对FastSpeech2融合语音特征方法，在不同融合方式上都提高了模型所生成语音在听觉上的表现，并且在MCD值和SSIM值上的表现均有提高。虽然FastSpeech2模型在RTF上有比最终模型有更好的性能，但由于本文提出的方法在生成语音的自然度和流畅度上有更好的表现，且RTF值相比基线模型的损失并不大，本文还是以获得了最高MOS评分、MCD值以及SSIM值的方法作为最终模型。

4.3.4 梅尔频谱图分析

为了分析本文提出的方法是否能借助预训练的语音模型在细节部分进行更精细地建模，本文对具有相同内容的语音进行了梅尔频谱图的分析。如图6所示。

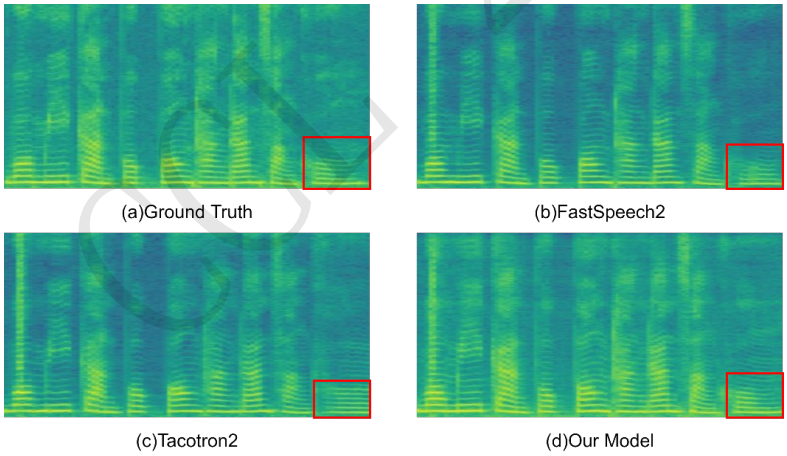

Figure 6: 梅尔频谱图分析

对图6中的红色部分进行梅尔频谱图分析，对于同样内容的语音，它的ground truth转换的梅尔频谱图如图中(a)所示，用FastSpeech2生成的梅尔频谱图如图中(b)所

示，可以看出在所选区域的梅尔频谱图有明显失真。而用Tacotron2模型生成的梅尔频谱图的对应位置可以发现实际上虽然Tacotron2合成该音频保留了发音，但是在结构上与原始音频相差很大，如图中(c)所示。而本文提出的模型可以完整保留该部分发音，在梅尔频谱图结构细节上与原始谱图保持了高度一致，如图中(d)所示。

4.3.5 ABtest实验

为了更直接地对比不同方法在主观上的优劣，本文进行了ABtest实验，该实验是让听者盲听两个不同模型对同一文本合成出来的音频，并选择出音频较优一方。参与该测试的一共有30人。测试结果如图7所示。

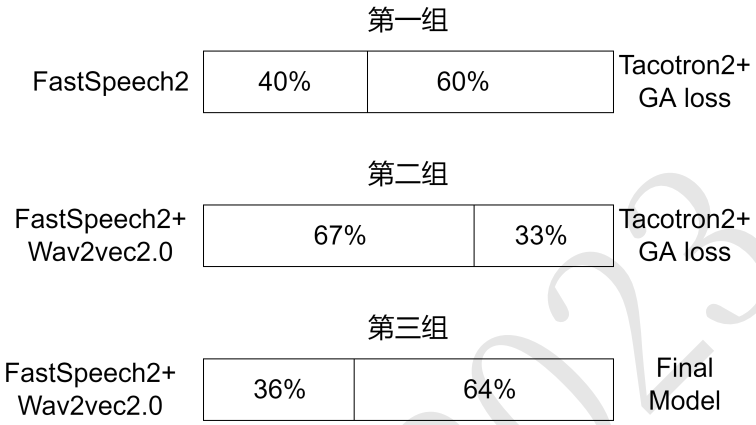

Figure 7: ABtest实验

测试显示，在第一组ABtest实验中，有超过一半的听众选择了自回归架构的Tacotron2模型合成的语音，说明在该数据集上仅仅基于FastSpeech2所合成的模型生成的语音在主观上不如Tacotron2。而通过第二组对比表明，基于FastSpeech2融合语音特征的模型所合成的语音的效果就超过了Tacotron2，这说明对于非自回归模型来说，融合语音特征的效果是有利于模型生成更自然、流畅的语音的。通过第三组实验可以对比出不同融合方式下，听众对语音优劣的主观选择，从而帮助本文定位出最适合融合语音特征的方法。

5 结论

针对通用语音合成方法在老挝语上表现较差的问题,本文提出基于离散化自监督表征增强的老挝语非自回归语音合成方法，结合老挝语的语言语音特点，在老挝语音素粒度上标注时长信息，使用非自回归架构建模声学模型提高老挝语语音合成速度，通过自监督学习的预训练语音模型来提取语音内容和声调信息的离散化表征，融入到声学模型中增强模型的语音生成能力，增强合成音频的流畅性和自然性，并且保证推理时长几乎不增加。实验证明，本文方法合成音频达到了4.03的MOS评分。

第二十二届中国计算语言学大会论文集，第90页-第101页，哈尔滨，中国，2023年8月3日至5日。

参考文献

Nguyen Thi Ngoc Anh, Nguyen Tien Thanh, et al. 2022. Development of a high quality text to speech system for lao. In *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–5. IEEE.

Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu. 2021. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*.

Stephen Grossberg. 2013. Recurrent neural networks. *Scholarpedia*, 8(2):1888.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. 2018. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans.

John Kominek, Tanja Schultz, and Alan W Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE.

Tuan Manh Lai, Yang Zhang, Evelina Bakhturina, Boris Ginsburg, and Heng Ji. 2021. A unified transformer-based framework for duplex text normalization. *arXiv preprint arXiv:2108.09889*.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Will Styler. 2013. Using praat for linguistic research. *University of Colorado at Boulder Phonetics Lab*.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4784–4788. IEEE.

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.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.