基于神经编解码语言模型的老挝语韵律建模方法

易宁静[1,2]，王琳钦[1,2]，高盛祥*[1,2]，余正涛[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
897079505@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.com,
ztyu@hotmail.com

摘要

为了赋予合成语音类似人类语言的丰富韵律和节奏变化，现有方法普遍采用基于随机数的时长预测器。这些方法通过使用随机数初始化的潜在变量来模拟人类说话的多样节奏变化。然而，由于依赖于随机数噪声的局限性，这些方法合成的语音往往仍然缺乏真实语音的多样性和韵律变化的丰富性。与之前方法不同，本文提出了一种基于神经编解码语言模型（VALL-E）的韵律建模方法，本文利用先验速度和音调时序变化曲线建模韵律变化分布，有效融入神经编解码语言模型训练过程中，并且在推理阶段可通过控制先验时序曲线控制生成语音的韵律。实验证明，本文方法合成英语音频达到了4.05的MOS评分，合成老挝语音频达到了3.61的MOS评分。基于神经编解码语言模型的老挝语韵律建模方法，能很好的在速度和音调方面实现韵律的可控性。

关键词： 韵律；老挝语；神经编解码语言模型；可控性

A Method for Lao Prosody Modeling Based on Neural Codec Language Model

Ningjing Yi[1,2], Linqin Wang[1,2], Shengxiang Gao*[1,2], Zhengtao Yu[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
897079505@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.coma,
ztyu@hotmail.com

Abstract

In order to give synthesized speech rich prosody and rhythm changes similar to human language, existing methods generally use duration predictors based on random numbers. These methods simulate the diverse rhythmic variations of human speech by using latent variables initialized with random numbers. However, due to the limitations of relying on random number noise, the speech synthesized by these methods often still lacks the diversity of real speech and the richness of prosodic changes. Different from previous methods, this paper proposes a prosody modeling method based on the neural codec language model (VALL-E). This paper uses a priori speech speed and pitch timing change curves to model the distribution of prosody changes, effectively integrating into the neural codec language During the model training process

*高盛祥（通信作者）：gaoshengxiang.yn@foxmail.com
基金项目：国家自然科学基金（62376111，U23A20388，U21B2027，62366027）；云南省重点研发计划（202303AP140008，202302AD080003，202401BC070021，202103AA080015）；云南省科技人才与平台计划（202105AC160018）

and during the inference stage, the prosody of the generated speech can be controlled by controlling the prior timing curve. Experiments have shown that the method in this article achieves a MOS score of 4.05 when synthesizing English audio, and a MOS score of 3.61 when synthesizing Lao audio. The Lao prosody modeling method based on the neural encoding and decoding language model can achieve good controllability of prosody in terms of speaking speed and pitch.

Keywords: Prosody , Lao , Neural Codec Language Model , Controllability

1 引言

老挝与中国比邻，2021年12月3日中老铁路的开通进一步加深了中国与老挝的国际交流，开展老挝语的韵律建模研究对促进两国政策沟通有重要支撑，符合国家“一带一路”建设等战略需求。近年来，语音合成模型在合成声音的自然度方面有了极大的提高，跳词复读的问题也得到了解决。因此，如何在合成语音中实现人类语音的表现力是一个研究热点，而让语音富有表现力的关键，就是对语音进行韵律建模。

相较于中、英等语言，老挝语与它们的发音存在一定的区别。老挝语的字符读法分为元音和辅音两类，元音分长短两类，共28个，辅音共27个。另一方面，老挝语是音调语言，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，这使得老挝语的语音合成需要在音节及音调上准确建模。最近，(Anh and Thanh, 2022)首次实现了基于神经网络的老挝语语音合成，但这项研究工作只是在基准模型上实现了老挝语的语音合成，在合成语音的韵律方面并没有达到很好的效果。因此，针对老挝语的韵律建模仍是一个值得探索的问题。

韵律是语言中多种特征的融合，韵律建模内容可包括速度、音高、持续时间等等信息，主要使合成的语音更加拟人化。对于韵律建模任务，按照韵律调整的粒度可分为两类：(1)粗粒度：粗粒度为句子层面的迁移调控。(2)细粒度：细粒度为短语，单词层面的调控。在粗粒度方面，(Skerry-Ryan et al., 2018)等人提出的参考韵律嵌入方法，首次使用无监督学习来进行韵律的迁移，把参考的语音编码成一个向量，向量包含参考语音的韵律信息。(Wang et al., 2018)等人提出的风格token可以对韵律信息进行解耦，使其每个token控制一种风格。但存在韵律信息无法完全解耦，每个token仍包含多种信息的问题。在细粒度方面，(Lee and Kim, 2019)等人在帧级别和音素级别，使用参考编码器进行细粒度的韵律建模。但缺点是对未见说话人进行韵律迁移效果较差。(Klimkov et al., 2019)提出了对参考音频进行单独的信息抽取方法，并通过变分自编码器模型(VAE)对韵律信息进行预测，改进了对未见说话人的韵律迁移效果。这些工作在韵律迁移方面已经取得了较好的效果，但在韵律的可控性方面并没有受到太多关注。在语言交流中，韵律的可控性对于传达说话者的情感、意图甚至语义至关重要。例如，放慢语速和提升音调有助于突出某一个词语，强调这个词语的重要性。因此，实现对韵律的可控性能使语音合成技术在许多应用场景中发挥更好的效果，这正是本文所关注的。

为了解决老挝语韵律建模任务中对语言特征缺乏可控性等问题，受ControlVC(Chen and Duan, 2022)的启发，本文提出了一种可以对速度和音调进行时变控制的老挝语语音合成模型。本文在神经编解码语言模型VALL-E(Wang et al., 2023)的架构体系上，利用先验速度和音调时序变化曲线建模韵律变化分布，随后分别在英语和老挝语数据集上进行训练，最终提出韵律可控的语音合成模型。本文的贡献如下：

(1)提出了韵律可控的语音合成方法，实现了在速度和音调多种语言特点上的可控性韵律建模，解决了语音合成任务中合成语音韵律不丰富、不可控的问题。

(2)构建不同的字符转音素方法，利用大规模多语言、多后端字符转音素系统Epitran构建老挝语音素字典，实现了在神经编解码语言模型上的低资源语言老挝语的韵律建模。

(3)在585小时的英语数据集上的合成语音达到4.05的MOS评分，在50小时左右的老挝语数据集上的合成语音达到3.61的MOS评分，

2 相关工作

(1)传统的端到端(End-to-End)语音合成模型通常由文本分析、声学模型、声码器三个部分组成，基于端到端语音合成模型的韵律建模通常是在声学模型部分进行修改。有通过全局风格

标记(Wang et al., 2018)的方法，训练不需要真实标签，模型学习将各种噪声和说话者因素分解为单独的风格标记；有通过韵律调控(Lee and Kim, 2019)的方法，即在嵌入网络中引入时间结构，从而实现对合成语音说话风格的细粒度建模；有通过隐变量(Kim et al., 2021)而非频谱串联起来声学模型和声码器的方法，在隐变量上进行随机建模并利用随机时长预测器，对于同样的输入文本，能够合成不同韵律的语音，提高了合成语音的多样性。但由于其依赖于随机数噪声的局限性，合成的语音仍然缺乏真实语音的多样性和韵律变化的丰富性。ControlVC(Chen and Duan, 2022)提出了一种对音调和速度进行时变控制的神经语音转换模型。ControlVC使用控制曲线对音调和速度进行时变控制，使用预先训练的编码器来生成音调和语言嵌入，使用说话人编码器生成说话人嵌入，最后使用声码器将其组合并转换为目标语音。但其训练数据量小，ControlVC使用的公共英语数据集仅包含约44小时的英语语音。且上下文学习能力弱，并不能在推理过程中转换出未见说话人的语音。

端到端语音合成模型的优点是简化了语音合成系统中的流程，减少了多个阶段的处理步骤，使得系统更加简洁高效。但该方法合成语音的质量还有待提升。

(2)神经编解码语言模型(Neural Codec Language Model)引入了一种用于语音合成的语言建模方法，将语音合成视为条件语言建模任务。与端到端语音合成模型相比，神经编解码语言模型的中间表示不再是梅尔谱图，而是离散的编码。具体来说，神经编解码语言模型引入了一种基于提示的语音合成方法，以EnCodec(Défossez et al., 2022)的音频编解码器作为中间表示。通过将语音波形编码为离散声学代码，这些模型能够重建高质量的波形，即使对于未见说话人也是如此。基于神经编解码语言模型的韵律建模也已经开展了一些工作，例如SC VALL-E(Kim et al., 2023)，它从文本句子和提示音频中获取输入，不再是简单地模仿提示音频的特征，而是通过控制属性来产生不同的声音。

神经编解码语言模型的优点是训练数据量大，且上下文学习能力强。对于语音合成任务而言，如果模型无需微调即可为训练中未出现过的说话人合成高质量的语音，则该模型被认为具有上下文学习能力。而神经编解码语言模型无需微调，只需输入未见说话人的3秒语音，就能合成出高质量的音频。

3 方法

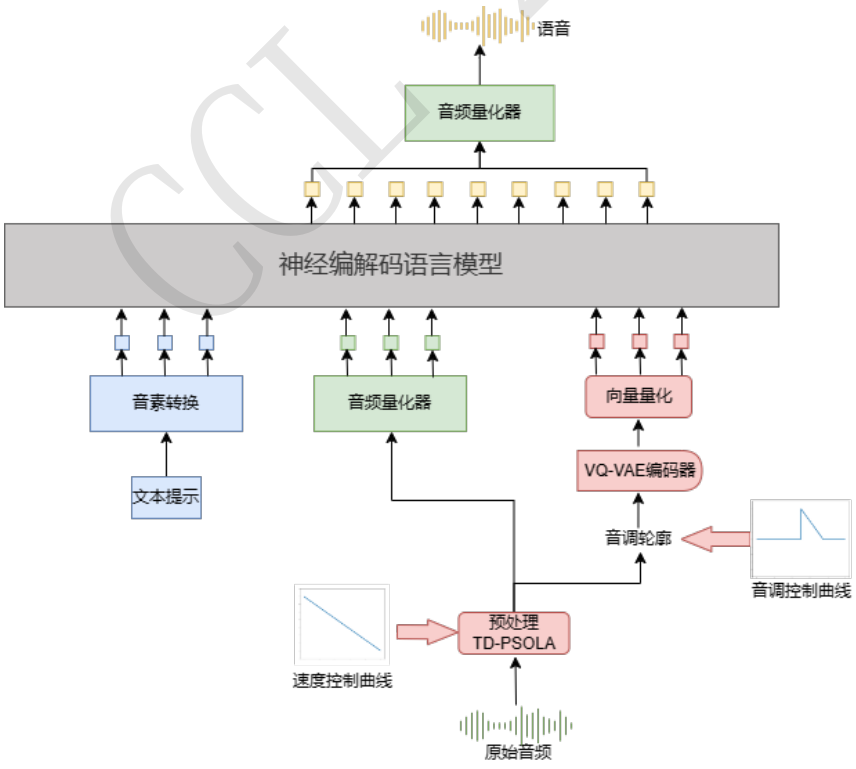

图 1: 模型总图

本文旨在利用控制曲线在英语和老挝语语音合成中实现韵律的时变控制。本文基于VALL-E的模型结构，受ControlVC(Chen and Duan, 2022)的启发，如图1所示，系统由三个部分组成：速度预处理和音素转换部分、音频量化和音调嵌入部分、音频合成部分。在速度预处理和音素转换部分，采用TD-PSOLA算法，根据速度控制曲线修改输入语音的语速，并且将输入的文本转换为音素。在音频量化和音调嵌入部分，将预处理后的音频分别送入神经音频编解码器和VQ-VAE编码器，分别转换为离散声学代码及离散音调嵌入。最后，在音频合成部分，将离散代码重构为语音波形。

3.1 速度预处理和音素转换

在预处理阶段，我们使用时域音调同步重叠相加（TD-PSOLA）算法(Charpentier and Stella, 1986)，根据输入的速度控制曲线修改语音的语速。我们首先对输入的音频重采样至24kHz，再对音频进行分段，并使用控制曲线相应位置指示的拉伸比率对每一段音频的时间进行拉伸或收缩。遵循ControlVC，如图2所示，速度控制曲线的三种模式，分别为上升、下降和抛物线。具体来说，首先我们对原始音频进行分段，再使用np.array函数创建整数序列，并根据线性公式或抛物线公式计算每一段的速度，使语音的速度可以随着时间越来越快，越来越慢，也可像抛物线一样先快后慢。在这一阶段，语音的音色、音调均被保留。

基准模型VALL-E使用的字符转音素方法并不支持老挝语，因此，我们构建特殊的字符转音素模块来进行老挝文本的字符转音素。在这一模块中，我们引入了Epitran(Mortensen et al., 2018)，Epitran是一个用于字符转音素的大规模多语言、多后端系统，支持包括老挝语在内的多种语言，它采用语言正字法中的单词标记并输出音素表示，我们再将其输出的音素表示构成老挝语音素词典。最终达到捕捉老挝语语言的差异性，模型更好地学习老挝语语言知识的效果。

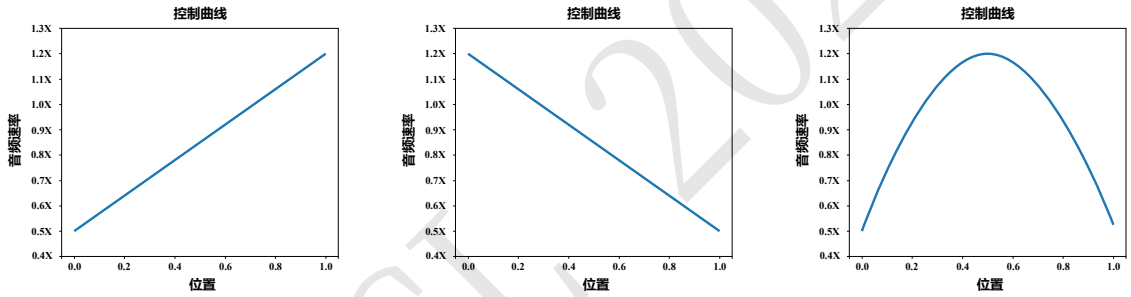

图 2: 速度控制曲线

3.2 音频量化和音调嵌入

由于音频通常存储为16位整数值序列，因此需要生成模型在每个时间步输出$2^{16} = 65536$个概率来合成原始音频。此外，超过万的音频采样率导致序列长度过长，使得原始音频合成变得更加困难。为此，需要音频量化来压缩整数值和序列长度。在本文中，我们遵循AudioLM(Borsos et al., 2023)，利用神经编解码器模型来表示离散标记中的音频。为了压缩音频以进行网络传输，编解码器模型能够将输入的语音波形编码为离散声学代码，对于训练中未出现的说话者，也能重建出高质量波形。神经编解码器可以减少时间步长以提高效率，并且它包含丰富的说话人信息和声学信息。在本文中，我们采用预先训练的神经音频编解码器模型EnCodec(Défossez et al., 2022)作为我们的量化器。EnCodec是一种卷积编码器-解码器模型，其输入和输出都是可变比特率的采样率为24kHz的音频。编码器将24kHz的输入波形产生75Hz的嵌入，采样率降低了320倍。其中量化器共有八层，每个嵌入均通过残差向量量化(RVQ)进行建模。

我们采用YAAPT算法(Kasi and Zahorian, 2002)提取经过速度控制的语音的音调序列$(p_1, ..., p_T)$，其中T 是帧。然后将该音调序列乘以输入的音调控制曲线以获得修改的音调序列$(p'_1, ..., p'_T)$。同样遵循ControlVC，输入的音调控制曲线示例如图3所示，模式为stress模式，该曲线表示修改后的语音与原始语音的基频(F0s，单位为Hz)的比率。例如，1.2x F0率表

示修改后的语音的F0预计为原语音的1.2倍。这里的实现思路是，在音调序列中，选取一段区间作为上升的部分，对该部分音调进行线性调整，使得音调值随着位置的增加而线性上升。修改后的音调序列作为基于VQ-VAE编码器的音调嵌入网络(Dhariwal et al., 2020)的输入，以获得修改后的音调嵌入。编码器产生潜在向量序列$(h_1, ..., h_T)$，并将其映射到码本中各自最接近的代码。然后，我们采用码本向量的整数索引来形成音调嵌入序列$z^{(p)} = (z_1^{(p)}, ..., z_T^{(p)})$。

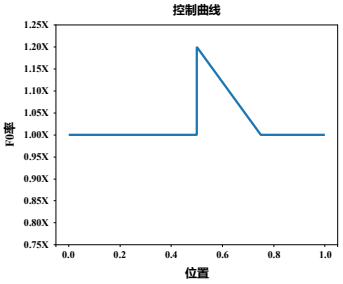

图 3: 音调控制曲线

3.3 训练策略

在这项工作中，我们通过采用具有两个编码器-解码器的编解码语言模型的分层方法来实现条件编解码语言模型的功能。这两个模型采用不同的建模方法：自回归和非自回归。我们首先应用自回归编解码语言模型从语音样本的第一个量化器生成离散代码。接下来，我们利用非自回归编解码语言模型来预测剩余的离散代码。我们可以将我们的模型表示为$\theta = \theta_{ar}, \theta_{nar}$，其中$\theta_{ar}$和$\theta_{nar}$分别表示自回归编解码语言模型和非自回归编解码语言模型的参数。

(1)自回归编解码器

在我们的方法中，我们采用自回归编解码语言建模来从语音样本的第一个量化器生成离散代码。公式为：

$$p(c^y_{:,1}|c^x_{:,1}, p, z^{(p)}; \theta_{ar}) = \prod_{t=0}^{T} p(c^y_{t,1}|c^y_{1:t-1}, c^x_{:,1}, p, z^{(p)}; \theta_{ar}) \quad (1)$$

其中$c_{:,1}$表示来自第一个量化器的离散代码，x表示语音样本，p为x对应的音素转换。通过利用自回归建模，我们可以有效地处理韵律建模过程中语音长度的变化。我们通过使用基于Transformer的编码器-解码器架构来实现自回归编解码语言模型。编码器输入是$c^x_{:,1}$和p和$z^{(p)}$的串联。

(2)非自回归编解码器

我们采用非自回归编解码语言建模来从语音样本的第2-8个量化器生成剩余的离散代码。公式为：

$$p(c^y_{:,j}|c^y_{:,j-1}, c^x_{:,j}, p, z^{(p)}; \theta_{nar}), j \in [2, 8] \quad (2)$$

其中$c_{:,j}$表示来自j的离散代码，j为第j个量化器，θ_{nar}表示非自回归编解码语言模型的参数。我们同样采用基于Transformer的编码器-解码器模型作为我们的非自回归编解码语言模型，除了解码器的自注意力掩码之外，它与我们的自回归编解码器语言模型的架构类似。我们采用$c^x_{:,j}$和p和$z^{(p)}$的串联作为编码器输入。

自回归模型和非自回归模型的结合能够使合成的语音在质量和推理速度之间取得一个良好的平衡。一方面如果只用非自回归的话，模型不知道合成语音的长度应该多长，而由于自回归模型的输入序列最后有特殊的$< eos >$标记，因此自回归模型对声学序列长度的预测具有灵活性，使用自回归模型能得到语音的长度。由于第一层量化器已经得到语音长度了，那后面的量化器就可以只用非自回归，因为非自回归模型的推理速度优于自回归模型，这样就把时间复杂度从O(T)降到了O(1)。

3.4 音频合成

给定经过速度控制的语音样本$x = (x_1, x_2, ..., x_n)$及其对应的音素转换$p = (p_1, p_2, ..., p_n)$和音调嵌入序列$z^{(p)} = (z_1^{(p)}, ..., z_T^{(p)})$。首先，神经音频编解码器的声学量化器将每个语音样本编码

为离散声学代码$c^x = E_{codec}(x)$，其中$c^x \in \mathcal{N}^{n^2 \times 8}$,$n^2$表示下采样的语音长度。然后，音频合成的过程可以看作是使用多项式采样策略对

$$c^{\hat{y}} \sim p(c^y | c^x, p, z^{(p)}; \theta) \tag{3}$$

进行采样，其中θ表示条件编解码语言模型的参数，$c^{\hat{y}}$是生成的离散代码序列。最后，我们再使用神经编解码器的解码器合成语音波形$\hat{y}$:

$$\hat{y} = D_{codec}(c^{\hat{y}}) \tag{4}$$

4 实验

4.1 实验设置

为了实验的公平，所有语音数据均采样至24kHz。本实验分别在英语数据集和老挝语数据集上进行训练。在训练英语数据集时，自回归编解码器epoch轮数设置为20，非自回归编解码器epoch轮数设置为40。在训练老挝语数据集时，自回归编解码器epoch轮数设置为15，训练非自回归编解码器epoch轮数设置为30。Batch size大小随机设定，初始学习率设置为0.05，实验设备为ubuntu22.04，CUDA版本为11.7，python版本为3.10.14，torch版本为1.13.1。

4.2 数据集

我们分别在公开的英语数据集LibriTTS(Zen et al., 2019)和私有的老挝语数据集上进行训练。LibriTTS中包含大约585小时、2456个说话者的阅读英语语音。LibriTTS语料库专为语音合成研究而设计，语音都有其对应的文本，且排除了具有显着背景噪声的语音。私有的老挝语数据集包含大约50小时、380个说话者的老挝语语音，老挝语的语音由母语为老挝语的人士进行录制，每一条语音同样有其对应的文本，其训练集和验证集的大小比为4：1。

4.3 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：平均意见得分(MOS,Mean Opinion Score)，又称平均满意度得分，是一种用于评估音频自然度的指标。MOS评分采取 5 个级别对被测音频的质量进行评价，1分代表很差，2分差，3分一般，4分好，5分非常好，待测音频的得分是在所有试听人员的评分上求平均得到的。指标二：比较平均选项得分(CMOS,Comparison Mean Option Score)，在CMOS测试中，试听人员每次听两个音频，并使用一个分数来评估后者与前者相比的感觉。CMOS 的范围从-3（我们的模型比基准模型差很多）到3（我们的模型比基线模型好很多），间隔为1。

客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：梅尔倒谱失真(MCD,Mel Cepstral Distortion)，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，MCD的值越低代表合成音频与真实音频之间的偏差越小，即模型的效果越好。指标二：基频F0的均方根误差(RMSE)，RMSE的值越低表示合成音频与真实音频的基频轮廓越接近，效果越好。

4.4 对比实验

为了证明本文所提方法对韵律可控的同时，也能保证语音的自然度与流畅度，本文与以下不同语音合成模型进行对比。为了验证本文方法的有效性，分别基于LibriTTS数据集和老挝语数据集，设计本文方法与韵律相关的语音合成模型进行比较的不同实验。其中，GST-Tacotron(Wang et al., 2018)与VAE-Tacotron(Zhang et al., 2019)都是仅采用文本和参考音频作为输入的模型。GST-Tacotron使用了一种“全局风格标记”的方法来实现韵律建模。VAE-Tacotron引入VAE以无监督方式来学习韵律的潜在表达，以便于韵律控制。VITS(Kim et al., 2021)是一种结合变分推理、标准化流和对抗训练的高表现力语音合成模型。VITS在隐变量上进行随机建模并利用随机时长预测器，输入同样的文本，能够合成不同韵律的语音。VALL-E(Wang et al., 2023)是本文的基准模型。

对上述几种模型合成的语音进行主观评估，进行MOS和CMOS评分。对实验结果进行客观评估，为了方便使用MCD和RMSE评估，本文所提方法合成的语音内容与基准模型合成的语音内容一致，均方根误差(RMSE)使用基频F0进行计算。英语的实验结果如表1所示，老挝语的实验结果如表2所示，其中Ground Truth为数据集中真实的音频。

模型	MCD	RMSE	MOS	CMOS
GST-Tacotron	7.77	57.36	3.82(±0.10)	-1.32
VAE-Tacotron	7.91	54.88	3.89(±0.10)	-0.80
VITS	7.40	50.83	3.93(±0.06)	-0.41
VALL-E	7.25	47.39	3.96(±0.05)	-0.30
Our Model	**7.13**	**45.21**	**4.05(±0.07)**	**0.00**
Ground Truth	-	-	4.50(±0.03)	+0.25

表 1: 英语数据集上的各项评价指标得分

分析表1结果可知，本文所提方法合成语音的MOS评分达到4.05，相比基准模型VALL-E提升了0.09，并且我们提出的方法以+0.30的CMOS优于基准模型。同时可以看出本文提出的模型在MCD指标上相比基准模型VALL-E降低了0.12，在RMSE指标上相比基准模型降低了2.18。我们的方法在4项评价指标上均是最优的。由于GST-Tacotron和VAE-Tacotron在使用较短的参考音频时无法准确复制参考音频的说话人风格，而我们的方法基于神经编解码语言模型，它只需要3秒的参考音频，就能合成出与参考音频说话风格相似的语音。因此，在参考音频较短的情况下，我们的方法合成的音频相比GST-Tacotron和VAE-Tacotron与真实音频偏差更小、更接近，听觉上更加良好。VITS采用基于随机数的时长预测器，形成随机噪声，但随机噪声无法很好地模拟真实的韵律分布，导致其在听感上不如我们的方法。VALL-E能很好地学习参考音频的音色，但其缺乏对声音的韵律建模。根据主观评价指标和客观评价指标结果，表明我们提出的方法可以根据基准模型合成更自然、更真实的语音。

模型	MCD	RMSE	MOS	CMOS
VITS	7.78	58.26	3.52(±0.10)	-0.30
VALL-E	7.66	**52.15**	3.58(±0.12)	-0.14
Our Model	**7.52**	53.81	**3.61(±0.15)**	**0.00**
Ground Truth	-	-	4.45(±0.04)	+0.40

表 2: 老挝语数据集上的各评价指标得分

分析表2结果可知，本文所提方法合成的老挝语语音的MOS评分达到3.61，相比基准模型VALL-E提升了0.03，并且我们提出的方法以+0.14的CMOS优于基准模型。同时可以看出本文提出的模型在MCD指标上相比基准模型VALL-E降低了0.14。虽然在客观评价指标RMSE上结果不如基准模型，但差距并不大，这是可以接受的。VITS采用基于随机数的时长预测器，形成随机噪声，但随机噪声无法很好地模拟真实的韵律分布，导致其在听感上不如我们的方法。VALL-E能很好地学习参考音频的音色，但其缺乏对声音的韵律建模。总体来说，对于老挝语而言，所提出的方法仍可以根据基准模型合成更自然、更真实的语音。

考虑到本文所使用的私有老挝语数据集时长只有约50小时，数据量的有限可能无法展现模型的优势和特点，我们将老挝语数据集扩增至100小时并再次进行实验。

数据量	MCD	RMSE	MOS	CMOS
Our Model-50h	7.52	53.81	3.61(±0.15)	0.00
Our Model-100h	7.46	51.94	3.65(±0.21)	+0.18

表 3: 老挝语不同数据集规模上的各评价指标得分

由表3结果可以发现，数据量的增大确实能提升模型的性能，证明了本文方法的在不同数据集规模的有效性。

4.5 可控性测试

为了证明本文所提出方法的可控性，我们向试听人员展示了同一句英文文本的受控音频的合成结果，以及相应的控制曲线图。然后，试听人员对受控的合成音频的音调或速度变化的准确程度进行评估，评分范围为1 到5，其中1 表示“完全不准确”，3 表示“准确”，5“非常准确”。

每一轮包含同一合成文本的速度控制、音调控制和音调+速度控制，每位试听人员均进行6轮测试。

		速度	音调
仅速度控制	真实曲线	3.30±0.13	-
仅速度控制	虚假曲线	3.16±0.10	-
仅音调控制	真实曲线	-	**3.14±0.21**
仅音调控制	虚假曲线	-	2.95±0.15
速度+音调控制	真实曲线	**3.34±0.08**	3.06±0.15

表 4: 95%置信区间下可控性测试的MOS结果

在表4中，对于单因素(仅速度或仅音调)控制，所呈现的控制曲线有15%的可能性是虚假曲线，虚假曲线即真实曲线的翻转或移位版本。将向试听人员展示虚假曲线后得到的评分作为基线评分，对比真实曲线的评分可得，在同时进行速度和音调控制时，速度变化最准确，与基线评分相比提升0.18；在进行单因素音调控制时，音调变化最准确，与基线评分相比提升0.19。这是由于评测的时候存在一些较短的音频，仅有1-2秒的时长，试听人员可能更难评估音调的可控性。在测试中，速度控制采用上升模式，音频速度会随时间逐渐变快。音调控制采用stress模式，选取一段音频区间作为音调上升的部分，因此对于短音频，音调上升部分很短，尤其在速度变化的对比下，音调在听感上变化更加不明显。而单音调控制时，没有速度的变化，试听者在听感上更能察觉音调的变化。这就导致在速度音调同时控制时，音调评分略低于仅音调控制，但仍高于基线评分。这表明我们的方法能够同时控制这两个因素，即实现了韵律的可控性。

4.6 定性分析

给定两个老挝语文本和一段录音，我们分别在基准模型VALL-E和我们的模型上运行推理过程两次，并在图4和图5中可视化其波形。

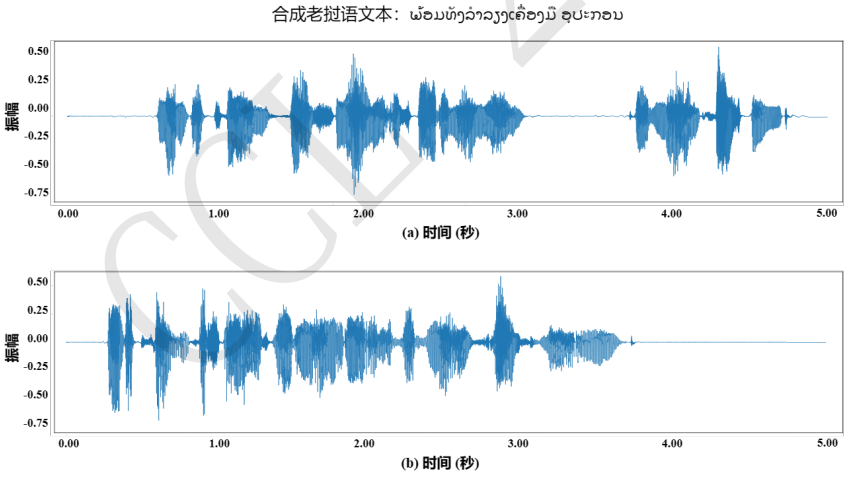

图 4: 老挝语文本1合成的声音波形

在图4中，我们观察到两个样本具有不同的持续时间。这是因为4(a)是在基准模型VALL-E上合成的样本，没有速度控制，而4(b)是经过抛物线模式的速度控制曲线修改后的样本。可以观察到，速度的差异性导致两个样本即使合成的是相同的老挝语文本，仍具有不同的持续时间。

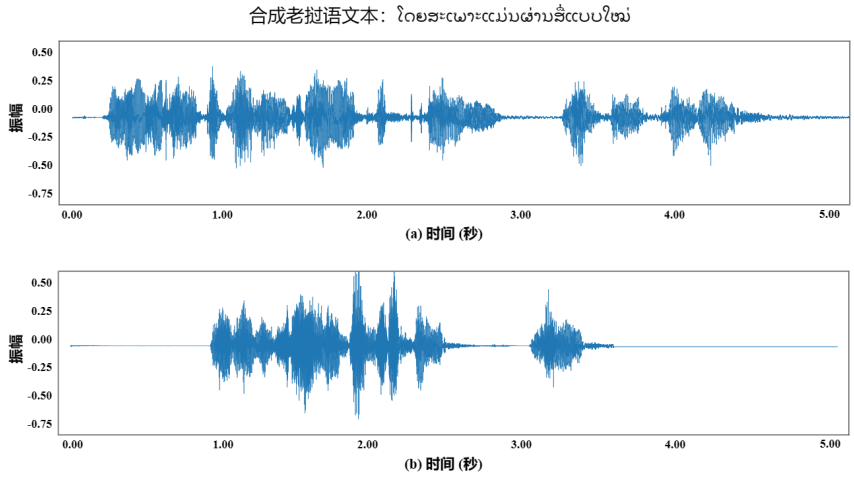

图 5: 老挝语文本2合成的声音波形

在图5中，我们观察到两个样本的振动的频率不同。类似的，5(a)是在基准模型VALL-E上合成的样本，没有音调控制，而5(b)是经过stress模式的音调控制曲线修改后的样本。可以观察到，音调的差异性导致两个样本即使合成的是相同的老挝语文本，仍具有不同的振动频率。

5 结论

针对现有方法中合成的语音仍然缺乏真实语音的多样性和韵律变化的丰富性的问题，本文提出基于神经编解码语言模型的老挝语韵律建模方法。利用先验速度和音调时序变化曲线建模韵律变化分布，实现对语言特征的时变控制，并且在训练中使用自回归架构和非自回归架构相结合的方式。在公开英语数据集LibriTTS和私有老挝语数据集上，主观和客观评价结果都证明，本文提出的方法能够在合成高质量音频的同时进行速度和音调控制。

参考文献

Anh, Nguyen Thi Ngoc, Thanh, Nguyen Tien, and others. 2022. *Development of a high quality text to speech system for Lao*, volume 1. IEEE, 2022. 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.

Borsos, Zalán, Marinier, Raphaël, Vincent, Damien, Kharitonov, Eugene, Pietquin, Olivier, Sharifi, Matt, Roblek, Dominik, Teboul, Olivier, Grangier, David, Tagliasacchi, Marco, and others. 2023. *Audiolm: a language modeling approach to audio generation*, volume 1. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. IEEE.

Charpentier, Francis, and Stella, M. 1986. *Diphone synthesis using an overlap-add technique for speech waveforms concatenation*, volume 11. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2015–2018. IEEE.

Chen, Meiying, and Duan, Zhiyao. 2022. *ControlVC: Zero-shot voice conversion with time-varying controls on pitch and speed*. arXiv preprint arXiv:2209.11866.

Défossez, Alexandre, Copet, Jade, Synnaeve, Gabriel, and Adi, Yossi. 2022. *High fidelity neural audio compression*, volume 1. arXiv preprint arXiv:2210.13438.

Dhariwal, Prafulla, Jun, Heewoo, Payne, Christine, Kim, Jong Wook, Radford, Alec, and Sutskever, Ilya. 2020. *Jukebox: A generative model for music*, volume 1. arXiv preprint arXiv:2005.00341.

Kasi, Kavita, and Zahorian, Stephen A. 2002. *Yet another algorithm for pitch tracking*, volume 1. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–361. IEEE.

Kim, Daegyeom, Hong, Seongho, and Choi, Yong-Hoon. 2023. *SC VALL-E: Style-Controllable Zero-Shot Text to Speech Synthesizer*, volume 1. arXiv preprint arXiv:2307.10550.

Kim, Jaehyeon, Kong, Jungil, and Son, Juhee. 2021. *Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech*, volume 1. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Klimkov, Viacheslav, Ronanki, Srikanth, Rohnke, Jonas, and Drugman, Thomas. 2019. *Fine-grained robust prosody transfer for single-speaker neural text-to-speech*, volume 1. arXiv preprint arXiv:1907.02479.

Lee, Younggun, and Kim, Taesu. 2019. *Robust and fine-grained prosody control of end-to-end speech synthesis*, volume 1. IEEE, 2019. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5911–5915.

Mortensen, David R, Dalmia, Siddharth, and Littell, Patrick. 2018. *Epitran: Precision G2P for many languages*. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Skerry-Ryan, RJ, Battenberg, Eric, Xiao, Ying, Wang, Yuxuan, Stanton, Daisy, Shor, Joel, Weiss, Ron, Clark, Rob, and Saurous, Rif A. 2018. *Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron*, volume 1. PMLR, 2018. International Conference on Machine Learning, pages 4693–4702.

Wang, Chengyi, Chen, Sanyuan, Wu, Yu, Zhang, Ziqiang, Zhou, Long, Liu, Shujie, Chen, Zhuo, Liu, Yanqing, Wang, Huaming, Li, Jinyu, and others. 2023. *Neural codec language models are zero-shot text to speech synthesizers*, volume 1. arXiv preprint arXiv:2301.02111.

Wang, Yuxuan, Stanton, Daisy, Zhang, Yu, Ryan, RJ-Skerry, Battenberg, Eric, Shor, Joel, Xiao, Ying, Jia, Ye, Ren, Fei, and Saurous, Rif A. 2018. *Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis*, volume 1. PMLR, 2018. International Conference on Machine Learning, pages 5180–5189.

Zen, Heiga, Dang, Viet, Clark, Rob, Zhang, Yu, Weiss, Ron J, Jia, Ye, Chen, Zhifeng, and Wu, Yonghui. 2019. *Libritts: A corpus derived from librispeech for text-to-speech*. arXiv preprint arXiv:1904.02882.

Zhang, Ya-Jie, Pan, Shifeng, He, Lei, and Ling, Zhen-Hua. 2019. *Learning latent representations for style control and transfer in end-to-end speech synthesis*, volume 1. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE.