计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700138

基于音素大语言模型及扩散模型的低资源越南语语音合成

邹 睿 杨 鉴 张 凯
云南大学信息学院 昆明 650504
(june_ouui@163.com)

摘 要 随着深度学习技术的发展及语音合成研究的深入,汉语、英语等通用、高资源语言的合成语音已越来越接近于自然语音。越南语与汉语有密切联系,是一种声调语言,属于南亚语系越芒语族越语支。因受制于可获取的语料数据规模以及相关研究的深入程度,越南语语音合成离自然语音还有明显差距。在低资源前提下,提出了两种提高越南语语音合成自然度的方法:1)基于预训练的音素大语言模型 XPhoneBERT 构建音素编码器,在数据集有限的情况下,显著提高越南语语音合成的韵律表现力;2)改进轻量化扩散语音合成模型 LightGrad 中的 U-Net 结构,增加嵌套跳跃路径,使模型在低资源条件下得到充分训练、捕获更有效的信息、提高噪声预测的准确性,从而提升语音合成质量。实验结果表明,采用上述提出的方法,越南语语音合成系统的客观、主观评测性能有明显的提升,MCD(梅尔倒谱失真)和 MOS(平均意见得分)分别达到 6.25 和 4.22,相比于基线系统的 7.44 和 3.56 有明显的下降和提升。

关键词:语音合成;越南语;低资源;大语言模型;扩散模型
中图分类号 TP391

Low-resource Vietnamese Speech Synthesis Based on Phoneme Large Language Model and Diffusion Model

ZOU Rui, YANG Jian and ZHANG Kai
School of Information Science & Engineering, Yunnan University, Kunming 650504, China

Abstract With the development of deep learning technology and the progression of speech synthesis research, synthetic speech in widely spoken and high-resource languages such as Chinese and English has increasingly approached natural speech. Vietnamese, a tonal language closely related to Chinese, belongs to the Vietic branch of the Austroasiatic language family of South Asian languages. Due to the scale of available corpus data and the depth of related research, Vietnamese speech synthesis is still significantly short of natural speech. At the premise of low resources, two methods are proposed to improve the naturalness of Vietnamese speech synthesis: 1) The phoneme encoder is constructed based on pre-trained phoneme large language model XPhoneBERT, which significantly improves the prosodic expressiveness of Vietnamese speech synthesis with limited data set. 2) Improve the U-Net structure in the lightweight diffusion TTS model LightGrad, add nested jump paths, so that the model can be fully trained under low resource conditions, capture more effective information, improve the accuracy of noise prediction, and thus improve the quality of speech synthesis. Experiment results show that the objective and subjective evaluation performance of the Vietnamese speech synthesis system has been significantly improved by using the proposed method. MCD and MOS are up to 6.25 and 4.22 respectively, which are significantly decreased and increased respectively, compared with 7.44 and 3.56 of the baseline system.

Keywords Speech synthesis, Vietnamese, Low resources, Large language model, Diffusion model

1 引言

语音合成(Text-to-speech,TTS)是人工智能的重要分支之一,被广泛应用于各个场景,如智能家居、车载导航、有声阅读、资讯播报等。随着深度学习的快速发展,基于神经网络的语音合成系统取得了显著的成果,例如 Tacotron 1/2[1-2],Deep Voice 3[3],FastSpeech 1/2[4-5]等,这些端到端模型简化了文本分析模块,直接将字符/音素序列作为输入,并使用 Mel 谱图简化声学特征。后来又开发了完全端到端的语音合成系统,直接从文本生成波形,如 ClariNet[6],FastSpeech 2s[5]和 EATS[7]。与传统的基于串联合成和统计参数合成的语音合成系统相比,基于神经网络的语音合成系统所合成的语音在清晰度和自然度方面有了很大的提升。

扩散概率模型(Diffusion Probabilistic Models,DPMs)[8]是一类新的生成模型,展现出强大的生成能力。相较于 GANs[9]和基于流[10]的生成模型,扩散模型有更好的稳定性、可控性和多样性,具有更高的参数效率,还体现出了数学范式的重要性。随着扩散概率模型的推广,语音合成系统有了新的进展。Grad-TTS[11]是一种基于扩散模型的语音合成系统,主要包含编码器、持续时间预测器和基于 U-Net 的解码

基金项目:国家重点研发计划资助项目(2020AAA0107901)
This work was supported by the National Key Research and Development Program(2020AAA0107901).
通信作者:杨鉴(jianyang@ynu.edu.cn)

器，只需 10 次反向扩散迭代就能生成高质量的 Mel 谱图，在 GPU 上的速度超过了 Tacotron 2。LightGrad[12] 在 Grad-TTS 的基础上作了改进，提出了轻量级的 U-Net 解码器，推理过程中采用 DPM-Solver[13] 快速求解器，并实现了流式推理，减少了模型参数，加快了推理过程。

虽然基于神经网络的语音合成系统已经展现出良好的性能，但其仍需要大量的数据集进行训练。目前只有汉语、英语等主流语言的语音合成技术已经迈向相对成熟的阶段，低资源语言的语音合成研究仍然有限，并且在商业化语音服务中的应用较少。越南语又称京语或国语，与汉语有密切联系，是一种声调语言，属于南亚语系越芒语族越语支，是越南的官方语言。随着经济全球化的发展，对非通用语的语音合成需求增多，研究针对越南语的语音合成技术有着十分重要的意义。因为创建数据集的成本很高，现有的公开越南语数据集资源相对匮乏，并且越南语的语音韵律和语调较为复杂，因此在语音合成系统中往往难以得到高质量的越南语语音。

语音的节奏、重音、语调等韵律信息对应着音节时长、响度、音高的变化，在人类言语交际中起着重要的感知作用。越南语使用拉丁字母书写，共计 29 个字母，有 9 个变音符号，其中 4 个符号用于添加元音，5 个符号用于表示声调。相较于汉语，越南语声调变化更复杂，总共包含 6 个声调，分别为平声、玄声、问声、跌声、锐声、重声，平声无符号表示。越南语的音节由 3 个部分组成：开头的辅音、不可或缺的主要元音和声调。每个音素、每个声调发音准确与否都会直接影响语义的表达，为了合成更准确、自然的越南语语音，韵律的支持尤为重要。但在低资源语音合成中，低资源数据集往往难以让大型的深度学习模型得到充分的训练，导致模型无法完全掌握语音的细节和多样化的韵律特征，从而影响合成语音的可懂度和自然度，该问题通常采用数据增强[14]、迁移学习[15]的方法来解决，但存在一些局限性，可能会引入一些噪声的变化，或因不同迁移数据集间的语言差异，导致合成语音失真或不自然。越南语因复杂的语音韵律和语调，在低资源语音合成研究中更具挑战性，Lam 等[16] 采用基于实例的迁移学习来构建越南语语音合成系统，减少了训练所需数据量，但合成语音在韵律方面的表现仍有提升空间；Phung 等[17] 通过插入韵律标注来预处理数据，该方法会增加人工标注成本。

为弥补上述方法的不足，本文提出了一个基于 LightGrad 和音素级语言大模型的越南语语音合成系统，主要目的是在低资源条件下合成高自然度和可懂度的越南语语音，增强韵律表达。本文所做的主要工作如下：1）针对合成越南语语音韵律表现力不足的问题，使用预训练大语言模型 XPhoneBERT[18] 作为音素编码器，在数据集有限的情况下提升系统在韵律方面的性能；2）改进基线系统中的 U-Net 结构，增加嵌套跳跃路径[19]，在特征提取时捕获更多细粒度细节，提高系统的准确性，从而合成高质量的语音。

2 相关理论及模型

2.1 扩散概率模型

扩散概率模型是一种参数化马尔可夫链，被训练用来生成与真实数据匹配的样本，包含正向扩散过程和反向扩散过程。扩散模型的基本思想是：在正向扩散过程中通过迭代破坏原始数据，将高斯噪声添加到真实数据中，最终将数据分布转换为高斯分布，反向扩散过程也是去噪的过程，逐渐生成与原始数据分布相似的新样本，这一步骤通常需要使用可训练神经网络来实现。

2.1.1 正向扩散

用 X_0 表示从数据集中采样得到的原始样本，用 X_T 表示经过 T 步加噪之后得到的近似纯噪声的样本。扩散模型的正向扩散过程创建了一个随机过程 $\{X_t\}_{t=0}^{T}$，可由随机微分方程（Stochastic Differential Equation，SDE）描述为：

$$\mathrm{d}X_t=\frac{1}{2}(\mu-X_t)\beta_t\mathrm{d}t+\sqrt{\beta_t}\,\mathrm{d}W_t,\ t\in[0,T] \tag{1}$$

其中，μ 是高斯先验分布 $N(\mu,I)$ 的均值；β_t 被称为噪声调度，是非负函数；W_t 为布朗运动。

2.1.2 反向扩散

反向扩散的轨迹与正向扩散的轨迹密切相关，但时间顺序相反，相当于对 X_T 进行反向求解来恢复原始数据分布 X_0，反向扩散的 SDE 可被描述为：

$$\mathrm{d}X_t=\left(\frac{1}{2}(\mu-X_t)-\nabla\log p_t\right)\beta_t\mathrm{d}t+\sqrt{\beta_t}\,\mathrm{d}\widetilde{W}_t \tag{2}$$

其中，$\widetilde{W}_t$ 为逆时布朗运动，p_t 为随机变量 X_t 的概率密度函数。此外，Song 等[20] 表明还可用常微分方程（Ordinary Differential Equation，ODE）来求解反向扩散，可被描述为：

$$\mathrm{d}X_t=\frac{1}{2}((\mu-X_t)-\nabla\log p_t)\beta_t\mathrm{d}t \tag{3}$$

对于式(2)和式(3)中的 $\nabla\log p_t(t\in[0,T])$，可以用一个神经网络 $s_\theta(X_t,t)$ 来估计，通常被称为分数匹配。

2.1.3 损失函数

训练神经网络 $s_\theta(X_t,t)$ 时使用的扩散损失可以参考 $L2$ 损失，其表达式为：

$$\mathcal{L}_{\mathrm{diff}}=\mathbb{E}_{X_0,t}[\mathbb{E}_{\epsilon_t}\|\sqrt{\textstyle\sum_t}s_\theta(X_t,\mu,t)+\epsilon_t\|^2] \tag{4}$$

2.2 LightGrad 模型

非自回归 TTS 模型 LightGrad 是一个轻量级扩散模型，其推理过程如图 1 所示。LightGrad 包含编码器、利用单调对齐搜索（Monotonic Alignment Search，MAS）训练的持续时间预测器、轻量级 U-Net 解码器和 HiFi-GAN[21] 声码器。输入的文本序列经过编码器被转换为特征序列，与持续时间预测器的输出进行对齐，再通过逐步加噪将特征序列转换成噪声序列，最后由解码器逐步去噪，将高斯噪声重建为高质量的 Mel 谱图，再由声码器转换成波形。其中，编码器由 1 个 prenet，6 个多头自注意力的 Transformer 块以及最后的线性投影层组成。持续时间预测器由两个卷积层和一个预测持续时间对数的投影层组成。轻量级 U-Net 解码器使用深度可分离卷积[22] 来减少模型参数和计算量。

在推理过程中，LightGrad 使用无需训练的 DPM-Solver-1 作为扩散模型的求解器，以加快推理过程。此外，LightGrad 利用流式推理[23] 来减少运行时占用的内存，进一步降低推理延迟。

基于以上轻量化的结构，LightGrad 将模型参数量缩减为 5.61×10^6，远低于其基线系统 Grad-TTS 的参数量 14.85×10^6，同比降低了 62.2%，同时保证了生成语音的质量。对低资源的数据集而言，在轻量化的 U-Net 解码器网络中更容易训练至收敛。本文以 LightGrad 为基线系统，目的是在低资源的越南语数据集条件下获得更好的训练效果，有效减小训

练不充分对实验结果造成的影响。

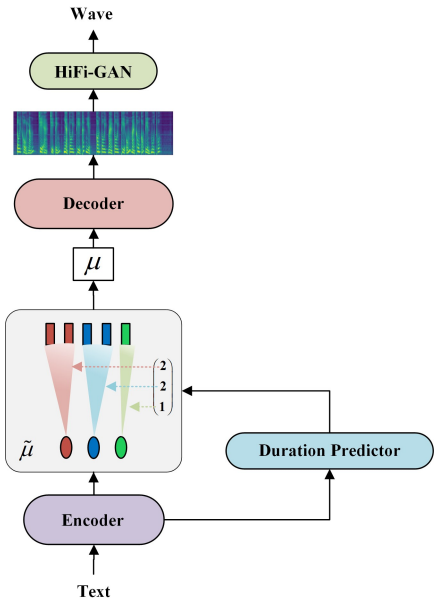


图 1　LightGrad 推理过程

Fig. 1　LightGrad inference process

3　越南语语音合成方法

本文基于 LightGrad 提出的越南语语音合成系统架构如图 2 所示,包含 XPhoneBERT 音素编码器、持续时间预测器、改进的 U-Net 和预训练的 HiFi-GAN 声码器。该系统以原始文本为输入,使用越南语 G2P(Grapheme-to-Phoneme)模型将输入文本转换成音素序列。输出的音素序列通过预训练的 XPhoneBERT 编码器被编码为特征序列,可以提高语音合成系统在自然度和韵律方面的性能。在使用解码器进行噪声预测时,改进的 U-Net 通过增加跳跃连接路径获得有效的信息,以更准确地预测噪声,从而合成更准确的越南语语音。

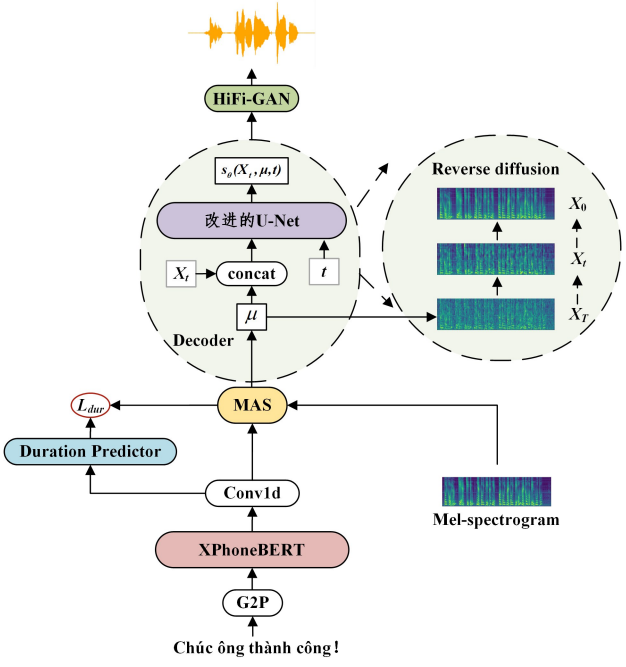


图 2　越南语语音合成系统训练过程

Fig. 2　Training process of Vietnamese speech synthesis system

3.1　XPhoneBERT 音素编码器的引入

韵律是影响合成语音质量的一个重要因素,不仅影响自然度,还会影响听者对语音的感知和理解。通常情况下,在语音合成中,如果训练数据充足,则可以通过大量的训练让语音合成系统学习到原始数据中的韵律表示,从而合成高自然度和可懂度的语音,但这往往对实验条件要求很高。当训练数据有限时,合成的语音又难以达到令人满意的效果。在使用低资源越南语数据集的条件下,为了提高系统合成语音的韵律表现力,本文采用大规模预训练的多语言音素表示模型 XPhoneBERT 作为音素编码器。

大语言模型 BERT 凭借双向上下文表示学习和预训练微调框架,提供了强大的语言理解和生成能力,在多领域展现出了良好的性能。目前使用越南语数据集进行预训练的大语言模型并不多见,输入一般为字符,且一般用于词性标注、依赖解析、命名实体识别等自然语言处理任务,而 XPhoneBERT 的输入为音素,能用于学习下游语音合成任务的音素表示,可作为音素编码器将音素序列转换为特征序列,并且用多语言数据集进行大规模的预训练,能更准确地提取语言特征。XPhoneBERT 与 BERT-base[24] 的模型架构一致,是基于 Transformer 的双向编码器,能更好地捕捉上下文语义,并且多层双向结构能更全面地捕获音素之间的长距离依赖关系,有助于生成自然流畅的韵律。XPhoneBERT 是作为一个模型单独预训练的,其预训练方法与 RoBERTa[25] 一致,使用动态屏蔽策略,在预训练期间不断调整被屏蔽的音素位置,使模型学习更稳健的音素表征。预训练时使用的多语言语料库中包含 94 种语言的 3.3 亿个音素级句子,越南语有 12 300 句。在预训练之前,首先对数据集文本的词和句子进行分割、重复删除以及文本规范化处理,再将文本转为音素,并分割音素。

在本文提出的语音合成系统前端,首先使用越南语的 G2P 模型将越南语文本转换成音素序列,并将已预训练的 XPhoneBERT 作为编码器,直接对音素序列进行编码,输出特征序列。此外,为了提高 XPhoneBERT 模型在越南语单语种上的韵律表现力,实验过程中冻结了模型前九层的参数,用低资源越南语数据集对后三层的参数进行微调。使用预训练的 XPhoneBERT 能间接地为音素表示提供额外的上下文信息,更好地捕捉长距离依赖关系,这使得越南语语音合成系统能更好地学习原始音频中的韵律表达,以增强合成语音的韵律,并且有助于在数据集资源有限的条件下,生成韵律更自然的语音。

3.2　U-Net 解码器的改进

在扩散模型中,一般使用 U-Net 进行噪声预测。U-Net 包含编码器、解码器和跳跃连接,下采样用于对输入信息进行压缩编码、提取特征、上采样,以恢复出期望输出的噪声尺度,跳跃连接可以让下采样层的输出特征融合到同层的上采样层中。UNet++[19] 是一种基于嵌套和密集跳跃连接的新架构,其通过一系列嵌套的密集卷积块将编码器和解码器连接在一起。与常用的普通跳跃连接不同,嵌套跳跃连接能更好地融合编码器和解码器子网络之间的不同特征,减少两者特征映射之间的语义差距,能满足在训练过程中更精确地提取特征的需求。

为了在训练时更精确地提取特征，增强解码器的去噪能力，本文基于 UNet++对基线系统中的 U-Net 作了改进，结构如图 3 所示，其中下采样块、中间块、上采样块和最终卷积块保持其原有数量和结构不变。图中红色部分是在原普通跳跃连接的基础上增加的嵌套跳跃路径，在$X^{1,0}$和$X^{1,2}$中间添加了卷积块$X^{1,1}$，$X^{1,1}$由两层 SepResBlock 和一层 LA Layer 组成。SepResBlock 包括两个 2D 的深度可分离卷积(SepConv2d)、群归一化和 Mish 激活函数[26]，并通过使用一个额外的线性层加入步长嵌入。LA Layer 用线性自注意[27]对输入序列进行处理，其复杂度是输入序列长度的线性函数。$X^{1,1}$的输入由$X^{1,0}$的输出和$X^{2,0}$上采样的输出拼接而成，输入通道数大小为 256。$X^{0,1}$只包含两个 2D 卷积和一个 GroupNorm，其作用是将输出结果转变为 Mel 谱图的维度。

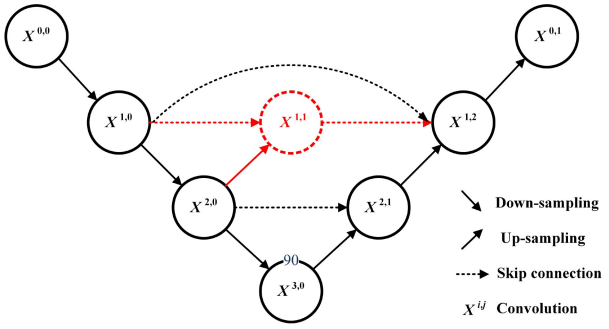


图 3 改进的 U-Net 结构图(电子版为彩图)

Fig. 3 Diagram of improved U-Net structure

改进的 U-Net 可以优化训练过程，更好地处理学习任务。增加的嵌套跳跃路径改变了编码器与解码器子网络的连通性，能够融合不同尺度的特征，使来自解码器和编码器网络的特征映射在语义上更相似。该方法同时使用长连接和短连接，使模型在低资源条件下得到充分训练，更有效地捕获细粒度细节，在训练时能更准确地提取特征。改进的 U-Net 能预测更准确的噪声，让去噪生成的 Mel 谱更接近真实的 Mel 谱图，从而生成更高质量的语音。

4 实验

4.1 实验数据

实验中使用的越南语音频数据由一位专业的女性播音员录制，总时长约为 6.9 h，共包含 3996 个音频片段，均为单声道，采样率为 48 kHz，16 位 PCM 编码，音频前后各保留 50 ms 的静音段。用于训练、验证和测试的样本数量分别为 3596，200，200 语句。

4.2 实验设置

本文设计的越南语语音合成系统包含 XPhoneBERT 音素编码器、持续时间预测器、改进的 U-Net 和预训练的 HiFi-GAN 声码器，持续时间预测器和声码器的结构与基线系统 LightGrad 相同。实验参数设置参照 LightGrad 的设置，在单个 GPU 上训练的迭代次数增加至 2×10^6，批大小同为 16，优化器选用 Adam，学习率设置为 0.0001，前向过程 T 设置为 1，使用与 LightGrad 相同的噪声调度。推理过程中，去噪步骤设置为 4 步，采用流式推理，每次生成 0.5 s 的 Mel 谱图块，最后由 HiFi-GAN 声码器将 Mel 谱图转换为高保真度的语音波形。XPhoneBERT 文本编码器中 Transformer 块的数量为 12，隐藏大小为 768，自注意头的数量为 12。

为验证本文提出的越南语语音合成系统的有效性，设置了如表 1 所列的 4 组实验，另外，加入 XPhoneBERT 后，特别设置了 Exp 3 和 Exp 4 来对比微调前后的结果，以验证微调的必要性。

表 1 越南语语音合成实验设置

Table 1 Experimental settings of Vietnamese speech synthesis

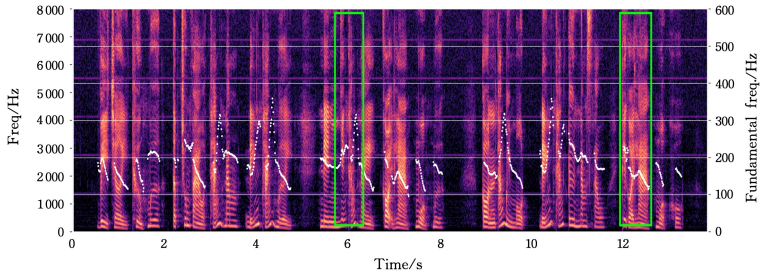
实验名称	实验方法
Exp 1	基线系统
Exp 2	基线系统＋改进的 U-Net
Exp 3	基线系统＋改进的 U-Net＋XPhoneBERT(冻结)
Exp 4	基线系统＋改进的 U-Net＋XPhoneBERT(微调)

本文采用客观评价指标梅尔倒谱失真(Mel Cepstrum Distortion，MCD)和主观评价指标平均意见得分(Mean Opinion Score，MOS)来评估合成语音的质量。其中，MCD 是计算两个梅尔倒谱序列的差异大小，MCD 值越小则表示合成语音越接近原始音频。MOS 评分需要试听者依据听感对听到的音频进行打分，分值范围为 1～5 分，包含 5 个等级，一般高于 4 分则认为语音质量较好。本文邀请到 10 位越南语专业的在读硕士进行打分，从测试集中随机选取 15 条合成语音呈现给试听者，同时加入真实音频进行测评，最终结果取平均分。

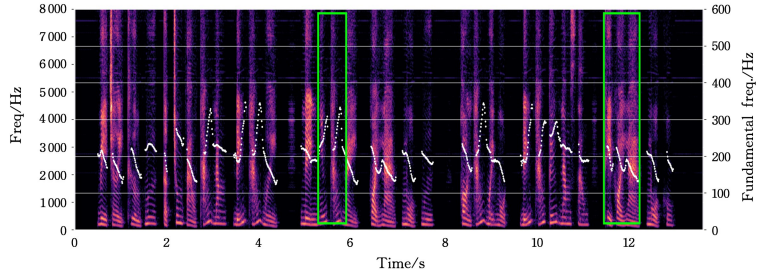
4.3 实验结果与分析

为了清楚观察本文提出的越南语语音合成系统所合成的语音在韵律方面的表现力，实验时随机选取文本“Vì trò'i thu'ò'ng mu'a t ập trung vào tháng năm đến tháng mu'ò'i, nên nh ũ'ng tháng này g ọi là mùa mu'a; còn tháng mu'ò'i môt đến tháng bốn năm sau thì gọi là mùa khô.”，借助 Praat 软件分别绘制其对应原始音频、基线系统合成音频和本文系统合成音频的语谱图和基频曲线，如图 4 所示。由图可知，相较于基线系统合成的音频，本文提出的越南语语音合成系统合成音频的停顿点更少，高频部分的能量损失也更少，表明其在语音流畅度和清晰度方面都优于基线系统合成的音频，更接近原始音频。同时，从图中标注的基频变化可知，在读“nh ũ'ng”和“gọi”时，基线系统合成音频的基频更高，重声和跌声的发音还不够准确，而本文系统合成的音频在音高方面与原始音频差异明显更小。上述情况表明越南语语音合成系统在使用 XPhoneBERT 音素编码器之后，更好地学习到了原始音频中的韵律表达，合成的语音听起来更流畅，声调更准确，由此验证了本文使用的音素编码器在提升合成语音韵律表现力方面的有效性。

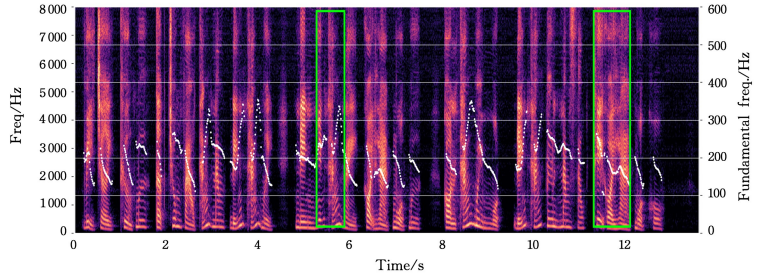
MCD 和 MOS 的评分结果如表 2 所列，由结果可知，本文提出的越南语语音合成系统所合成的语音质量优于基线系统，更接近自然语音。对比 Exp 1 和 Exp 2 的结果可知，改进 U-Net 结构后，相较于基线系统，MCD 值更低，MOS 评分值更高，生成的语音更准确，证明了该结构的有效性。Exp 3 的实验结果表明，在微调前，使用 XPhoneBERT 并未提升系统性能，其原因是多语种数据集预训练的音素编码器，直接用于单语种的音素序列编码时，会导致特征提取不够准确，因此需要微调以适应单语种模型。而 Exp 4 的实验结果证明了微调可以进一步增强系统的韵律表现力，有助于生成更高自然度的语音，也可以证明在数据集有限的条件下，使用 XPhoneBERT 作为音素编码器，能生成韵律更自然的语音。



(a)原始音频



(b)基线系统合成音频



(c)本文提出的越南语语音合成系统合成音频

图 4　语谱图及基频曲线

Fig. 4　Spectrogram and fundamental frequency curve

表 2　实验结果

Table 2　Experiment results

实验方法	MCD(↓)	MOS(↑)
原始音频	—	4.75
Exp 1	7.44	3.56
Exp 2	6.65	3.85
Exp 3	6.71	3.53
Exp 4	6.25	4.22

结束语　本文提出了一个基于 LightGrad 的越南语语音合成系统,在低资源越南语数据集的条件下,使用预训练的大语言模型 XPhoneBERT 作为音素编码器,并通过微调增强了合成语音的韵律表现力。在原 U-Net 结构中增加嵌套跳跃连接,以在特征提取时捕获更多细粒度细节,使模型在低资源条件下得到充分训练,从而更准确地预测噪声。实验结果表明,本文提出的越南语语音合成系统在自然度、准确度和韵律方面有更好的性能,能合成高质量的语音。本文提出的系统仍包含级联的 HiFi-GAN 声码器,下一步工作可以探索端到端的越南语语音合成系统。

参 考 文 献

[1] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards End-to-End Speech Synthesis[C]// Interspeech. 2017.

[2] SHEN J, PANG R, WEISS R J, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2018). IEEE, 2018.

[3] PING W, PENG K W, GIBIANSKY A, et al. Deep voice 3: Scaling text-to-speech with convolutional sequence learning[J]. arXiv: 1710.07654, 2017.

[4] REN Y, RUAN Y, TAN X, et al. FastSpeech: Fast, robust and controllable text to speech[C]// Advances in Neural Information Processing Systems. 2019.

[5] REN Y, HU C, TAN X, et al. FastSpeech 2: Fast and high-quality end-to-end text to speech[J]. arXiv: 2006.04558, 2020.

[6] PENG K W, CHEN J. Clarinet: Parallel wave generation in end-to-end text-to-speech[J]. arXiv: 1807.07281, 2018.

[7] DONAHUE J, DIELEMAN S, BIŃKOWSKI M, et al. End-to-end adversarial text-to-speech[J]. arXiv: 2006.03575, 2020.

[8] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27.

[10] KIM J, KIM S, KONG J, et al. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search[J]. Advances in Neural Information Processing Systems, 2020, 33: 8067-8077.

[11] POPOV V, VOVK I, GOGORYA N, et al. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech[C]// International Conference on Machine Learning(2021).

[12] CHEN J. LightGrad: Lightweight Diffusion Probabilistic Model

for Text-to-Speech[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2023). 2023:1-5.
[13] LU C, ZHOU Y, BAO F, et al. DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps [J]. Advances in Neural Information Processing Systems, 2022, 35:5775-5787.
[14] LIANG Z, SHI H, WANG J, et al. EM-TTS: Efficiently Trained Low-Resource Mongolian Lightweight Text-to-Speech[J]. arXiv:2403.08164, 2024.
[15] JEONGM, KIM M, CHOI B J, et al. Transfer Learning for Low-Resource, Multi-Lingual, and Zero-Shot Multi-Speaker Text-to-Speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.
[16] LAM T Q, et al. Instance-based transfer learning approach for Vietnamese speech synthesis with very low resource[C]// Future of Information and Communication Conference. Cham: Springer International Publishing, 2022.
[17] PHUN V L. Data processing for optimizing naturalness of Vietnamese text-to-speech system[C]// 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques(O-COCOSDA). IEEE, 2020.
[18] NGUYEN L T, THINH P, DAT Q N. XPhoneBERT: A Pre-trained MULTILINGUAL Model for Phoneme Representations for Text-to-Speech[J]. arXiv:2305.19709, 2023.
[19] ZHOU Z, SIDDIQUEE M M R, TAJBAKHSH N, et al. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation [J]. IEEE Transactions on Medical Imaging, 2020, 39(6):1856-1867.
[20] SONG Y, SOHL-DICKSTEIN J, KINGMAD P, et al. Score-based generative modeling through stochastic differential equations[J]. arXiv:2011.13456, 2020.
[21] KONG J, KIM J, BAE J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in Neural Information Processing Systems, 2020, 33: 17022-17033.
[22] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017:1800-1807.
[23] ElLINAS N, VAMVOUKAKIS G, MARKOPOULOS K, et al. High quality streaming speech synthesis with low, sentence-length-independent latency[J]. arXiv:2111.09052, 2021.
[24] DEVLIN J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C] // North American Chapter of the Association for Computational Linguistics. 2019.
[25] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized bert pretraining approach[J]. arXiv:1907.11692, 2019.
[26] MISRA D. Mish: A Self Regularized Non-Monotonic Activation Function[J]. British Machine Vision Conference, 2020.
[27] ZHUORAN S, MINGYUAN Z, HAIYU Z, et al. Efficient Attention: Attention with Linear Complexities[C]// 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA, 2021:3530-3538.

ZOU Rui, born in 2000, postgraduate. Her main research interests include speech synthesis, recognition and understanding.

YANG Jian, born in 1964, Ph.D, professor. His main research interests include speech synthesis, recognition and understanding.