计算机科学
COMPUTER SCIENCE

# 基于跨语言迁移学习及联合训练的泰语语音合成

张欣瑞，杨鉴，王展

引用本文

张欣瑞，杨鉴，王展. 基于跨语言迁移学习及联合训练的泰语语音合成[J]. 计算机科学，2024，51(6A): 230500174-7.

ZHANG Xinrui, YANG Jian, WANG Zhan. Thai Speech Synthesis Based on Cross-language Transfer Learning and Joint Training [J]. Computer Science, 2024, 51(6A): 230500174-7.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500174

# 基于跨语言迁移学习及联合训练的泰语语音合成

张欣瑞　杨　鉴　王　展
云南大学信息学院　昆明 650504
(1090525272@qq.com)

**摘　要**　随着深度学习和神经网络的快速发展,基于深度神经网络的端到端语音合成系统因性能优异成为主流。然而近年来,泰语语音合成相关研究还不充分,主要原因是大规模泰语数据集稀缺且该语言拼写方式有其特殊性。为此,在低资源前提下基于 FastSpeech2 声学模型和 StyleMelGAN 声码器研究泰语语音合成。针对基线系统中存在的问题,提出了 3 个改进方法以进一步提高泰语合成语音的质量。(1)在泰语语言专家指导下,结合泰语语言学相关知识设计泰语 G2P 模型,旨在处理泰语文本中存在的特殊拼写方式;(2)根据所设计的泰语 G2P 模型转换的国际音标表示的音素,选择拥有相似音素输入单元且数据集丰富的语言进行跨语言迁移学习来解决泰语训练数据不足的问题;(3)采用 FastSpeech2 和 StyleMelGAN 声码器联合训练的方法解决声学特征失配的问题。为了验证所提方法的有效性,从注意力对齐图、客观评测 MCD 和主观评测 MOS 评分 3 方面进行测评。实验结果表明,使用所提泰语 G2P 模型可以获得更好的对齐效果进而得到更准确的音素持续时间,采用"所提泰语 G2P 模型+联合训练+迁移学习"方法的系统可以获得最好的语音合成质量,合成语音的 MCD 和 MOS 评分分别为 7.43±0.82 分和 4.53 分,明显优于基线系统的 9.47±0.54 分和 1.14 分。

**关键词**:语音合成;低资源;泰语 G2P 模型;迁移学习;联合训练
**中图分类号**　TP391

## Thai Speech Synthesis Based on Cross-language Transfer Learning and Joint Training

ZHANG Xinrui, YANG Jian and WANG Zhan
School of Information Science & Engineering, Yunnan University, Kunming 650504, China

**Abstract**　With the rapid development of deep learning and neural network, end-to-end speech synthesis system based on deep neural network has become the mainstream because of its excellent performance. However, in recent years, there are not enough researches on Thai speech synthesis, which is mainly due to the scarcity of large-scale Thai datasets and the special spelling of the language. This paper studies Thai speech synthesis based on the FastSpeech2 acoustic model and StyleMelGAN vocoder under the premise of low resources. Aiming at the problems existing in the baseline system, three improvement methods are proposed to further improve the quality of Thai synthesized speech. (1)Under the guidance of Thai language experts and combined with relevant knowledge of Thai linguistics, the Thai G2P model is designed to deal with the special spelling in Thai text. (2)According to the phonemes represented by the international phonetic alphabet converted by the designed Thai G2P model, languages with similar phonemes input units and rich data sets are selected for cross-language transfer learning to solve the problem of insufficient Thai training data. (3)The joint training method of FastSpeech2 and StyleMelGAN vocoder is used to solve the problem of acoustic feature mismatch. In order to verify the effectiveness of the proposed methods, this paper measures the attention alignment map, objective evaluation MCD and subjective evaluation MOS score. Experimental results show that using the Thai G2P model designed in this paper can obtain better alignment effect and thus more accurate phoneme duration, and the system using the "Thai G2P model designed in this paper+joint training+transfer learning" method has the best speech synthesis quality, and the MCD and MOS scores of the synthesized speech are 7.43 ± 0.82 and 4.53 points, which are significantly better than the 9.47±0.54 and 1.14 points of the baseline system.

**Keywords**　Speech synthesis, Low resource, Thai G2P model, Transfer learning, Joint training

## 1　引言

语音合成即给定文本产生语音,它是人工智能的一个重要研究方向且有着广泛的应用,如导航、手机以及智能家具等支持语音的设备,因此其受到学术界和工业界的持续关注。传统的语音合成系统效率低、性能差,难以适应当代社会的

基金项目:国家重点研发计划(2020AAA0107901);国家自然科学基金(61961043)
This work was supported by the National Key Research and Development Program of China(2020AAA0107901) and National Natural Science Foundation of China(61961043).
通信作者:杨鉴(jianyang@ynu.edu.cn)

发展，因此其逐渐被基于深度学习和神经网络的语音合成系统所取代，如 Tacotron[1]，Tacotron2[2]，FastSpeech[3] 和 FastSpe-ech2[4] 等。其中 Tacotron 和 Tacotron2 属于自回归语音合成系统，它们往往会受到合成速度慢、鲁棒性低以及没有相关语言特征导致不自然等问题的困扰。为解决这些问题，微软和浙江大学联合提出了非自回归语音合成系统 FastSpeech 和 FastSpeech2 并取得优异的效果。FastSpeech 主要包含编码器、长度调节器和解码器。其中编码器用于提取文本的序列表示，长度调节器用于解决音素和声谱图序列之间的长度不匹配问题以及控制语音速度和部分韵律，解码器则用于生成语音的 Mel 谱图。它的改进版本 FastSpeech2 在此基础上增加了一些语音特征信息预测器，包括音高、能量和更准确的持续时间。并且考虑到音高对语音韵律十分重要但由于其波动较大导致很难预测，因此 FastSpeech2 使用连续小波变换将音高轮廓转换为音高谱图并在频域中预测音高，这可以提高预测音高的准确性。FastSpeech2 减少了文本序列和 Mel 谱图之间的信息差距，更好地处理了非自回归语音合成系统中一对多映射问题，使得系统的合成效果达到了较高的水平。

泰语属汉藏语系壮侗语族，是泰国的官方语言，除泰国本土外，主要由分布在老挝、缅甸、越南西北、中国西南等地的傣泰民族使用。泰语的使用人口在籍约有 6000 多万并不算少，但近年来有关泰语语音合成的研究却不多见，大都还停留在传统的基于统计参数模型或基于 HMM 的泰语语音合成系统[5]。构建这些传统的泰语语音合成系统不但对语言学知识有严苛要求[6]而且还需有高质量大规模的数据集才有效果[7]。然而目前基于深度学习和神经网络的语音合成系统主要应用于通用语，如英语和汉语。这是因为其也需要规模较大的高质量数据集来训练，而对于缺少高质量数据集的非通用语泰语而言，直接把该语音合成系统应用在泰语语音合成上并不可取。并且在实验过程中我们还发现泰语本身存在一些复杂的语言问题影响着合成音频的质量，如泰语书写顺序和发音顺序不一致以及同一字符后接特定字符会产生变音等情况。这些关键因素致使泰语语音合成的研究相对落后。

目前在语音合成领域中我们所看到的大都是两阶段语音合成系统，例如 FastSpeech2 将输入文本转换为 Mel 谱图，接着声码器从 Mel 谱图生成语音。这些模型分开训练然后结合起来进行推理。其中声码器大多是用英语或汉语等通用语训练而成的，这里便存在着一个问题即该声码器若用于像泰语这样的非通用语合成将会缺少目标语言所特有的声学特征，也就是说声码器的训练阶段和推理阶段存在着声学特征不匹配的问题，这将会导致合成语音质量下降。

因此，本文设计了一个基于 FastSpeech2 和 StyleMel-GAN 声码器的泰语语音合成系统作为基线，并做出了以下改进来提高合成的泰语音频质量：(1)请教泰语专家结合相关语言学知识设计泰语 G2P 模型添加到系统前端，将泰语字符转换为国际音标表示的音素并调整位置从而解决泰语书写顺序和发音顺序不一致以及特定情况字符发生变音等问题；(2)根据系统输入的音素单元，对比其他通用语选择音素输入单元更为相似的语言进行跨语言迁移学习，从而解决泰语资源问题，弥补其训练数据的不足；(3)采用 FastSpeech2 和 StyleMelGAN 声码器联合训练的方法来解决声码器训练阶段和推理阶段声学特征不匹配的问题。

本文第 2 章介绍本文设计的泰语语音合成基线系统以及该系统中存在的问题；第 3 章介绍本文为解决基线系统中存在的问题所提出的模型及方法；第 4 章介绍实验数据与平台、实验设计、评测方法和实验结果。

## 2 泰语语音合成基线系统

为了实现以上研究目标，本文首先设计并实现了一个基于 FastSpeech2 和 StyleMelGAN 声码器的泰语语音合成基线系统，该系统用到的模型及方法介绍如下。

### 2.1 FastSpeech2

FastSpeech2 是目前十分主流的非自回归语音合成模型，它主要由编码器、语音特征信息预测器和解码器组成，模型如图 1 所示。编码器的作用是将音素嵌入序列转换为音素隐藏序列，它由多个 Feed-Forward Transformer(FFT)模块堆叠而成，每个 FFT 模块由一个自注意力和一维卷积网络构成，其中自注意力网络由多头注意力组成以提取交叉位置信息，如图 2 所示。FastSpeech2 的语音特征信息预测器将持续时间、音高和能量等不同的语音特征信息添加到音素隐藏序列中，提取真实音频中的相关信息进行训练，这样得到的合成音频自然度更高，如图 3 所示。相关语音特征信息介绍如下：1)音素持续时间，它决定了语音的声音长度；2)音高是传达情感的关键特征，对语音韵律有很大影响；3)能量则是直接影响语音音量的关键因素。值得一提的是，其中真实音频的音素持续时间可以通过外部强制对齐得到，也可以通过教师模型获取，而由于泰语语言的特殊性我们选用后者的方法进行提取真实音频的音素持续时间，这也将在 3.1 节详细说明。FastSpeech2 的解码器和编码器有着相同的基本结构，其作用是将调整后的隐藏序列并行转换为 Mel 谱图序列。最后将 Mel 谱图输入到声码器当中，即可得到合成的音频。

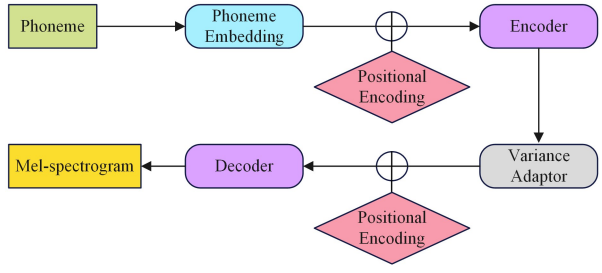



图 1 FastSpeech2 模型图

Fig. 1 FastSpeech2 model

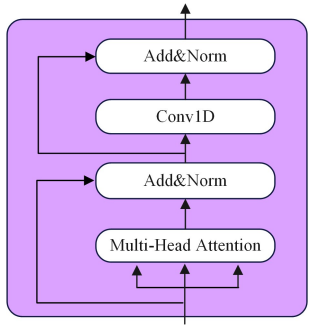



图 2 FFT 模块

Fig. 2 FFT module

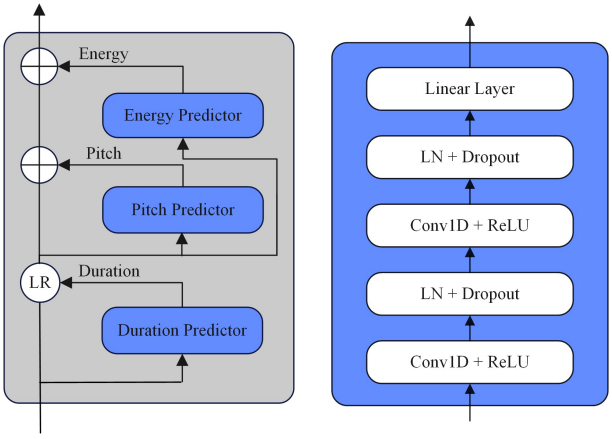



图 3　语音特征信息预测器

Fig. 3　Speech feature information predictor

## 2.2　StyleMelGAN 声码器

基于 GAN 的声码器是近年来的研究热点，主要原因是这类声码器相较于 WaveNet[8] 和 WaveGlow[9] 参数规模小、生成速度快，但合成质量稍差，如 MelGAN[10] 和 Parallel WaveGAN[11]。然而仅包含 $3.86\times10^6$ 个可训练参数的 StyleMelGAN[12] 却做到了兼顾参数小、生成速度快以及合成质量较高等优点并且其还可以起到减轻合成语音金属机械感的作用。

StyleMelGAN 声码器主要由生成器和判别器组成。生成器的作用是生成目标语音的 Mel 谱图，如图 4 所示。相较于其他基于 GAN 的声码器，它加入了噪声矢量，这对目标语音的低频部分信息起到了补充作用，从而改善了合成语音金属机械感重的问题。生成器包括 8 个上采样阶段和一个最终激活模块。其中每个上采样阶段包括一个 TADE[13]（Temporal Adaptive DE-normalization）残差模块和一个对信号进行二倍上采样的层，最终激活模块包括一个 TADE 残差模块和一个卷积层。训练过程中，生成器以目标语音的梅尔谱图为条件，通过 TADE 将其插入到生成器的每个 TADE 残差模块中，利用目标语音的声学特征对低维噪声矢量进行处理，最终输出合成的目标语音。判别器的作用是将合成语音与真实语音不断进行对抗性训练，以此逐渐减小损失从而减小合成语音与真实语音之间的差别，如图 5 所示。StyleMelGAN 采用 4 个判别器，每个判别器都由 PQMF、DBlock 模块、信号 4 倍下采样模块以及卷积模块等构成，其中每个 DBlock 模块都由一维卷积层和 LeakyReLU 激活函数构成。在训练过程中，先对输入的语音波形进行切片，接着通过 PQMF[14] 获得语音信号的子频带，最后不断优化对抗损失并进行判别直到判别通过为止。这里使用 PQMF 主要是为了可以针对特定频域进行合成从而减小计算量，以此来达到提升合成速率和质量的目的。

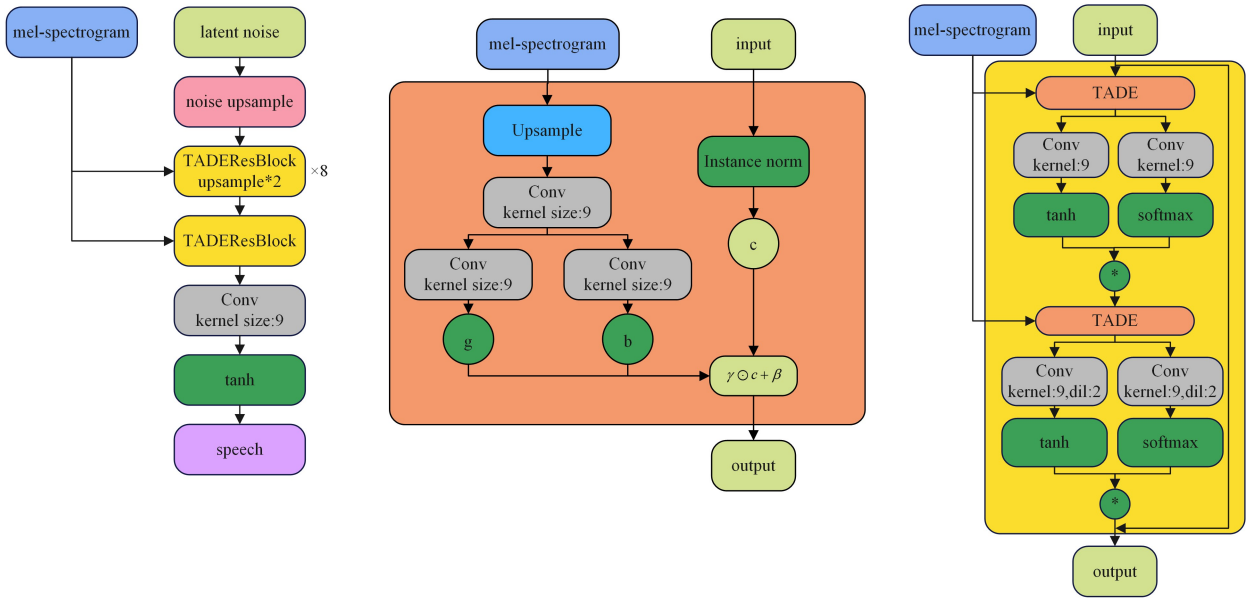



图 4　StyleMelGAN 生成器

Fig. 4　StyleMelGAN generator

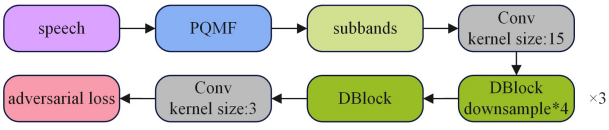



图 5　StyleMelGAN 判别器

Fig. 5　StyleMelGAN discriminator

## 2.3　基线系统

前端合成器选用 FastSpeech2，后端声码器选用预训练的 StyleMelGAN，接着将二者搭建成一个用于泰语语音合成的系统作为基线，如图 6 所示。其中 FastSpeech2 的编码器及解码器层数和单元数分别为 4 和 1 536，注意力头数和维度分别为 2 和 384。

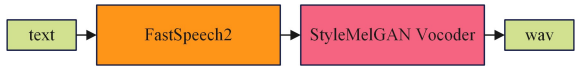



图 6　基线系统框图

Fig. 6　Baseline system

在实验过程中我们发现上述泰语语音合成基线系统存在一些问题，影响着合成语音的质量，具体问题如下：

（1）泰语本身的语言问题，即文本书写顺序与发音顺序不一致和同一字符后接特定字符会产生变音的情况。文本书写

顺序与发音顺序不一致会使得教师模型不能正确地学习对齐关系从而导致得不到准确的音素持续时间，这就会影响后续音频合成的质量。同一字符后接特定字符会产生变音，使得系统不能准确地学习到该字符在不同情况下的正确发音，进而造成合成音频可懂度降低。

(2)现有的泰语数据集规模很小，该程度的训练数据量无法训练出可以生成高质量泰语音频的泰语语音合成系统。

(3)声码器大多是用英语或汉语等通用语训练而成的，由于其在训练阶段未见过泰语，因此若将声码器直接用于泰语的语音合成则会产生声学特征不匹配的问题，从而致使合成语音质量下降。

为解决上述 3 个问题，本文根据泰语专家讲解结合相关语言学知识设计了泰语 G2P 模型并采用了迁移学习以及联合训练的方法旨在提高泰语合成语音的质量。

## 3 改进方法

### 3.1 依据泰语语言学知识设计的泰语 G2P 模型

泰语文字由字符构成且一般不使用标点符号。泰文字符的 Unicode 编码范围为 0E00-0E7F 代码点，共 128 个代码位，目前已使用 87 个，其余 41 个作为保留暂未使用，即可以把泰文看成是由 87 个字符构成的。87 个字符可构成 44 个辅音字母(现代泰语实际用到的有 42 个)，32 个元音字母(现代泰语实际用到的有 30 个)，5 个声调以及一些并不发音的特殊字符[15]。其中声调符号位于辅音字母的右上方且第一声调不标符号。值得一提的是，元音字母可以在辅音字母的上下左右出现甚至可以拆成两部分把辅音字母置于中间，这就导致泰语中很多字符排列顺序和实际发音顺序不同，此外泰语中还会出现某个字符由于后接特定字符而出现不同发音的特殊情况。例如，图 7 泰文“เรา”中“เ”和“ร”在国际音标里的读音分别是[au]和[r]，前者是作为一个整体发音的，然而实际书写顺序却是“ร”夹在“เ”中间，这就会导致字符排列顺序和实际发音顺序不同。并且“เ”和“า”在国际音标中分别发[e:]和[a:]的音，然而当“เ”后接“า”时就会出现“เ”发音为[au]的特殊情况。


图 7 泰语文本举例

Fig. 7 Examples of Thai texts

由于上述泰语语言的特殊性，若使用外部强制对齐的方法提取真实音频的音素持续时间难度很大不易实现，因此我们通过教师模型来获取，正如 2.1 节所述。

由于泰文可以由 87 个字符完整表示，因此我们可以采用 87 个字符直接作为系统输入，如图 8 所示。但这种单纯的字符输入并不能解决字符排列顺序和发音顺序不匹配的问题，也不能处理某个字符由于后接特定字符而出现不同发音的特殊情况。


图 8 字符形式的系统输入

Fig. 8 System input as characters

借鉴汉语的文本处理方式，即将汉字转换为拼音作为系统输入的方法[16]，本文准备在系统前端加入泰语 G2P 模型，将字符转换为音素从而对泰语文本进行处理。G2P 模型的全称是 Grapheme-to-Phoneme 模型，用于将字符转换为音素。截至目前有关泰语 G2P 模型的研究并不多见，espeak_ng_thai 便是其中具有代表性的一个。本文的教师模型采用 espeak_ng_thai 将泰语字符转换为音素去学习对齐关系但得到的对齐效果却很差。这是因为虽然 espeak_ng 支持 100 多种语言且包含泰语，但开发者讲述由于缺乏泰语语言学相关知识，espeak_ng_thai 只是初步建立，并不完善，且还有很大的提升空间。鉴于此，我们请教泰语专家结合语言学相关知识设计了一个泰语 G2P 模型并将其应用于教师模型，得到了良好的对齐效果，这在第 4 章的实验中得到了证实。具体做法是先对文本进行归一化处理将其他形式的文字统一转换为泰语文字，如阿拉伯数字等。接着使用 ThaiLMCut[17] 泰语分词模型对文本进行分词，然后用最大正向匹配算法去匹配泰语文本中产生变音的特殊情况再匹配正常的元音字母、辅音字母、声调符号和特殊字符，将它们按照泰语字符的国际音标表示规则转换为音素并调整位置使其和发音顺序一致后再输入系统，如图 9 所示。其中词边界用“*”表示，“〈sos〉”和“〈eos〉”分别代表输入的起止符号，4 个声调符号则由数字“1”“2”“3”“4”表示。采用这种方式方法既可以解决字符排列顺序和发音顺序不同的问题，还可以处理某个字符由于后接特定字符而出现不同发音的特殊情况。综上所述，在系统前端加入本文所设计的泰语 G2P 模型更有利于合成高质量的泰语语音。

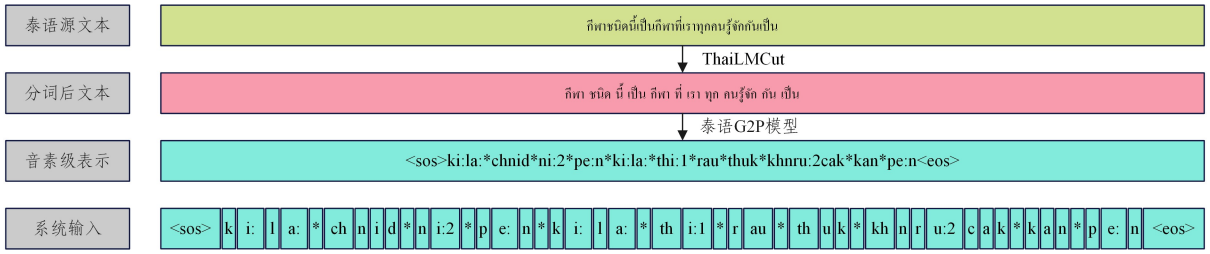



图 9 音素形式的系统输入

Fig. 9 System input in phoneme form

### 3.2 适配泰语音素级输入的跨语言迁移学习

语音合成的迁移学习通常被使用在两种比较相似的语言或输入之间，其中一种资源较丰富，数据集易获取，另一种资源较稀缺数据集规模很小，不足以单独作为训练系统的数据集。而迁移学习简单来说就是利用二者的相似特性来弥补低资源语言数据的不足[18-19]。具体做法是先用数据资源丰富的语言进行训练得到一个该语言的语音合成系统，目的是让系统学习其语言特性；接着通过共享网络权重和参数大小让目标语言语音合成系统学习到二者间相似的语言特性和语音共同的特征信息；最后使用目标语言的低资源数据集再次训练得到最终的目标语言语音合成系统。这样做除了可以解决数据不足的问题，对于目标语言语音合成系统而言迁移学习

还起到了加速训练的作用。这是因为在相似语言语音合成系统的基础上进行微调训练，部分特征信息已经被学习到了，所以在训练过程中收敛速度会加快。

目前数据资源丰富的语言主要为英语和汉语等通用语，然而对于数据集稀缺的非通用语泰语而言并没有一种合适的语言可以迁移学习。这是因为泰语的输入字符与英语和汉语的输入完全没有相似之处。这对低资源泰语语音合成的研究无疑是致命的。但加入 3.1 节中本文设计的泰语 G2P 模型后系统输入变成了国际音标表示的音素，这样便可以把也将输入转换为国际音标音素形式的英语或汉语作为迁移学习的源语言。这里我们选择采用 espeak_ng_english_us_vits G2P 模型的英语语音合成系统进行迁移学习，该系统输入就是国际音标表示的音素，和本文设计的泰语 G2P 模型转换得到的音素输入单元较为相似。

采用上述模型及方法便可以很好地解决泰语没有合适语言迁移学习的问题，以此弥补了数据资源的不足并在训练过程中加速了收敛，进而有利于低资源泰语语音合成的研究。

### 3.3 FastSpeech2 和 StyleMelGAN 的联合训练

文献[20]中提出将 FastSpeech2 和 HiFi-GAN[21] 声码器进行联合训练来解决声学特征失配的问题。其原理在于将 HiFi-GAN 声码器的生成器部分和 FastSpeech2 构建成一个文本到语音的生成器，接下来再和 HiFi-GAN 声码器的判别器部分进行对抗训练和判别分析，以此来解决由于声码器未见过目标语言而导致声学特征不匹配的问题并起到简化训练流程的作用。但我们发现采用 FastSpeech2 和 HiFi-GAN 声码器联合训练方法得到的泰语合成音频由于缺少语音低频部分信息致使其金属机械感较重，且又因为 HiFi-GAN 声码器参数较大导致训练过程中收敛较慢，于是本文对此进行改进。由于 StyleMelGAN 声码器具备参数小、生成速度快以及可以减轻合成语音金属机械感等优点，因此我们采用 FastSpeech2 和 StyleMelGAN 声码器联合训练的方法进行改进从而解决上述问题。其中声码器的训练往往也需要大规模的数据集，显然泰语数据集并不满足此条件。所以我们对 StyleMelGAN 声码器采用了预训练的方法，加载预训练模型参数并在此基础上使用泰语的小数据集对联合后的模型进行微调训练来解决该问题。这种方法不但可以弥补声码器训练数据的不足而且在训练过程中还起到了加速收敛的作用。总而言之，联合训练就是将合成器的 T2M(Text-to-Mel)和声码器中生成器的 M2W(Mel-to-Waveform)两阶段合并成一阶段 T2W(Text-to-Waveform)进行统一的训练，所以联合训练损失 $L_{T2W}$ 为合成器损失 $L_{T2M}$ 与声码器中生成器损失 $L_{M2W}$ 求和，损失函数公式如下：

$$L_{T2W}=\lambda_{T2M}L_{T2M}+\lambda_{M2W}L_{M2W} \tag{1}$$

其中，$\lambda_{T2M}$ 和 $\lambda_{M2W}$ 为超参数，作为控制两个损失大小的比例因子，文中都设置为 1.0。

综上所述，采用改进后联合训练方法的泰语语音合成系统在性能上优于分别进行单独训练的两阶段系统且还简化了训练流程，其框架如图 10 所示。

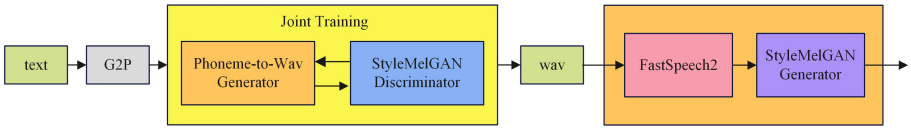



图 10　FastSpeech2 和 StyleMelGAN 声码器的联合训练

Fig. 10　Joint training of FastSpeech2 and StyleMelGAN vocoders

## 4　实验

### 4.1　实验数据与平台

实验中使用的泰语音频数据为一位母语是泰语的女性说话人录制，时长为 8 小时，单声道，采样率为 22 050 Hz，16 位 PCM 编码，音频前后保留 50 ms 的静音段。文本音频对共 4 982 句，其中 4 782 句用于训练，100 句用于验证，100 句用于测试，预训练模型使用的是英语数据集 LJSpeech。

本文涉及到的实验均是在 Linux 操作系统下的 ESPnet(End-to-End Speech Processing Toolkit)平台[22] 上完成的，ESPnet 是一个端到端的语音工具箱，其包含了端到端语音识别、语音合成、语音翻译、语音增强等功能。

### 4.2　实验设计

首先我们设计了如表 1 所列的 3 组实验来比较不使用 G2P 模型、使用 espeak_ng_thai 和使用本文设计的泰语 G2P 模型 3 种教师模型的对齐效果，从而证实采用本文设计的泰语 G2P 模型有助于更好地学习对齐关系，得到更准确的音素持续时间，这有利于后续持续时间预测器的训练。这里使用的教师模型是自回归模型 Tacotron2。

表 1　教师模型设置

Table1　Teacher model settings

| 实验名称 | G2P 模型名称 |
|---|---|
| Exp 1 | None G2P |
| Exp 2 | espeak_ng_thai |
| Exp 3 | 本文设计的泰语 G2P 模型 |

接着我们又设计了 7 组实验来证实本文提出的模型及方法对低资源条件下高质量泰语语音合成的有效性，实验设置如表 2 所列。

表 2　泰语语音合成实验设置

Table 2　Thai speech synthesis experiment settings

| 实验名称 | 实验方法 |
|---|---|
| Exp 4 | None G2P＋基线系统 |
| Exp 5 | NoneG2P＋基线系统＋迁移学习 |
| Exp 6 | espeak_ng_thai＋基线系统 |
| Exp 7 | espeak_ng_thai＋基线系统＋迁移学习 |
| Exp 8 | 本文设计的泰语 G2P 模型＋基线系统 |
| Exp 9 | 本文设计的泰语 G2P 模型＋基线系统＋迁移学习 |
| Exp 10 | 本文设计的泰语 G2P 模型＋联合训练＋迁移学习 |

### 4.3　评测方法

我们随机挑选了 100 句泰语音频作为测试集用于 Exp 4 至 Exp 10 的合成语音质量评测，本文采用梅尔倒谱失真(Mel Cepstrum Distortion，MCD)[23] 和平均意见得分(Mean Opinion Score，MOS)分别作为客观评测和主观评测的标准。MCD 为合成语音与真实语音的 MCEPs 之间的欧氏距离，记为：

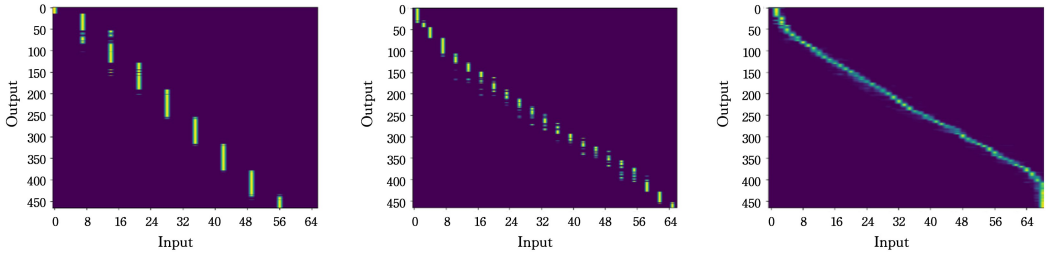
$$MCD[\mathrm{dB}]=\frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{N}(C_d-C_d^{syn})^2} \tag{2}$$

其中，$N$ 为 MCEPs 的维数，$c_d$ 和 $c_d^{syn}$ 分别为真实语音和合成语音的 MCEPs 的第 $d$ 维系数。MOS 评分即听者根据对音频的听觉感受进行评分，评分范围为 0.0～5.0 分，分值越高则表示音频的可懂度和自然度越高。我们邀请了 10 位泰语专

业的在读硕士研究生对用于评测的合成语音以及相应的真实语音进行评分。

### 4.4 结果分析

首先在测试集中随机挑选一句文本，观察其通过 Exp 1，Exp 2 和 Exp 3 得到的对齐图可以发现，采用本文设计的泰语 G2P 模型的确可以提升对齐效果，如图 11 所示。这里的注意力对齐图是体现合成质量的一个指标，其横轴代表输入序列，纵轴代表合成的语音帧。注意力对齐图表示输入序列和合成的语音帧之间的对应关系，最理想的状态是对齐线呈对角线趋势，此时代表解码器生成的语音帧与输入序列一一对应。图 11(a)和图 11(b)中对齐线不呈对角线趋势且断点很多，这表示合成语音与输入序列对齐效果很差，而采用了本文方法的图 11(c)中对齐线呈现了一个十分接近对角线的趋势且几乎不存在断点，这表示合成语音与输入序列得到了一个良好的对齐关系，以此证实了本文设计的泰语 G2P 模型的有效性。这为得到更准确的音素持续时间以及后续持续时间预测器的训练奠定了良好的基础。




(a)Exp 1 (b)Exp 2 (c)Exp 3

图 11 对齐图

Fig. 11 Alignment images

接着对 Exp 4 至 Exp 10 中挑选的合成语音进行 MCD 和 MOS 评分并取平均值，结果如表 3 所列。

表 3 实验结果

Table 3 Experimental results

| | MCD | MOS |
|---|---|---|
| 真实音频 | 0.00±0.00 | 4.84 |
| Exp4 | 9.47±0.54 | 1.14 |
| Exp5 | 10.38±0.58 | 1.04 |
| Exp6 | 9.64±0.56 | 2.31 |
| Exp7 | 9.73±0.58 | 2.43 |
| Exp8 | 9.29±0.68 | 2.76 |
| Exp9 | 8.05±0.78 | 3.87 |
| Exp10 | 7.43±0.82 | 4.53 |

通过比较表 3 中的实验结果可以得出结论：(1)Exp 4，Exp 6 和 Exp 8 的实验结果表明相同条件下相较于不使用 G2P 模型或使用 espeak_ng_thai G2P 模型的系统，采用本文设计的泰语 G2P 模型的系统所合成的泰语语音 MCD 平均值更小、MOS 平均分更高；(2)Exp 4 和 Exp 5 的实验结果表明简单的迁移学习并不能很好地解决泰语低资源问题从而提高合成音频质量，而 Exp 8 和 Exp 9 的实验结果则证实了结合本文设计的泰语 G2P 模型采用针对国际音标表示的音素输入单元进行跨语言迁移学习的方法很好地解决了上述问题；(3)Exp 9 和 Exp 10 的实验结果证实了我们使用的联合训练方法对于低资源泰语语音合成的有效性，采用了本文提出的 3 个改进方法的 Exp 10 得到的泰语合成语音 MCD 平均值最小、MOS 平均分最高，即该合成语音可懂度和自然度最高、合成效果最好。

**结束语** 本文基于 FastSpeech2 和 StyleMelGAN 声码器构建了泰语语音合成基线系统，在此基础上请教语言专家结合相关语言学知识设计了泰语 G2P 模型，解决了泰语中书写顺序和发音顺序不一致以及同一字符后接特定字符会产生变音的复杂问题，并根据设计的泰语 G2P 模型转换的音素选择合适的语言及输入单元进行跨语言迁移学习，弥补了泰语训练数据的不足，接着采用 FastSpeech2 和 StyleMelGAN 声码器联合训练的方法解决了声学特征失配的问题，从而提高了泰语合成语音的质量。接下来，我们将探索如何将情感等语音特征更精准地考虑进去，进一步提高泰语合成语音的自然度。

## 参 考 文 献

[1] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[J]. arXiv: 1703.10135, 2017.

[2] SHEN J, PANG R, WEISS R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2018: 4779-4783.

[3] REN Y, RUAN Y, TAN X, et al. Fastspeech: Fast, robust and controllable text to speech[C]// Proceesing of the 33rd International Conference on Advances in Neural Information Processing Systems. 2019: 3171-3180.

[4] 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.

[5] CHOMPHAN S, KOBAYASHI T. Implementation and evaluation of an HMM-based Thai speech synthesis system[C]// Eighth Annual Conference of the International Speech Communication Association. 2007.

[6] TESPRASIT V, CHAROENPORNSAWAT P, SORNLERTLAMVANICH V. A context-sensitive homograph disambiguation in Thai text-to-speech synthesis[C]// Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers. 2003: 103-105.

[7] WAN V, LATORRE J, CHIN K K, et al. Combining multiple high quality corpora for improving HMM-TTS[C]// Thirteenth Annual Conference of the International Speech Communication Association. 2012.

[8] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[J]. arXiv: 1609.03499, 2016.

[9] PRENGER R, VALLE R, CATANZARO B. Waveglow: A flow-

based generative network for speech synthesis[C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2019). IEEE, 2019: 3617-3621.

[10] KUMAR K, KUMAR R, DE BOISSIERE T, et al. Melgan: Generative adversarial networks for conditional waveform synthesis [J]. arXiv: 1910.06711, 2019.

[11] YAMAMOTO R, SONG E, KIM J M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2020). IEEE, 2020: 6199-6203.

[12] MUSTAFA A, PIA N, FUCHS G. Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization[C]// 2021 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2021). IEEE, 2021: 6034-6038.

[13] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2337-2346.

[14] NGUYEN T Q. Near-perfect-reconstruction pseudo-QMF banks [J]. IEEE Transactions on Signal Processing, 1994, 42(1): 65-76.

[15] QIN Y Y. Analysis of Thai phonetics teaching and teaching strategies for Chinese students in the primary stage[D]. Nanning: Guangxi University, 2017.

[16] LIU J, XIE Z, ZHANG C, et al. A novel method for Mandarin speech synthesis by inserting prosodic structure prediction into Tacotron2[J]. International Journal of Machine Learning and Cybernetics, 2021, 12: 2809-2823.

[17] SEEHA S, BILAN I, SANCHEZ L M, et al. Thailmcut: Unsupervised pretraining for thai word segmentation[C]// Proceedings of The 12th Language Resources and Evaluation Conference. 2020: 6947-6957.

[18] FAHMY F K, KHALIL M I, ABBAS H M. A transfer learning end-to-end arabic text-to-speech (tts) deep architecture[C]// Artificial Neural Networks in Pattern Recognition: 9th IAPR TC3 Workshop (ANNPR 2020). Winterthur, Switzerland, Cham: Springer International Publishing, 2020: 266-277.

[19] XU J, TAN X, REN Y, et al. Lrspeech: Extremely low-resource speech synthesis and recognition[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2802-2812.

[20] HAYASHI T, YAMAMOTO R, YOSHIMURA T, et al. Espnet2-tts: Extending the edge of tts research[J]. arXiv: 2110.07840, 2021.

[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] WATANABE S, HORI T, KARITA S, et al. Espnet: End-to-end speech processing toolkit[J]. arXiv: 1804.00015, 2018.

[23] KUBICHEK R. Mel-cepstral distance measure for objective speech quality assessment[C]// Proceedings of IEEE Pacific RIM Conference on Communications Computers and Signal Processing. IEEE, 1993: 125-128.

**ZHANG Xinrui**, born in 1999, postgraduate. His 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.