2024 年 第 24 卷 第 19 期
2024, 24(19): 08176-10

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2305948
引用格式:蔡姗, 王林, 谭棉, 等. 基于子音节表征的苗语语音合成方法[J]. 科学技术与工程, 2024, 24(19): 8176-8185.
Cai Shan, Wang Lin, Tan Mian, et al. Sub-syllable representation-based hmong language text-to-speech method[J]. Science Technology and Engineering, 2024, 24(19): 8176-8185.

基于子音节表征的苗语语音合成方法

蔡姗[1,2], 王林[1,2*], 谭棉[1,2], 郭胜[1,2], 吴磊[1,2], 王飞[2,3]
(1. 贵州民族大学数据科学与信息工程学院, 贵阳 550025; 2. 贵州省模式识别与智能系统重点实验室, 贵阳 550025;
3. 贵州民族大学人文科技学院, 贵阳 550025)

摘 要 少数民族语言的语音合成有助于民族文化的传承、保护和发展,目前相关研究成果较少。针对不同声调的相同词发音相似时易出现语音合成错误的问题,提出了一种基于子音节表征的苗语语音合成方法,该方法利用子音节作为训练基元来表征苗语发音信息,以区分学习不同音节间的相似发音。根据文本序列和梅尔谱图之间对齐的单调性,引入单调对齐损失来指导注意力模块进行更准确的对齐学习,以减少因注意力机制的自回归性带来的跳词、重复等合成现象。为验证所提方法的有效性,以自建苗语语音合成语料库 HmongSpeech(下载链接:http://sxjxsf.gzmu.edu.cn/info/1728/1214.htm)作为基准数据集,与典型的语音合成方法进行对比实验。实验结果表明,所提方法能够降低不同声调的相同词发音相似时导致的合成错误率,词错误率仅为 0.96%,较基线方法改善了 6.25%。

关键词 苗语语音合成; 子音节; 单调对齐; 语料库; 梅尔谱图
中图法分类号 TP391; **文献标志码** A

Sub-syllable Representation-based Hmong Language Text-to-Speech Method

CAI Shan[1,2], WANG Lin[1,2*], TAN Mian[1,2], GUO Sheng[1,2], WU Lei[1,2], WANG Fei[2,3]
(1. College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China;
2. Key Laboratory of Pattern Recognition and Intelligent System of Guizhou Province, Guiyang 550025, China;
3. College of Humanities & Sciences of Guizhou Minzu University, Guiyang 550025, China)

[**Abstract**] Speech synthesis of minority languages contributes to the preservation, protection and development of national culture, while the research results in this field are currently limited. To address the problem of speech synthesis errors where words with different tones sound similar, a sub-syllable representation-based text-to-speech method for the Hmong language was proposed. The method utilized sub-syllables as training primitives to accurately represent the pronunciation information of the Hmong language, enabling distinctive learning of similar sounds across different syllables. According to the monotonicity of alignment between text sequence and Mel-spectrogram, a monotonic alignment loss was introduced to guide the attention module to learn alignment more accurately, thereby reducing synthesis phenomena such as word skipping and repetition inherent in the autoregressive attention mechanism. To verify the effectiveness of the proposed method, a self-built Hmong language speech synthesis corpus, HmongSpeech (download link: http://sxjxsf.gzmu.edu.cn/info/1728/1214.htm), was utilized as the benchmark dataset. Comparative experiments were conducted with typical speech synthesis methods. The experimental results show that the proposed method successfully reduces the synthetic error rate caused by the similar pronunciation of words with different tones. Notably, the word error rate is only 0.96%, outperforming the baseline method by 6.25%.

[**Keywords**] Hmong language text-to-speech; sub-syllable; monotonic alignment; corpus; Mel-spectrogram

语音合成也叫文本转语音(text-to-speech,TTS),是一种将输入文本转化为可理解语音的技术[1],被广泛应用在人们的日常生活中,如滴滴语音播报、手机语音助手、智能教育等,是人机交互的重要组成部分。

收稿日期: 2023-08-04; **修订日期**: 2023-11-27
基金项目: 国家自然科学基金(62162012);贵州省科技计划(黔科合基础-ZK[2022]一般 195,黔科合基础-ZK[2023]一般 143,黔科合平台人才-ZCKJ[2021]007);贵州省教育厅自然科学研究项目(黔教技[2023]061 号,黔教技[2023]012 号,黔教技[2022]015 号);贵州省青年科技人才成长项目(黔教合 KY 字[2021]115,黔教合 KY 字[2021]110);贵州省模式识别与智能系统重点实验室开放课题(GZMUKL[2022]KF01,GZMUKL[2022]KF05);贵州省高层次创新型人才项目(黔科合平台人才-GCC[2023]027);教育部产学合作协同育人项目(221001766110209)
第一作者: 蔡姗(1996—),女,汉族,贵州纳雍人,硕士研究生。研究方向:语音合成及图像处理。E-mail:2291084203@qq.com。
* **通信作者**: 王林(1965—),男,苗族,贵州安顺人,博士,教授。研究方向:模式识别与图像处理。E-mail:wanglin@gzmu.edu.cn。

苗语作为中国少数民族语言之一,其语言文化仅通过口传心授,随着时间的推移,使用苗语交流的人越来越少。因此,对苗语的语音合成研究不仅可以促进苗族语言文化的传承、保护和发展,而且有助于少数民族地区的民汉双语教学教育发展。

传统的语音合成方法主要有基于波形拼接[2]和基于统计参数[3]的语音合成。基于波形拼接的语音合成不仅需要一个拥有大量语音片段的语音数据库做支撑,且该方法存在合成语音拼接感较强、合成结果不流畅的问题。而基于统计参数的语音合成方法结构复杂,对于特定语言的研究需要大量相关领域的专家知识,且合成语音存在机械感。

为简化语音合成框架、提高语音合成的质量和减少人工参与及文本预处理所需的语言学知识,深度学习[4]技术被广泛运用在语音合成领域,越来越多基于神经网络的语音合成方法相继而出。基于神经网络的语音合成可以进一步分为自回归和非自回归的方法。自回归模型[5]是一种序列到序列的生成模型,即对于当前时刻的预测依赖于前一时刻的输出,取得了一定的成果。其中,Wang 等[5]提出的 Tacotron 模型是第一个发展较为成熟的端到端语音合成方法,Tacotron 模型去除了复杂的文本前端,直接输入原始文本就能输出语音波形,简化了传统语音合成方法的复杂性。Shen 等[6]提出的 Tacotron2 模型简化了 Tacotron 中复杂的文本特征提取网络,并引入位置敏感注意力机制替换原来的内容注意力机制,减少了合成语音错误率。此外,Tacotron2 用 Wave Net[7]替换 Tacotron 中简单的 Griffin-Lim[8]算法,提高了合成语音质量。虽然这种逐帧生成的语音质量较好,但由于算法本身存在的递归性,使得推理速度较慢、效率较低。对此,Li 等[9]提出的 Transformer TTS 模型结合了 Transformer[10]和 Tacotron2[6]的优势,以并行的方式提取文本特征,加快了模型的推理速度,但解码器仍是自回归的。且基于注意力机制的隐式对齐学习由于注意力机制[11]的自回归性容易出现对齐错误,对于一些较长或复杂的句子,合成的语音往往会出现漏词、重复、错词等情况,导致最终合成的语音质量较差。为解决自回归模型存在的局限性,众多学者研究了非自回归模型。如,Elias 等[12]针对推理速度慢的问题,提出了 Parallel tacotron 模型,该模型由变分自编码进行增强,并使用轻量级卷积作为自注意力提高了合成的效率。Łańcucki[13]提出的 Fastpitch 模型通过引入教师网络来学习更准确的注意力对齐,同时加快了合成速度。非自回归模型以并行的方式生成声学特征,提高了语音合成的速度和效率。与自回归模型不同,非自回归模型不采用注意力机制来进行对齐学习,而是利用持续时间来建模文本和声学特征之间的对应关系[14]。但持续时间的获取来自一个预先训练好的自回归模型或外部对齐器,这增加了模型训练的复杂性。

随着人机交互技术的不断发展,语音合成朝着韵律[15]、情感[16]、多说话人[17]、多语言[18]等方向发展。但是,目前的语音合成主要集中于英语、汉语等主流语言[19],对少数民族语言、地区方言及小语种的研究相对较少[20]。Zu 等[21]提出基于非自回归声学模型 Fastspeech2[22]的藏语语音合成研究,解决目前藏语语音合成方法存在合成速度慢、重复单词或跳词、无法精细控制语速和韵律等问题。Huybrechts 等[23]针对低资源语言的语音数据获取昂贵的问题,采用语音转换的数据增强方式增加数据量,以提高低资源场景下的语音合成质量。丁云涛等[24]针对 Griffin-Lim 算法恢复语音波形自然度较低的问题,提出一种基于 WaveNet 的藏语语音合成方法,该方法具有更好的合成效果。刘瑞等[25]针对现有的蒙古语语音合成模型存在的合成效率低和语音保真度低的问题,提出完全非自回归的蒙古语语音合成模型,设计了两种时长对齐方法,实时生成高保真的蒙古语语音。杨琳等[26]提出基于迁移学习的越南语语音合成,解决了因数据资源不足使得合成质量差的问题。从语言的特殊性角度来看,一方面是以少数民族语言为母语的人口占比相对较少,其语料存在电子资源匮乏且难以获取的问题;另一方面是少数民族语言的语料特点不同于主流语言,不易分析,使得众多少数民族语言的语音合成发展迟缓。

苗语作为西南世居民族交流使用的语言之一,历史上没有本族文字。现代苗文是一种拉丁字母文字,拼写方式为“声母 + 韵母 + 声调”,不同的声调表示不同的意思,这会存在不同声调的相同词(声韵母的组合称为词)发音相似情况,从而导致合成错误语音的问题。因此,现提出基于子音节表征的苗语语音合成方法(sub-syllable representation-based Hmong language text-to-speech, SRHTTS)。SRHTTS 根据苗语发音结构,以子音节作为训练基元,子音节包含声母、音调集(韵母 + 声调),通过学习带声调的韵母信息以区分不同音节间的相似发音。同时考虑输入文本与梅尔谱图间对齐的单调特性,引入单调对齐损失[27]来指导注意力模块进行更准确的对齐学习,以提高合成语音的发音准确性。本文所做主要工作如下。

(1)提出适用于苗语的语音合成模型 SRHTTS,

通过自建苗语语音合成语料库 HmongSpeech 验证所提方法能准确合成出输入文本的对应发音,为苗语语音语料库的自动构建提供了技术支持,并为其他少数民族语言的语音合成研究奠定基础。

(2)针对不同音节间的相似发音,设计一种子音节的训练基元来表征苗语发音信息,解决不同音节之间发音相似导致的合成错误问题。

(3)从注意力机制的非单调性角度出发,通过引入单调对齐损失惩罚注意力机制中偏离主对角线的注意力,建立文本和梅尔谱图间准确的对应关系,从而解决注意力机制自回归性导致的跳词、重复等合成问题。

1 苗语文字特征

苗语是苗族人的语言,属汉藏语系苗瑶语族苗语支。1956 年根据苗语方言差别较大的问题,创制了东部、中部、西部 3 种以拉丁字符为书写单元的方言文字,由声母、韵母和声调三部分拼写而成[28],其中韵母和声调统称为音调集。例如,“花”的苗文拼写为“bangx”,其组成结构如图 1 所示。苗语的声母、韵母与声调组合在一起形成一个音节,各音节间以空格进行划分,一个音节表示一个字,是一种单音节语言。

本文研究对象为中部苗语,也称黔东苗语,以黔东南州凯里市养蒿村苗语为标准音。中部苗语共有 32 个声母,26 个韵母和 8 个声调,如表 1 所示。

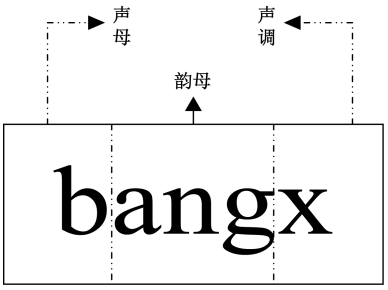

图 1 苗文书写结构

Fig. 1 Architecture of Hmong language writing

表 1 苗语的声韵母及声调

Table 1 Vocal vowels and tones of Hmong language

声母	韵母	声调
b, p, m, hm, f, hf, w, d, t, n, hn, dl, hl, l, z, c, s, hs, r, j, q, x, hx, y, g, k, ng, v, hv, gh, kh, h	i, e, a, o, u, ai, ei, ia, io, ie, iu, ang, en, ong, in, iang, iong, ee, ao, iee, iao, ui, ua, uai, un, uang	b:调值 33, x:调值 55, d:调值 35, l: 调值 22, t: 调值 44, s: 调值 13, k: 调值 53, f: 调值 31

2 子音节表征的苗语语音合成方法

2.1 问题描述

现代苗语采用拉丁字母进行文字表示,是一种带声调的单音节语言。不同声调所代表的含义有所不同,这会存在不同声调的相同词发音相似的情况。若直接将英语的语音合成模型应用到苗语上,模型会难以学习到苗语的声调信息,从而导致合成语音错误的问题。因此,本文提出了基于子音节表征的苗语语音合成方法来学习更准确的苗语发音。

该方法以端到端模型 Tacotron2[6] 为基础,端到端模型去除了复杂的文本前端模块,降低了语音合成对语言学知识的要求。SRHTTS 架构如图 2 所示,是一种编解码的结构。SRHTTS 在基线模型的原有结构上将字符嵌入替换为子音节嵌入,同时在梅尔谱图预测中引入单调对齐损失来惩罚注意力机制中的非单调对齐,从而建立文本和声学特征间更准确的映射关系,并用 HiFi-GAN 声码器替换原来复杂的 Wave Net 网络。

输入文本首先经过子音节提取模块获得子音节训练基元,子音节序列通过编码器提取具有上下文信息的文本隐藏表示特征,以捕捉文本序列间的长距离依赖关系[29-30],表达式为

$$f_e = \mathrm{ReLU}(F_3\mathrm{ReLU}\{F_2\mathrm{ReLU}[F_1\bar{E}(X)]\}) \tag{1}$$

$$H = \mathrm{EncoderRecurrency}(f_e) \tag{2}$$

式中:F_1、F_2、F_3 为 3 个卷积核;ReLU 为每一个卷积层上的非线性激活函数;$\bar{E}(\cdot)$ 表示对输入序列 X 进行 Embedding;H 为上下文文本隐藏特征,EncoderRecurrency(·)表示双向长短期记忆网络 Bi-LSTM。

其次,采用位置敏感注意力机制来获取注意力上下文向量,以学习输入文本与梅尔谱图间的对齐,表达式为

$$e_{ij} = \boldsymbol{v}_a^{\mathrm{T}}\tanh(\boldsymbol{W}s_{i-1} + \boldsymbol{V}h_j + \boldsymbol{U}f_{i,j} + \boldsymbol{b}) \tag{3}$$

式(3)中:e_{ij} 为注意力得分;s_{i-1} 为上一个时间步 $i-1$ 的解码器隐藏状态;h_j 为第 j 个编码器隐藏状态;$f_{i,j}$ 为上一个解码步的注意力权重经卷积获得的位置特征;$\boldsymbol{v}_a$、$\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{V}$ 和 $\boldsymbol{b}$ 为待训练参数。

最后将具有对齐信息的文本隐藏表示经线性映射输出预测梅尔谱图,再由 HiFi-GAN 将梅尔谱图转化为语音波形。整个过程可以表示为

$$\boldsymbol{a}_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{L}\exp(e_{ik})} \tag{4}$$

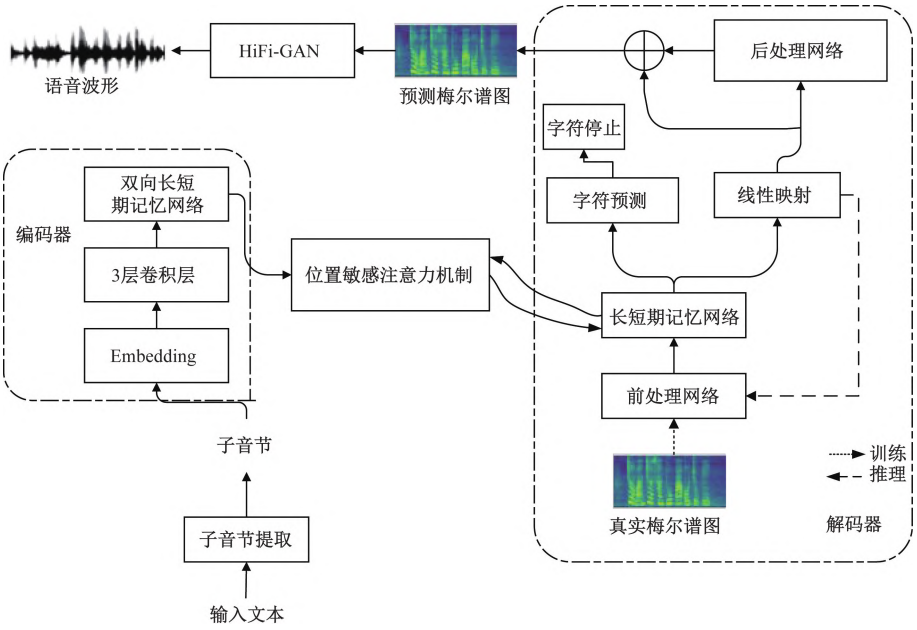

图 2　苗语语音合成模型

Fig. 2　Hmong language speech synthesis model

$$c_i = \sum_{j=1}^{L} a_{ij} h_j \tag{5}$$

$$y_i = \text{Decoder}(s_i, y_{i-1}, c_i) \tag{6}$$

$$z_i = \text{HiFi-GAN}(y_i) \tag{7}$$

式中:a_{ij} 为注意力权重;L 为输入序列的长度;c 为根据注意力机制得到的上下文向量;y 为预测梅尔谱图序列;z 为语音波形。

2.2　子音节的提取

由于基线模型是针对英语设计的,采用字符作为训练基元就能合成出较好的语音。但苗语与英语不同,苗语是一种带声调的单音节语言,由声母和音调集组成,这使得苗语词的组合方式复杂多样,且具有不同声调的相同词其发音相似。因此,基线模型的字符基元可能不适用于苗语,故以子音节作为模型的训练基元。如图 3 所示,通过将输入文本按照苗语的发音结构进行划分,分为声母和音调集,从而获得苗语文本的子音节,以作为模型的训练基元。

在提取输入文本的子音节后,根据构建的字典库进行子音节的位置索引,得到数字化的子音节序列以转化为计算机能够处理的数据形式。子音节是一种更细粒度的发音单元,模型首先学习字的声母发音,再学习不同声调的韵母发音,最后将学习到的声母和不同声调的韵母发音连接起来得到字的完整发音。这遵循了苗语的发音规则,有助于模型正确学习输入文本对应的发音,降低不同音节间发音相似导致的合成错误率。

2.3　损失函数

为了更好地学习文本和梅尔谱图之间的对齐,解决由注意力机制的自回归性导致合成语音出现

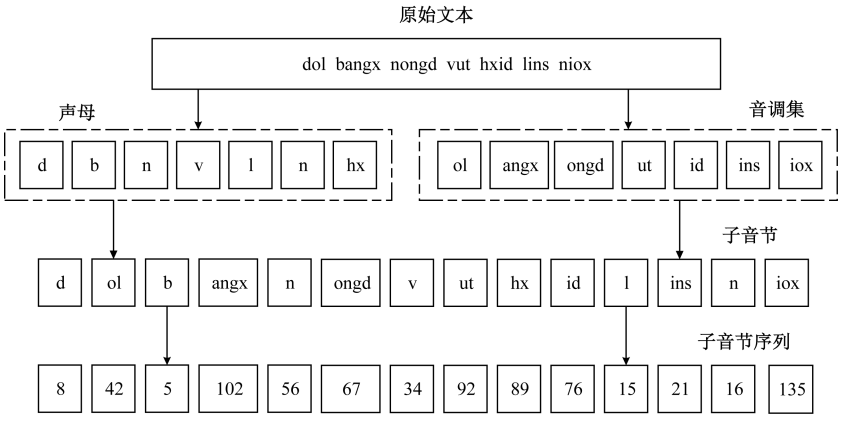

图 3　苗语文本的子音节提取

Fig. 3　Acquisition of sub-syllables in Hmong language text

漏词、重复等问题,本文受文献[27]中单调注意力的启发,在基线模型的注意力模块中添加单调对齐损失。单调对齐损失对偏离主对角线的注意力进行惩罚,使得注意力矩阵朝着对角矩阵靠拢,以建立输入文本序列与输出梅尔谱图序列之间准确的对应关系,计算公式为

$$\bar{a}_i = \sum_{j=1}^{L} \boldsymbol{a}_{ij} j \tag{8}$$

$$\text{Mono\_loss} = \sum_{i=1}^{N-1} \max\left(\frac{\bar{a}_i - \bar{a}_{i+1} + \delta L/N}{L}, 0\right) \tag{9}$$

式中:Mono_loss 表示单调对齐损失;$\bar{a}_i$ 为第 i 帧梅尔谱的对齐权重质心;L 为输入序列的长度;N 为梅尔谱图序列的帧长度;δ 为一个超参数,根据输入输出序列的长度动态调整对齐,当 $\bar{a}_{i+1} \geqslant \bar{a}_i$ 时满足单调性,即 $(\bar{a}_i - \bar{a}_{i+1} + \delta L/N)/L$ 为负值时,单调对齐损失为 0,此时满足单调性原则;$(\bar{a}_i - \bar{a}_{i+1} + \delta L/N)/L$ 为正值时,对不满足单调性的值进行惩罚。

基线模型中的损失函数包含两个部分,一是进入后处理网络前和经过后处理网络后的梅尔谱图分别与真实梅尔谱图之间的均方误差损失,二是停止符预测的交叉熵损失,表达式分别为

$$\text{Mel\_loss} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{real},i}^{\text{Mel}} - y_{\text{before},i}^{\text{Mel}}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{Real},i}^{\text{Mel}} - y_{\text{after},i}^{\text{Mel}}\right)^2 \tag{10}$$

$$\text{Stop Token\_loss} = -[y\lg p + (1-y) \times \lg(1-p)] \tag{11}$$

式中:Mel_loss 表示进入后处理网络前后关于梅尔谱图的损失;$y_{\text{real},i}^{\text{Mel}}$ 为从原始音频中提取的真实梅尔谱图;$y_{\text{before},i}^{\text{Mel}}$ 和 $y_{\text{after},i}^{\text{Mel}}$ 分别为进入后处理网络前、后解码器输出的粗糙和精细梅尔谱图;n 为每批的样本数;Stop Token_loss 表示停止符预测的损失,y 为停止符真实概率分布;p 为解码器线性层输出的预测分布。

根据式(8)~式(11),将基线模型的原有损失和单调对齐损失共同优化,总的损失函数表示为

$$\text{Loss} = \text{Mel\_loss} + \text{StopToken\_loss} + \lambda_1 \text{Mono\_loss} \tag{12}$$

式(12)中:λ_1 为权重系数,用于调整单调对齐损失在总损失中的影响。

3 实验结果与分析

为了验证 SRHTTS 的性能,本节设置了四组实验,分别用于验证子音节基元、单调对齐损失和方法的有效性及鲁棒性。

3.1 实验设置

文中实验采用自建的苗语语音合成语料库 HmongSpeech 作为基准数据进行验证分析(数据下载链接:http://sxjxsf.gzmu.edu.cn/info/1728/1214.htm)。HmongSpeech 数据集包含 4 650 个音频片段及对应的文本,音频采样率为 44 100 Hz,格式为.wav、16 位采样精度和单声道。表 2 给出了 HmongSpeech 中语音标注信息的一些例子,包括苗语的拉丁文拼写形式及其汉语翻译。文本的平均长度是 7 个音节,最短的文本是 3 个音节,最长的文本是 18 个音节。将数据集划分为训练集、验证集和测试集三部分,分别有 4 295、255 和 100 对 <音频,文本>。由于现代苗文不涉及特殊字符,故去除文本规范化操作。本文将苗文拉丁序列表示中的每个拉丁单词称为音节,将拉丁单词中的每个字母都称为字符,将声母和音调集统称为子音节。

表 2 苗语语音标注信息

Table 2 Hmong language speech annotation information

文本编号	苗文文本	译文
0001	dol bangx nongd vut hxid lins niox	这些花好看极了
0002	mongl gux pab nenk dul lol diod	到外面劈一点材来烧
0003	baib nenx laib mos det diot khob	给他帽子戴头上
0004	nenx ib det hmid lod yangx	他的一颗牙断了
0005	mongl liuk laib fab lol hot	去摘一个瓜来煮
0006	jox hlat nongd nongk hfab dad nenf	这根绳子要搓长一点

所有实验均在 Linux 系统上进行,基于 Tensorflow 和 Pytorch 的深度学习框架,并使用单块 GPU V100 的显卡训练所有模型,批量大小设置为 32,使用 Adam 作为优化器,权重衰减设置为 1×10^{-6},学习率为 1×10^{-3},总共训练 200×10^3 步。此外,使用 HiFi-GAN 作为声码器,为了更好地适应数据分布,重新在苗语数据上训练 HiFi-GAN 模型,不使用预训练模型。

本文设置的 4 组实验:①第一组研究不同训练基元和单调对齐损失的有效性;②第二组对比研究 SRHTTS 方法与其他方法的主观评价指标,以分析不同方法合成语音的可理解性;③第三组进行 SRHTTS 方法与其他方法的客观评价指标对比,以分析不同方法合成语音的保真性;④第四组研究 SRHTTS 方法与其他方法的词错误率和句错误率,以验证 SRHTTS 方法的鲁棒性。

3.2 评价指标

语音合成的评价指标主要分为主观和客观评价。主观评价是根据人的听觉感受对语音的质量进行评分,客观评价是通过机器对合成语音与真实语音之间的差距进行评测,更详细的介绍将在下面小节中呈现。

3.2.1　主观评价

语音合成的主观评价通常是采用平均意见得分(mean opinion score,MOS),即测试者根据自己的听觉感受来对被测语音样例的整体质量进行打分。主观评价的评分标准如表3所示。MOS是一种分级判断指标,采取5个级别对被测语音的质量进行评价。具体实施过程是先收集真实和合成的语音样本,再邀请多名母语者或非母语者对语音样本进行打分,最后统计评分求均值,得到不同方法的MOS值。分值越大表明合成的语音质量越好。

表3　语音主观评测标准

Table 3　Subjective speech evaluation criteria

音频级别	MOS值	评价标准
优	5	很好,听得清楚;延迟小,交流流畅
良	4	稍差,听得清楚;有点杂音
中	3	还可以,听不太清;可以交流
差	2	勉强,听不太清;交流需要重复多遍
劣	1	极差,听不懂;延迟大,交流不通畅

3.2.2　客观评价

语音合成的相似性指标采用梅尔倒谱失真(Mel-cepstral distortion,MCD),MCD通过逐帧的方式计算合成的梅尔谱特征与真实梅尔谱特征之间的谱距离来度量梅尔频率倒谱系数(Mel-scale frequency cepstral coefficients,MFCC)的重建性能,计算公式为

$$\mathrm{MCD}_K = \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sum_{k=1}^{K}\left(c_{t,k}-c'_{t,k}\right)^2} \tag{13}$$

式(13)中:$c_{t,k}$和$c'_{t,k}$分别为真实音频和预测音频的第t帧的第k个MFCC。MCD通常使用$K=13$维的MFCC特征的均方误差来计算。其思想是真实和合成的梅尔到谱序列之间的MCD越小,合成语音的自然性越接近真实语音。

语音合成的准确性可以用词错误率(word error rate,WER)来度量,词错误率考虑了语音合成中常出现的跳词、重复等情况,可以表示为

$$\mathrm{WER} = \frac{S+D+I}{N} \tag{14}$$

式(14)中:S为发音错误的数目;D为漏词数;I为重复数;N为总单词数。

3.3　实验结果与分析

3.3.1　消融实验

从测试集中随机抽取了50个句子进行消融研究以验证SRHTTS中几个关键技术的有效性,包括不同训练基元和有无单调对齐损失的对比。MCD和WER的评估结果如表4所示。

表4　不同训练基元和单调对齐损失的有效性评估结果

Table 4　Effectiveness evaluation results of different training primitives and monotonic alignment loss

方法	MCD	WER/%
Baseline_Syll	32.65	55.29
Baseline_Char	15.31	4.81
Baseline_Sub-Syll	14.25	3.33
Baseline_ Sub-syll_monoloss	14.04	1.44

据表4可知,对于训练基元的有效性分析,基于基线模型(Baseline)进行了字符(Baseline_Char)、音节(Baseline_Syll)和子音节(Baseline_Sub-syll)的对比,结果显示,由子音节作为训练基元的合成方法能够得到更低的MCD和WER值,分别为14.25和3.33%,表明其合成的语音与真实语音间的相似性更高,同时也降低了合成错误率。而以音节作为训练基元的合成方法的WER达到了55.29%,其合成语音中有一半的错误发音。此外,在子音节训练的基础上加入单调对齐损失(Baseline_ Sub-syll_monoloss)后,WER减少了1.89%,表明模型能够建立文本和梅尔谱图间准确的对应关系,从而学习到更准确的文本发音。

为了更直观地呈现所提方法的有效性以及挖掘存在发音错误的原因,通过可视化的方式对比分析了由不同训练基元和单调对齐损失得到的对齐效果及预测的梅尔谱图。不同训练基元实验的可视化结果如图4所示,其中,所有模型都训练了200×10^3步,采用的测试句子为:yaf bib xongs zab jex bib yaf linf ob xongs。由图4(b)可知,尽管训练了200×10^3步,使用音节作为训练基元的注意力图仍不是一个对角型,且与图4(a)的真实谱图相比,图4(c)得到一个错误的预测谱图(实线框部分是较明显的梅尔谱图特征对比)。而基于字符和子音节的方法能得到更准确的对齐和发音,预测的梅尔谱图[图4(e),图4(g)]准确展示了每个音节的发音特征。但相比于子音节基元,基于字符训练的对齐图[图4(d)]会出现断裂的情况,连贯性和稳定性不如基于子音节[图4(f)]的方法,这会使得合成的语音出现跳词、漏词等现象。

上述实验结果暗示了以音节作为模型输入时,编码维度过高,导致简单的特征提取器难以提取到有效的文本特征,且需要上百个小时的语音语料库才能基本覆盖所有的发音情况。然而,本文目前的语料库规模远不能达到此要求,因此造成合成的错误率达到55.29%。而以字符或子音节作为训练基元的编码维度仅有音节的0.17,能更好地发挥特征提取器的提取能力,且这种少字符集的学习方式可避免出现集外词的情况,模型学习到的发音信息更

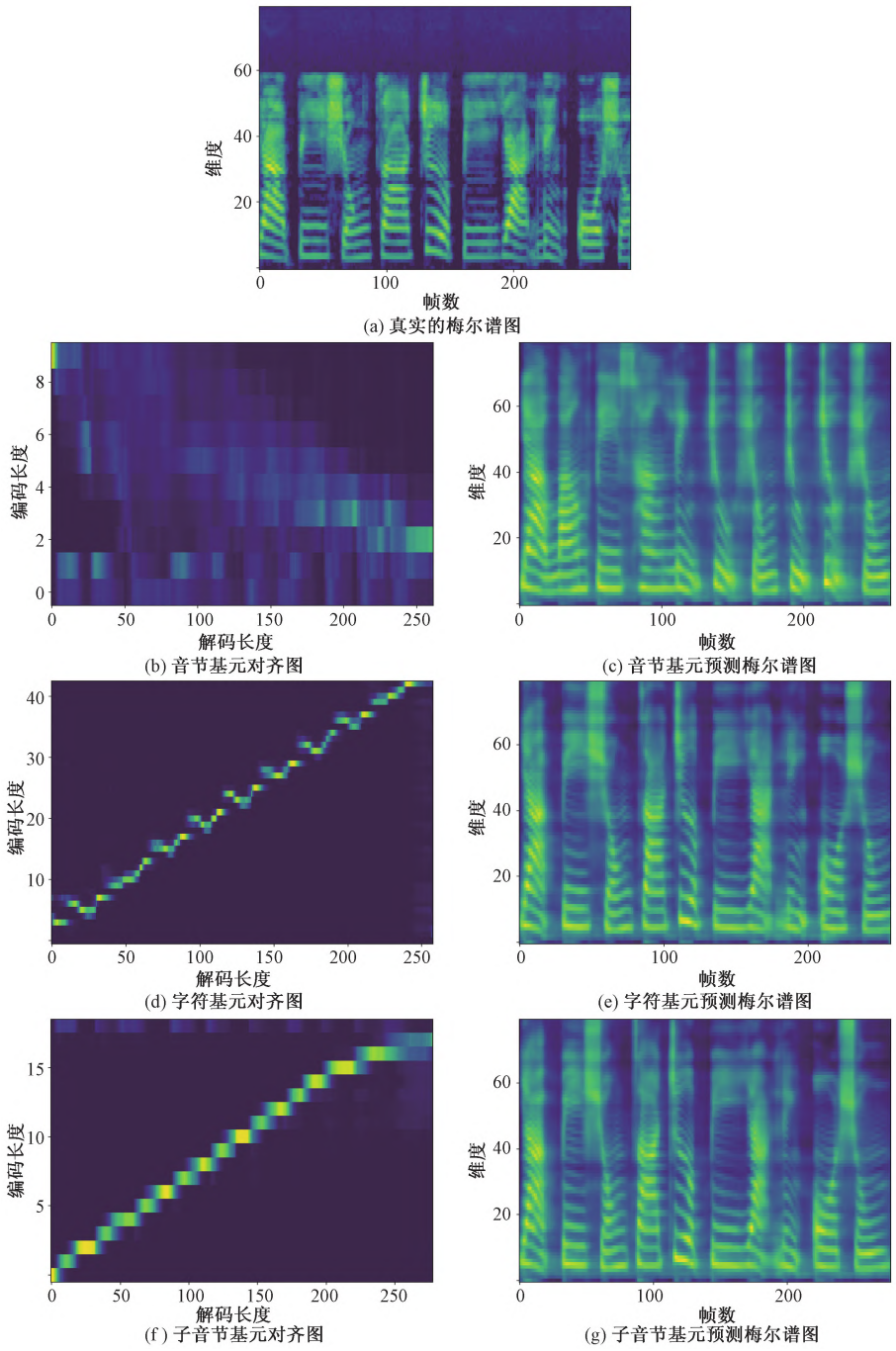

图 4 使用音节、字符和子音节作为训练基元,在没有单调对齐损失时的对齐

Fig. 4 Alignment without monotonic alignment loss using syllables, characters and sub-syllables as training primitives

多。此外,由于苗语是带声调的单音节语言,且其八个声调的字符表示与部分声母相同,这会使得模型混淆同形异音的声调和声母的发音,从而出现错误的发音。对于单调对齐损失的有效性分析,可视化结果如图 5 所示。

图 5(a)和图 5(b)分别表示没有单调对齐损失时训练了 1 500 步和 150×10^3 步的对齐图,结果表明,在 1 500 步时没有出现对齐,但随着迭代次数的增加,在 150×10^3 步时有较好的对齐,模型基本收敛。然而在边缘部分还存在一些干扰项,即当前输出的梅尔谱图帧在受到对应文本序列影响的同时也受到其余文本序列的干扰。图 5(c)和图 5(d)分别表示有单调对齐损失时训练了 1 500 步和 150×10^3 步的对齐图,结果显示,在加入单调对齐损失后,模型能在更早的训练阶段就显示出明显的对齐情况,表明单调对齐损失有助于注意力模块的快速学习。当模型训练到 150×10^3 步时,注意力对齐图呈完全对角状态,黄色的像素点很明亮,表明编码器

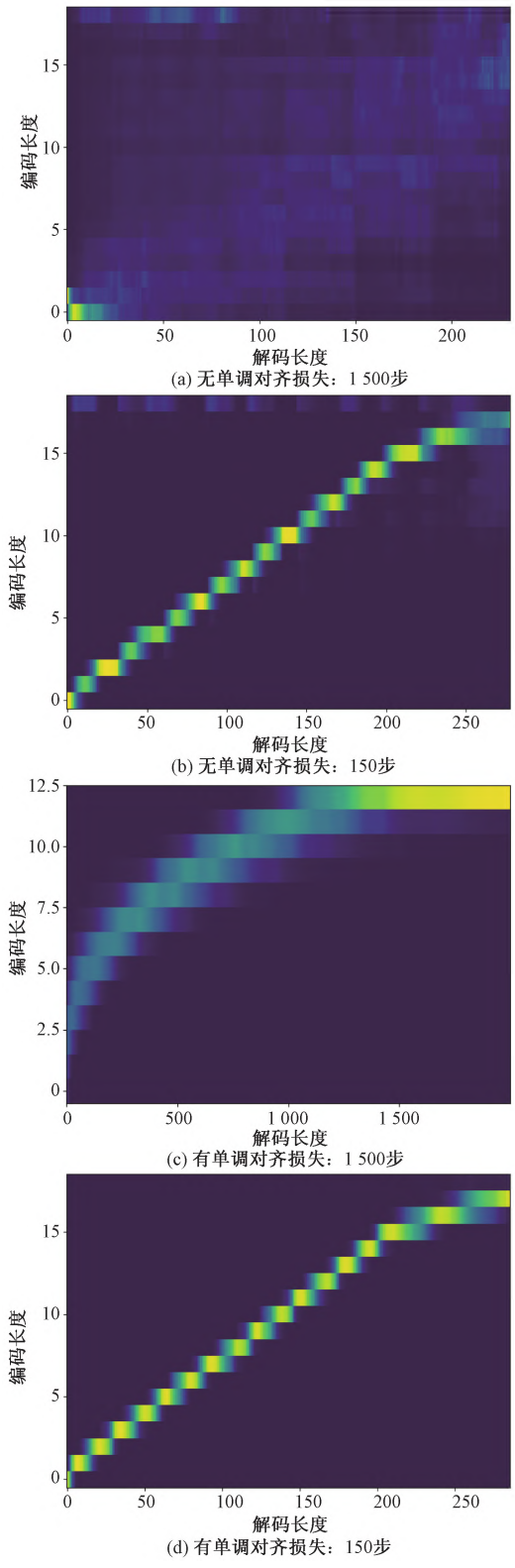

图 5　使用子音节作为训练基元,有无单调对齐损失时的对齐比较

Fig. 5　Comparison of alignment with and without monotonic alignment loss using sub-syllables as the training primitive

输出的文本隐藏表示特征与解码器输出的梅尔谱图之间建立了基本正确的对应关系,模型可以学习到输入文本的准确发音。

综上所述,经消融研究表明,以子音节作为训练基元,模型能有效学习不同音节的相似发音,降低了合成错误率;此外,单调对齐损失的引入能加快注意力机制的收敛速度,同时得到输入文本与梅尔谱图间更准确的对齐,减少了漏词、重复等合成现象。

3.3.2　合成语音质量评估

为了评估合成语音的质量,我们从测试集中随机选择了 30 个样本,将 Tacotron2 作为基线方法与本文提出的 SRHTTS 方法进行比较,同时也比较了第一个端到端的自回归语音合成方法 Tacotron。GT(ground truth)和 GT Mel(ground truth Mel)分别表示真实语音和由真实梅尔谱图合成的语音。为了比较的公平性,所有方法都采用 HiFi-GAN 作为声码器,且均训练至收敛。

对于主观指标的评估,邀请 10 名母语为黔东苗语的青年学生和 15 名非母语的研究生对真实的语音和合成的语音进行 MOS 评分。评估结果如图 6 所示。

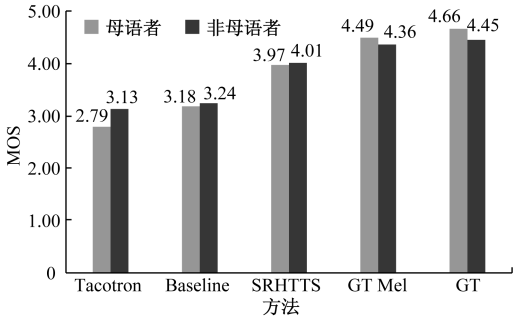

图 6　不同方法的主观评估结果

Fig. 6　Subjective evaluation results of different methods

由图 6 可知,与基线方法相比,母语者认为 SRHTTS 的语音可懂度比基线方法的好,达到 3.97 的 MOS 值;非母语者考虑了语音的自然度和清晰度,SRHTTS 依然有较高的 MOS 值,为 4.01。与其他方法相比,Tacotron 方法的 MOS 值仅为 2.79,且合成的语音存在杂音和浓厚的机械感,其语音内容难以被人理解。从真实梅尔谱图生成(GT Mel)的语音质量来看,SRHTTS 合成的语音与真实语音之间的差距较小。因此,以子音节作为训练基元及加入单调对齐损失有效提高了苗语语音合成的质量。

对于客观指标的评估,以 MCD 来度量合成语音与真实语音之间的相似性。由于合成的语音不存在静音段,通常来说比真实语音的时长短,故在计算 MCD 指标时,先采用动态时间规划(dynamic time warping,DTW)算法将合成语音与真实语音对齐。实验结果如图 7 所示。

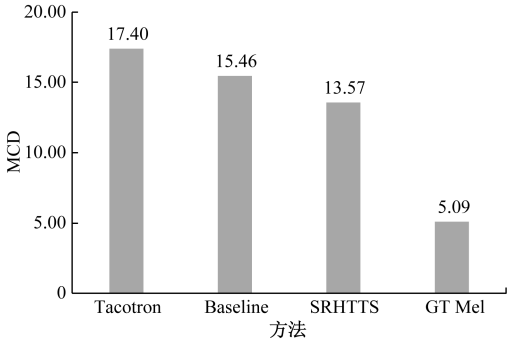

图 7 不同方法的客观评估结果

Fig. 7 Objective evaluation results of different methods

据图 7 可知,SRHTTS 的 MCD 值为 13.57,较基线方法改善了 1.89,与 Tacotron 相比改善了 3.83,说明 SRHTTS 方法合成的语音与真实语音的相似性更高,即 SRHTTS 能有效合成出较高质量的苗语语音。

由主观实验和客观实验结果可以充分表明 SRHTTS 方法在给定文本时能成功合成出对应的苗语语音,且合成的语音具有可理解性和自然性。此外,为了进一步验证所提方法的鲁棒性,将在下一子节中对 SRHTTS 方法的鲁棒性进行详细分析。

3.3.3 鲁棒性评估

为评估 SRHTTS 方法的鲁棒性,从测试集中随机选取了 30 个句子,共 208 个词,检验不同方法下合成语音的发音准确性,选取的句子没有在训练中出现过。各种发音错误的统计结果呈现在表 5 中。

表 5 不同方法下合成语音的鲁棒性评估

Table5 Robustness evaluation of synthesized speech under different methods

Method	重词数	跳词数	错词数	错句数	字错误率/%	句错误率/%
Tacotron	8	11	33	20	25.00	66.67
Baseline	2	1	12	12	7.21	40.00
SRHTTS	0	0	2	2	0.96	6.67

表 5 可知,SRHTTS 相较于 Tacotron 和 Baseline 的鲁棒性更强,仅有 0.96% 的字错误率和 6.67% 的句错误率,重词数和跳词数均为 0。因此,鲁棒性实验结果表明,SRHTTS 采用子音节和单调对齐损失共同训练时能够降低合成错误率及减少合成语音中词的重复和跳跃等现象。

4 结论

针对不同音节间发音相似导致合成错误的问题,本文提出基于子音节表征的苗语语音合成方法 SRHTTS,并构建了一种单说话人的苗语语音合成语料库。SRHTTS 以子音节序列作为输入,遵循了苗语的发音规则,能更好地表征苗语的发音信息。同时引入的单调对齐损失考虑了文本和梅尔谱图之间的单调性,更准确地学习输入文本的发音,提高合成的准确性。实验结果发现所提 SRHTTS 方法能有效学习不同音节的相似发音,降低了合成错误率,并减少漏词、重复等合成现象,从而提高了语音合成质量。所提方法也为其他以拉丁字符为书写单位的少数民族语言的语音合成研究提供思路。针对本文方法和语料库规模的不足,未来工作将主要集中于:①继续增加苗语语音合成语料库的文本与语音数量,解决因发音覆盖不全带来的合成问题;②考虑非自回归的序列建模方式,解决自回归性导致合成速度慢、效率低的问题。

参 考 文 献

[1] Nguyen B, Cardinaux F, Uhlich S. Autotts: end-to-end text-to-speech synthesis through differentiable duration modeling[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE Press, 2023: 1-5.

[2] Hunt A, Black A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Atlanta, GA, USA: IEEE Press, 1996: 373-376.

[3] Tokuda K, Nankaku Y, Toda T, et al. Speech synthesis based on hidden markov models[J]. Proceedings of the IEEE, 2013, 101(5): 1234-1252.

[4] Le Y, Bengio Y. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[5] Wang Y, Skerry-Ryan R, Stanton D, et al. Tacotron: towards end-to-end speech synthesis[C]//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017: 4006-4010.

[6] Shen J, Pang R, Ron J, et al. Natural TTS synthesis by conditioning wavenet on Mel spectrogram predictions[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada: IEEE Press, 2018: 4779-4783.

[7] Oord A, Dieleman S, Zen H, et al. WaveNet: a generative model for raw audio[C]//Proceedings of the 9th ISCA Speech Synthesis Wokshop. Sunnyvale, USA: Springer Press, 2016: 125-125.

[8] Griffin D, Lim J. Signal estimation from modified short-time Fourier transform[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Boston, Massachusetts, USA: IEEE Press, 1983: 804-807.

[9] Li N, Liu S, Liu Y, et al. Neural speech synthesis with transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, USA: AAAI Press, 2019: 6706-6713.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the Conference on Neural Information Processing Systems. Long Beach, USA: MIT Press, 2017: 5998-6008.

[11] 徐浩森, 姜囡, 齐志坤. 基于注意力机制的卷积循环网络语

音降噪[J]. 科学技术与工程, 2022, 22(5): 1950-1957.
Xu Haosen, Jiang Nan, Qi Zhikun. Speech denoising based on attention mechanism using convolution loop network[J]. Science Technology and Engineering, 2022, 22(5): 1950-1957.

[12] Elias I, Zen H, Shen J, et al. Parallel tacotron: non- autoregressive and controllable tts[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE Press, 2021: 5709-5713.

[13] Łańcucki A. Fastpitch: parallel text-to-speech with pitch prediction[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE Press, 2021: 6588-6592.

[14] Nguyen B, Cardinaux F, Uhlich S. Autotts: end-to-end text-to-speech synthesis through differentiable duration modeling[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE Press, 2023: 1-5.

[15] Cornille T, Wang F, Bekker J. Interactive multi-Level prosody control for expressive speech synthesis[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE Press, 2022: 8312-8316.

[16] Tang H, Zhang X, Wang J, et al. QI-TTS: questioning intonation control for emotional speech synthesis[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE Press, 2023: 1-5.

[17] He L, Sun C, Zhu R, et al. Multi-speaker emotional speech synthesis with limited datasets: two-stage non-parallel training strategy[C]//Proceedings of the 7th International Conference on Intelligent Computing and Signal Processing. Xi'an: IEEE Press, 2022: 545-548.

[18] Ye J, Zhou H, Su Z, et al. Improving cross-lingual speech synthesis with triplet trainingscheme[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE Press, 2022: 6072-6076.

[19] 安鑫, 代子彪, 李阳, 等. 基于 BERT 的端到端语音合成方法[J]. 计算机科学, 2022, 49(4): 221-226.
An Xin, Dai Zibiao, Li Yang, et al. End-to-end speech synthesis method based on BERT[J]. Computer Science, 2022, 49(4): 221-226.

[20] 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 Dilscovery and Data Mining. New York: ACM Press, 2020: 2802-2812.

[21] Zu B, Cai R, Cai Z, Pengmao Z. Research on tibetan speech synthesis based on fastspeech2[C]//Proceedings of the 3rd International Conference on Pattern Recognition and Machine Learning. Chengdu: IEEE Press, 2022: 241-244.

[22] Ren Y, Hu C, Tan X, et al. Fastspeech 2: fast and high-quality end-to-end text to speech[C]//Proceedings of the International Conference on Learning Representations. Virtual Event, Austria: IEEE Press, 2020: 1-15.

[23] Huybrechts G, Merritt T, Comini G, et al. Low-resource expressive text-to-speech using data augmentation[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE Press, 2021: 6593-6597.

[24] 丁云涛, 才让卓玛, 贡保加, 等. 一种基于 WaveNet 的藏语语音合成方法[J]. 计算机仿真, 2023, 40(1): 295-299, 538.
Ding Yuntao, Cai Rangzhuoma, Gong Baojia, et al. A WaveNet-based Tibetan speech synthesis method[J]. Computer Simulation, 2023, 40(1): 295-299, 538.

[25] 刘瑞, 康世胤, 高光来, 等. MonTTS: 完全非自回归的实时、高保真蒙古语语音合成模型[J]. 中文信息学报, 2022, 36(7): 86-97.
Liu Rui, Kang Shiyin, Gao Guanglai, et al. MonTTS: A fully non-autoregressive real-time and high-fidelity Mongolian speech synthesis model[J]. Journal of Chinese Information Processing, 2022, 36(7): 86-97.

[26] 杨琳, 杨鉴, 蔡浩然, 等. 基于迁移学习的越南语语音合成[J]. 计算机科学, 2023, 50(8): 118-124.
Yang Lin, Yang Jian, Cai Haoran, et al. Vietnamese speech synthesis based on transfer learning[J]. Computer Science, 2023, 50(8): 118-124.

[27] Rios A, Amrhein C, Aepli N, et al. On biasing transformer attention towards monotonicity[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Mexico City: NAACL Press, 2021: 4474-4488.

[28] 张学文, 王林, 冯夫健, 等. 基于卷积神经网络的苗语孤立词语音识别[J]. 软件导刊, 2022, 21(2): 21-26.
Zhang Xuewen, Wang Lin, Feng Fujian, et al. Isolated word speech recognition of hmong language based on convolutional neural network[J]. Software Guide, 2022, 21(2): 21-26.

[29] 郭潇楠, 王仁超, 毛三军, 等. 施工组织设计文档智慧辅助审查中的文本分类问题研究[J]. 科学技术与工程, 2022, 22(36): 16180-16188.
Guo Xiaonan, Wang Renchao, Mao Sanjun, et al. Document classification in intelligent aided review of construction organization design documents[J]. Science Technology and Engineering, 2022, 22(36): 16180-16188.

[30] 邹蕾, 崔斌, 樊超, 等. 基于双向编码文本摘要-长短期记忆-注意力的检察建议文本自动生成模型[J]. 科学技术与工程, 2021, 21(25): 10780-10788.
Zou Lei, Cui Bin, Fan Chao, et al. BERTSUM-LSTM- attention based model for the automatic generation of the procuratorial suggestions[J]. Science Technology and Engineering, 2021, 21(25): 10780-10788.