计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600045

基于迁移学习的越南语语音合成

杨 琳[1] 杨 鉴[1] 蔡浩然[1] 刘 聪[2]
1 云南大学信息学院 昆明 650504
2 科大讯飞股份有限公司人工智能研究院 合肥 230088
(yun20yl@mail.ynu.edu.cn)

摘 要 越南语是越南社会主义共和国的官方语言,属南亚语系越芒语族越语支。近年来基于深度学习的语音合成已经能够合成出高质量的语音,然而这类方法通常依赖于大规模的高质量语音训练数据。解决某些低资源非通用语语音训练数据不足问题的一种有效途径为:采用迁移学习方法并借用其他高资源通用语语音数据。在低资源条件下,以提高越南语语音合成质量为目标,选用端到端语音合成模型 Tacotorn2 作为基线模型,采用迁移学习方法研究不同源语言和不同文本字符嵌入方式、迁移学习方式对语音合成效果的影响;然后从主观和客观两方面对文中阐述的各种模型所合成的语音进行测评。实验结果表明,基于英语音素嵌入+越南语音素嵌入方式的迁移学习系统在合成自然可懂的越南语语音上取得了较好的结果,合成语音的 MOS 评分可达 4.11 分,远高于基线系统的 2.53 分。

关键词:越南语;语音合成;迁移学习;文本嵌入;端到端
中图法分类号 TP391

Vietnamese Speech Synthesis Based on Transfer Learning

YANG Lin[1],YANG Jian[1],CAI Haoran[1] and LIU Cong[2]
1 School of Information Science and Engineering,Yunnan University,Kunming,650504,China
2 AI Research Institute,iFLYTEK Co.,Ltd.,Hefei,230088,China

Abstract Vietnamese is the official language of the Socialist Republic of Vietnam. It belongs to the Vietnamese branch of the Viet Muang language family of the South Asian language family. In recent years,deep learning-based speech synthesis has been able to synthesize high-quality speech. However,these methods often rely on large-scale high-quality speech training data. An effective way to solve the problem of insufficient data for some low-resource non-lingua franca speech training is to adopt a transfer learning method and borrow other high-resource lingua franca speech data. Under the condition of low resources,with the goal of improving the quality of Vietnamese speech synthesis,the end-to-end speech synthesis model Tacotorn2 is selected as the baseline model,and the effects of different source languages,different text character embedding methods and transfer learning methods on the effect of speech synthesis are studied by transfer learning methods. Then,from both subjective and objective aspects,the speech synthesized by the various models described in this paper is evaluated. Experimental results show that the transfer learning system based on English phonetic module embedding+Vietnamese phonology embedding method has achieved good results in synthesizing naturally understandable Vietnamese speech,and the MOS score of synthetic speech can reach 4.11,which is much higher than the 2.53 of the baseline system.

Keywords Vietnamese,Speech synthesis,Transfer learning,Text embedding,End-to-end

1 引言

越南语,又称京语或国语,是越南社会主义共和国的官方语言,属南亚语系越芒语族越语支[1]。相比英语、汉语等通用语言,可获得的越南语电子化语音资源规模较小。此外,相比其他通用语,国内外在越南语语音合成方面的研究还不够充分。

语音合成(Text to Speech,TTS)的目的是从给定文本合成可理解和自然的语音[2],是语音、语言和机器学习领域的一个热点研究课题。语音合成技术主要经历了以下 3 个阶段:早期基于计算机的语音合成、基于统计参数的语音合成,以及基于神经网络的语音合成[3-5]。语音合成技术主要包括语言分析部分和声学系统部分,也称为前端部分和后端部分。语言分析部分主要对输入文本信息进行处理,包括分词、字素

到稿日期:2022-06-06 **返修日期**:2023-02-07
基金项目:国家重点研发计划(2020AAA0107901)
This work was supported by the National Key R&D Program of China(2020AAA0107901).
通信作者:杨鉴(jianyang@ynu.edu.cn)

转音素、韵律预测等，提取合成所需的信息；声学系统部分主要是根据语音分析部分提供的信息生成对应的音频，实现发声的功能。声学系统部分目前主要有3种技术实现方式，分别为：波形拼接、参数合成以及端到端的语音合成技术。前两种语音合成技术都需要花费大量的人力来完成数据准备工作，系统稳定性不高，合成语音不够自然。

随着深度学习的发展，基于神经网络的语音合成被提出，它采用深度神经网络作为语音合成的模型主干，不仅简化了合成系统的结构、减少了人工干预，也降低了对语言学相关背景的要求。例如谷歌提出的WaveNet[6]模型，其被视为现代第一个神经TTS模型。其他的模型如DeepVoice1/2[7-8]，Tacotron1/2[3-4]，FastSpeech1/2[9-10]等，真正实现了直接从文本合成语音。其中，谷歌在2017年初提出的Tacotron是第一个相对成熟的端到端语音合成系统，该系统以〈音频，文本〉数据对为输入，输出对应音频的梅尔频谱，再利用Griffin-Lim[11]算法将它转换为语音波形。Tacotron2相较于Tacotron优化了模型的编码器和解码器结构，使用了基于位置敏感的注意力机制，从而达到了更好的语音合成效果。

虽然基于神经网络的语音合成系统效果已经非常自然，但是它需要大量的数据进行训练。许多非通用语由于使用人数少、规范化程度不高、人力物力投入不足等原因，可用做语音合成的数据规模较小，这类语言的语音合成被统称为低资源语音合成。越南语属于非通用语，使用越南语的人数相对较少；此外，越南境内的越语又可分为3种方言，现代越南语的发音以河内腔（北方方言）为标准，但是不少的海外越南侨胞说的是西贡（南方方言）腔的越南语，越南语言学术界对于现代越南语的统一规范标准仍存在争议。

在低资源前提下，为了探索有效的非通用语语音合成方法，本文设计并实现了一个基于迁移学习的越南语语音合成系统。在该系统中，先用汉语普通话和英语这两种高资源语言的语音数据对Tacotron2基线系统进行预训练，然后将预训练模型迁移到越南语语音合成系统中。为了探讨不同源语言对迁移学习效果的影响，本文分别选用汉语普通话和英语作为迁移学习的源语言。在可选的文本嵌入方式中，本文对比了“汉语拼音字符嵌入＋越南语字符嵌入”“汉语声韵母嵌入＋越南语声韵母嵌入”“汉语音素嵌入＋越南语音素嵌入”“英语字符嵌入＋越南语字符嵌入”“英语音素嵌入＋越南语音素嵌入”5种不同组合的嵌入方式对迁移学习效果的影响。本文为低资源条件下快速开发高质量越南语语音合成系统奠定了良好基础，为研究其他低资源非通用语语音合成方法提供了良好借鉴。

2　相关理论及模型

2.1　Tacotron2

Tacotron2是由Google Brain于2018年提出的一个端到端语音合成框架。图1为Tacotron2的模型结构。模型由两部分组成：声谱预测网络和声码器。

声谱预测网络主要由编码器和包含注意力机制的解码器组成。编码器将字符序列转换成一个隐层表征，解码器接受这个隐层表征用以预测声谱图。其中编码器模块包含一个字符嵌入层、一个3层卷积、一个双向LSTM（长短期记忆网络）层。输入字符首先被编码成512维的字符向量，然后穿过一个3层卷积，每层卷积包含512个5×1的卷积核，即每个卷积核横跨5个字符，卷积层会对输入的字符序列进行大跨度上下文建模，最后一个卷积层的输出被传送到一个双向的LSTM层用以生成编码特征，这个LSTM包含512个单元（每个方向256个单元）。

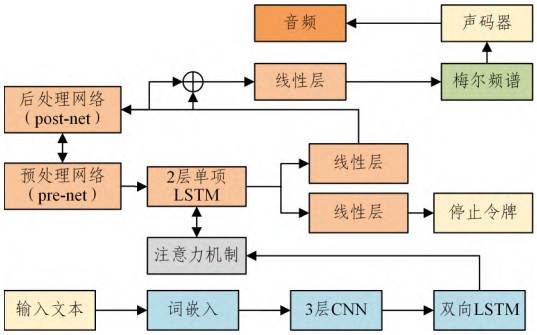


图1　Tacotron2模型结构

Fig. 1　Tacotron2 model structure

解码器是一个自回归的具有位置敏感注意力的递归神经网络，利用解码器生成的特征向量循环预测梅尔频谱，一次预测一帧。解码器预测出的频谱首先被传入一个pre-net网络，它也被称为信息瓶颈层，是一个两层预处理网络，每层包含256个隐藏ReLU单元。pre-net的输出和注意力上下文向量拼接在一起，传给一个两层堆叠的由1 024个单元组成的单向LSTM。LSTM的输出再次和注意力上下文向量拼接在一起，然后经过一个线性投影来预测目标频谱帧。最后，目标频谱帧经过一个5层卷积的post-net层来预测一个残差叠加到卷积前的频谱帧上，以改善频谱重构的整个过程。post-net每层由512个5×1卷积核组成。并行于频谱帧的预测，解码器LSTM的输出与注意力上下文向量拼接在一起，作为停止令牌，来预测输出序列完成与否的概率。

本文使用的声码器是Waveglow，用于把梅尔频谱特征表达逆变换为时域波形样本。

2.2　迁移学习

迁移学习（Transfer Learning）是机器学习中的一种方法，指一种学习对另一种学习的影响，或习得的经验对完成其他活动的影响[12]。迁移广泛存在于各种知识、技能与社会规范的学习中。随着越来越多的机器学习应用场景的出现，要得到表现较好的监督学习需要大量的标注数据，这是一项枯燥无味且需花费巨大人力物力的任务，因此迁移学习越来越受到人们的关注。文献[13]利用迁移学习进行土壤湿度的预测；文献[14]则利用多源迁移学习方法预测大坝裂缝。迁移学习的主要思想就是从相关领域中迁移标注数据或者知识结构，完成或改进目标领域或任务的学习效果[15]。

近年来，语音合成任务越来越国际化，语言交叉使用的现象也变得非常普遍。处理多语言交叉的文本，最直观的方案就是让一个人录制多种语言的训练语料，然后使用这些语料分别训练不同的模型，在语音合成阶段分别合成相应语言的语音。但是这种方法的缺点也很明显，一是语料录制难，花费金额大；二是合成的语音衔接不自然。现在主流的方案是

训练多语言模型，但是这种方案也需要一个发音人的多语言语料库。因此利用多人的单语言语料来训练多语言模型成为近期研究的热点，迁移学习就是解决这个问题的重要手段之一。另一方面，端到端语音合成模型需要大量的训练数据，许多非通用语可用的训练数据非常少，这会大大影响语音合成质量。直接录制大量该语言的训练数据不仅会耗费大量人力物力财力，而且某些语言由于使用人数太少无法获取到大量高质量的训练语料，迁移学习成为解决这类问题的有效方法之一。利用有大量训练数据的语料对模型进行“预训练”，再把学习到的知识迁移到低资源语言上，可以大大提高目标语言的语音合成质量。

到目前为止，有关语音合成方面迁移学习的文章很多。一方面集中于设计多语言合成的系统架构并不断对系统模型进行改进[16-17]；另一方面集中于设计多语言语音合成的输入格式[18]。

3 越南语语音合成模型

3.1 模型结构

模型采用 Tacotron2 为基线系统，先用源语言语音数据对模型进行预训练，再将预训练的模型用于越南语的语音合成。整体模型如图 2 所示。在文本嵌入模块，本文设计了不同的嵌入方式并对比了不同嵌入方式对迁移学习效果的影响。

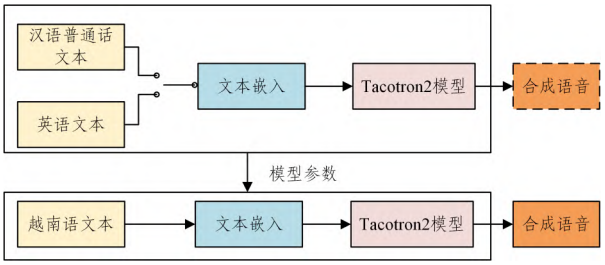


图 2 模型结构

Fig. 2 Model structure

3.2 源语言与预训练

深度神经网络模型在语音合成领域已经取得了很大的成功，但许多研究发现，它们面临的关键挑战之一是数据匮乏。深度神经网络通常具有大量参数，因此在训练数据较少的情况下，很容易陷入过拟合或者泛化能力差的困境。解决这个问题的关键点在于迁移学习的引入。迁移学习可以被大致划分为两个阶段：从一个或多个源任务中获取知识的预训练阶段以及将获取的知识转移到目标任务的微调阶段[19]。

目前有关迁移学习的研究如火如荼，但是，如何形式化地描述所要迁移的知识，使用何种方法进行迁移，以及如何选择一种可行的方法将知识从源任务转移到目标任务是非常重要的。为此，提出了各种预训练方法作为源任务和目标任务之间的桥梁，比如 BERT，XLNET，RoBERTa，BART 等。

目前，这些预训练模型的源语言大都基于大型英语语料库，并且在许多测试中取得了巨大的成功。但是，世界上的语言种类数以万计，为每一种语言都训练一个大型的预训练模型几乎是不可能的。因此，训练一个模型来学习多语言的表征方法是迁移学习目前要解决的主要问题。文献[20]提出的 mBERT 模型使用 104 种语言进行预训练。之后的研究表明，多语言语料规模越大，模型性能就越好。然而预训练模型语料规模的增大就意味着人力物力的进一步增加，因此，探索源语言与目标语言之间的关系是很有必要的。本文就这一关系进行探讨，期望可以在预训练语言和目标语言都很少的情况下快速搭建越南语语音合成模型。

现代越南语使用拉丁字母书写，是一种拼音化文字。越南语一共有 29 个字符，6 个声调。语音包括单元音 11 个，双元音 3 个，辅音 25 个。越南语构词的主要特点是每一个音节常常是一个有语义的单位，可以独立使用，这些单位又可作为构成多音节词的基础。绝大部分多音节词是双音节。

为讨论源语言与越南语的相似程度是否会对迁移学习的效果产生影响，本文分别选取了汉语普通话和英语作为源语言来对越南语进行迁移学习。其中汉语普通话与越南语都属于“有声调语言”，两者在句子结构、词汇构成、音节构成、声韵母发音和声调发音等方面都有很多共同点。据粗略统计，越南语大约有 70%的通用词语与汉语有亲缘关系[21]；英语除某些字符和越南语相同之外，其余方面两者的相似程度较低。

3.3 合成基元与文本嵌入方式

在文语转换系统中，可选择不同层级的文本或语音单元作为语音合成基本单元。以汉语为例，可选择汉字词、汉语拼音词、汉字、汉字拼音、声韵母、音素、拼音字符等。在神经网络语音合成系统中，还需采用文本嵌入技术将语音合成基元转换为向量表示。合成基元不同，所对应的文本嵌入方式也不同。

与汉语普通话类似，越南语也可选择越南语字、声韵母、音素、拼音字符等作为合成基元，英语则可以选用字符嵌入、词嵌入或者音素嵌入等作为合成基元。不同的文本嵌入方式会影响语音合成的质量，本文对这几种文本嵌入方式做了对比。具体嵌入方式如图 3 所示。

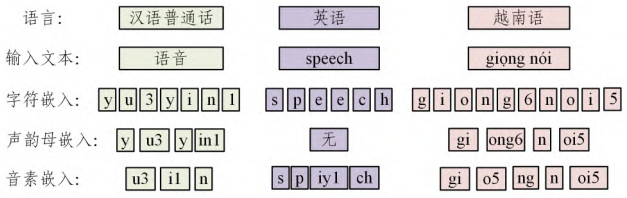


图 3 不同文本嵌入方式

Fig. 3 Different text embedding methods

3.3.1 汉语普通话文本嵌入方式

(1)拼音字符嵌入

汉语将汉字转换为拼音，将单独的拼音字符作为输入，声调编码为 1—5 的数字单独嵌入，轻声为 5。

(2)声韵母嵌入

汉语包括声母 21 个，韵母 35 个，声调编码为 1—5 的数字，随着韵母嵌入。

(3)音素嵌入

汉语的音素包括辅音音素 22 个，单元音音素 6 个，复元音音素 13 个，声调编码为 1—5 的数字，随着元音嵌入。

3.3.2　英语文本嵌入方式

(1)字符嵌入

英语包括 26 个英文字母，每个字符单独嵌入。

(2)音素嵌入

英语的音素嵌入包含 39 个音素以及 3 个重音标记，用 0，1，2 表示。其中，0 表示无重音，1 表示主重音，2 表示次重音。重音标记在元音后。

3.3.3　越南语文本嵌入方式

(1)字符嵌入

字符嵌入之前，由于越南语有 7 个自己独有的字符不在 26 个英文字母中，因此本文首先对这些字符做了预处理，替换规则如表 1 所列。越南语以替换之后的文本为输入文本，声调编码为 1—6 的数字并单独嵌入。

表 1　越南语特殊字符预处理规则

Table 1　Vietnamese special character preprocessing rules

越南语字符	đ	ă	â	ê	ô	ư	ơ
预处理注音	dd	ar	ae	ei	ou	aw	or

(2)声韵母嵌入

越南语语音体系复杂，韵母数量多达 100 多个，且越南语拉丁化时间较短，对于韵母的分类还没有明确的规定。本文中越南语的声母划分为 27 个，韵母划分为 162 个，声调编码为 1—6 的数字。由于越南语中主要是由韵母承担发音功能，因此声调随着韵母嵌入。

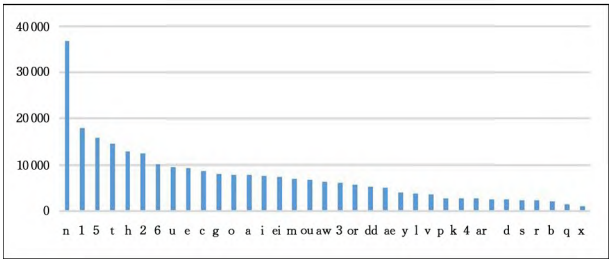
(3)音素嵌入

本文将越南语的音素划分为辅音音素 28 个，单元音音素 12 个，复元音音素 34 个，声调编码为 1-6 的数字。其中声调随着元音音素嵌入，原因有以下两点：1)在越南语中，韵母主要承担发音功能，而元音既可以充当单韵母也可以组合成复韵母；2)元音在带辅音韵尾的韵母中是主要的发音音素。

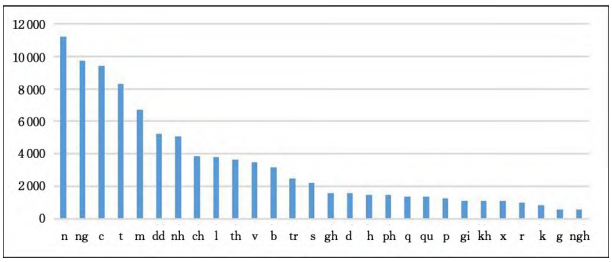
不同嵌入方式会使得标签集的规模不同。对于越南语来说，使用字符嵌入方式的标签集有 35 个，包括越南语 29 个字母和 6 个声调；使用声韵母嵌入方式的标签集有 999 个，包括声母 27 个和带声调的韵母 972 个；音素嵌入方式的标签集有 304 个，包括辅音音素 28 个和带声调的元音音素 276 个。不同文本嵌入方式也会使得标签在文本中的分布不同，本文选取了实验所使用的越南语中 3 960 句文本统计了不同嵌入方式的标签分布情况，如图 4 所示。

图 4(a)统计了越南语文本中所有字符标签出现的频次，可以看出字符标签分布很均匀，没有出现个别字符仅出现几次的情况，绝大部分字符标签频次在 5 000 次以上。图 4(b)统计了越南语文本中声母标签出现的频次，可以看出标签分布较为均匀，一半的标签出现频次均在 2 000 次以上，没有出现个别字符仅出现几次的情况；同时，声母也包括了越南语中的辅音音素，因此图 4(b)也代表了音素嵌入中辅音标签的出现频次。

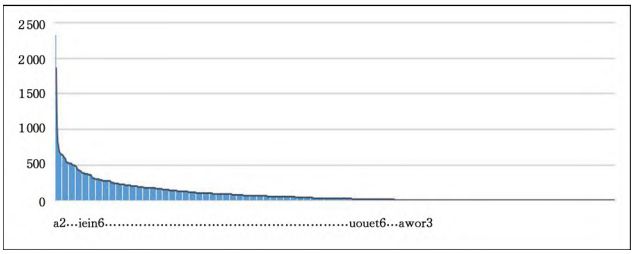
越南语的韵母体系庞大复杂，图 4(c)和图 4(d)分别统计了越南语声韵母嵌入时的韵母出现频次和越南语音素嵌入时的元音音素出现频次。对比图 4(c)和图 4(d)可以看出，相比声韵母嵌入时的韵母标签，使用音素嵌入时，元音音素的标签分布更为均匀且出现频次在 500 以上的标签更多。



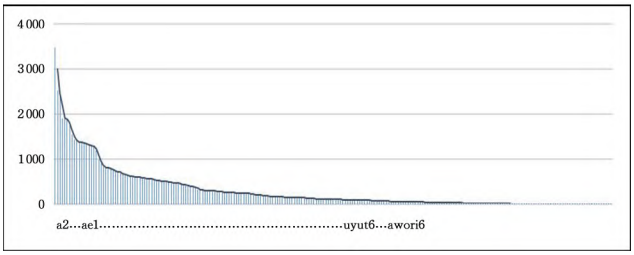
(a)越南语文本中字符标签频次统计



(b)越南语文本中声母(辅音音素)标签频次统计



(c)越南语文本中韵母标签频次统计



(d)越南语文本中元音音素标签频次统计

图 4　不同文本嵌入方式文本字符标签频次统计

Fig. 4　Frequency statistics of text character labels in different text embedding methods

总体来说，采用字符嵌入的标签集数量最少且分布比较均匀，但这种字符嵌入未考虑语言特点，将声调单独嵌入没有考虑到发音的协同性；采用声韵母嵌入和音素嵌入的标签仅在韵母和元音音素上相差较大，其中采用音素嵌入的元音音素标签分布更均匀；同时，采用音素嵌入的标签集规模也小于采用声韵母嵌入的标签集规模。

4　实验

4.1　实验平台及数据

实验中使用的数据库具体数据如表 2 所列。其中汉语普通话数据集时长 84 h，由标贝(北京)科技有限公司提供的时长约为 12 h 的开源中文语音合成数据库和希尔贝壳中文普通话语音数据 AISHELL-3 中的部分语音组成(时长为 72 h)[22]。前者是标准普通话女声，录制环境为专业录音室和录音软件。后者是多说话人语音，录制环境为安静的室内。英语数据集来自一个公开可用的 LJ Speech 数据集，它由

母语为英语的女性录制，时长为 24 h。实验中使用的越南语数据时长为 7h，由专业的女性播音员录制。所有音频的采样率统一为 22050 Hz。整个实验基于 PyTorch 深度学习框架搭建模型，使用一块英伟达 RTX3090 显卡来训练模型，模型的训练批次(batch size)都设置为 40。

表 2 实验数据

Table 2 Experimental data

数据集		时长/h	平行语料对(句)		
			总数据集	训练集	测试集
汉语普通话	标贝数据集	12	10000	8752	1248
	希尔贝壳数据集(部分)	72	63263	52301	10902
英语	LJ Speech 数据集(部分)	12	6653	5703	950
	LJ Speech 数据集	24	13100	11229	1817
越南语		7	3960	3880	80

4.2 实验设计

4.2.1 不同源语言迁移学习语音合成实验

为了探讨不同源语言对迁移学习效果的影响，本文设计了以下实验，具体方案如表 3 所列。

表 3 不同源语言迁移学习语音合成实验方案

Table 3 Experimental scheme of speech synthesis for transfer learning from different source languages

实验序号	源语言：文本嵌入方式，时长	目标语言：文本嵌入方式，时长	模型名称
1	无	越南语：字符嵌入，7 h	model1
2	汉语普通话：拼音字符嵌入，12 h	越南语：字符嵌入，7 h	model2
3	英语：字符嵌入，12 h	越南语：字符嵌入，7 h	model3

本次实验使用的汉语普通话数据集为上述提到的标贝数据集，英语数据集为 LJ Speech(部分)数据集，其余数据参数和表 2 一致。

4.2.2 不同文本嵌入方式迁移学习语音合成实验

为了探讨不同文本嵌入方式对迁移学习效果的影响，并且排除源语言模型训练不完全对迁移学习的影响，本次实验使用的汉语普通话数据库为 3.1 节中提到的 84h 的数据库；平行语料共有 73263 对〈文本，音频〉，其中 61053 对为训练集，12210 对为测试集；英语数据库为 LJ Speech 数据集；其余数据参数和表 2 中一致。本文设计了以下实验，如表 4 所列。

表 4 不同文本嵌入方式迁移学习语音合成实验方案

Table 4 Experimental scheme of speech synthesis for transfer learning from different embedding methods

实验序号	源语言：文本嵌入方式，时长	目标语言：文本嵌入方式，时长	模型名称
1	无	越南语：字符嵌入，7 h	model4
2	无	越南语：声韵母嵌入，7 h	model5
3	无	越南语：音素嵌入，7 h	model6
4	汉语普通话：拼音字符嵌入，84 h	越南语：字符嵌入，7 h	model7
5	汉语普通话：声韵母嵌入，84 h	越南语：声韵母嵌入，7 h	model8
6	汉语普通话：音素嵌入，84 h	越南语：音素嵌入，7 h	model9
7	英语：字符嵌入，24 h	越南语：字符嵌入，7 h	model10
8	英语：音素嵌入，24 h	越南语：音素嵌入，7 h	model11

4.3 评测方法

本文采用的评测方法分为客观评测和主观评测。

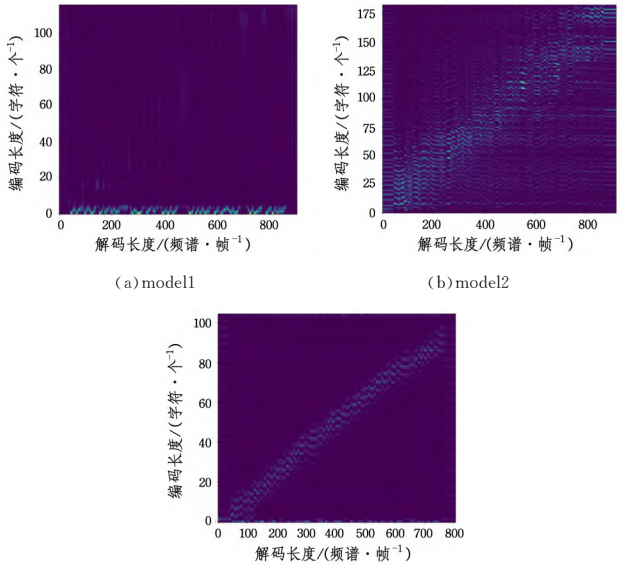
客观评测使用 Tacotron2 系统的注意力对齐结果图。注意力对齐结果呈对角线状，表示在生成音频序列时，解码器集中在正确的音素上，保证了每个字符的正确发音。通过经验观察，注意力对齐结果与合成音频的质量密切相关。虽然目前还无法用一种简单的方法去量化这种相关性，但是注意力对齐的结果可以大致反映模型的优劣。

主观评测使用模型偏好测试和 MOS(Mean Opinion Score)平均主观意见分。模型偏好测试邀请被测试者对不同模型合成的语音进行偏好选择，是语音合成实验中评测模型好坏的一种常用方法。MOS 评分则依靠人的听觉印象来对听到的语音进行打分。在国际标准中，统一使用 MOS 值来评价系统合成的话音质量。分值为 1～5 分，1 分最差，5 分最好。

4.4 实验结果与分析

4.4.1 不同源语言迁移学习语音合成实验结果

图 5 为可视化注意力对齐结果。对比图 5(a)和图 5(b)、图 5(c)可以看出，后两者的注意力对齐曲线比前者亮，且明显更集中在对角线上，这说明迁移学习的确可以提升越南语的语音合成质量；对比图 5(b)和图 5(c)可以看出，使用英语作为迁移学习的源语言比使用汉语普通话的效果更好。



(a) model1 (b) model2

(c) model3

图 5 不同源语言实验可视化注意力对齐结果

Fig. 5 Visualization attention alignment results in different source languages experiments

邀请 15 位越南语专业的同学对两个模型合成的 14 句语音进行模型偏好性测试，结果如图 6 所示。可以看出，使用英语做迁移学习的模型得到了更多的偏好。测试者认为，model3 合成的语音更流畅，听起来更像越南语，但两个模型合成语音的可懂度都较差。

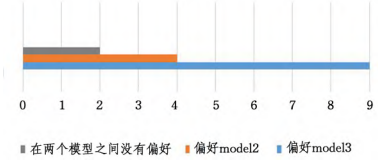


图 6 模型偏好测试

Fig. 6 Model preference test

综合客观评测和主观评测两方面的结果可以看出，选择英语作为源语言时模型的效果优于选择汉语普通话作为源语言时的模型效果。分析原因可能有以下几点：1）英语和越南语都属于拼音文字，相较于汉语普通话，英语的结构和韵律都较为简单，同样时长的训练数据，英语的模型训练更加充分，之后的迁移学习效果就更好；2）Tacotron2 模型本不依赖语言知识来合成语音，在合成基元为字符的情况下，弱化了语言之间的差异，更多的是学习基本的发音规则，这就导致汉语普通话和越南语的结构、声韵母等之间的相似性作用不大；3）虽然汉语普通话和越南语在声韵母的发音上有很多相似之处，例如，越南语中字母“a”的发音与汉语普通话中“a”发音相同，但在词嵌入层面没有捕捉到这种特性，从而让模型赋予它们相同的值，这也会削弱语言相关性对迁移学习效果的影响。

4.4.2　不同文本嵌入方式迁移学习语音合成实验结果

图 7 是可视化注意力对齐结果。对比图 7(a)、图 7(b)和图 7(c)可以看出，在不进行迁移学习的情况下使用音素嵌入的模型合成语音效果要好于字符嵌入和声韵母嵌入；对比图 7(d)、图 7(e)、图 7(f)和图(g)、图 7(h)可以明显看出，无论使用哪种语言作为迁移学习源语言，采用音素嵌入的效果都是最好的，字符嵌入次之，声韵母嵌入（仅针对汉语普通话）的效果最差。

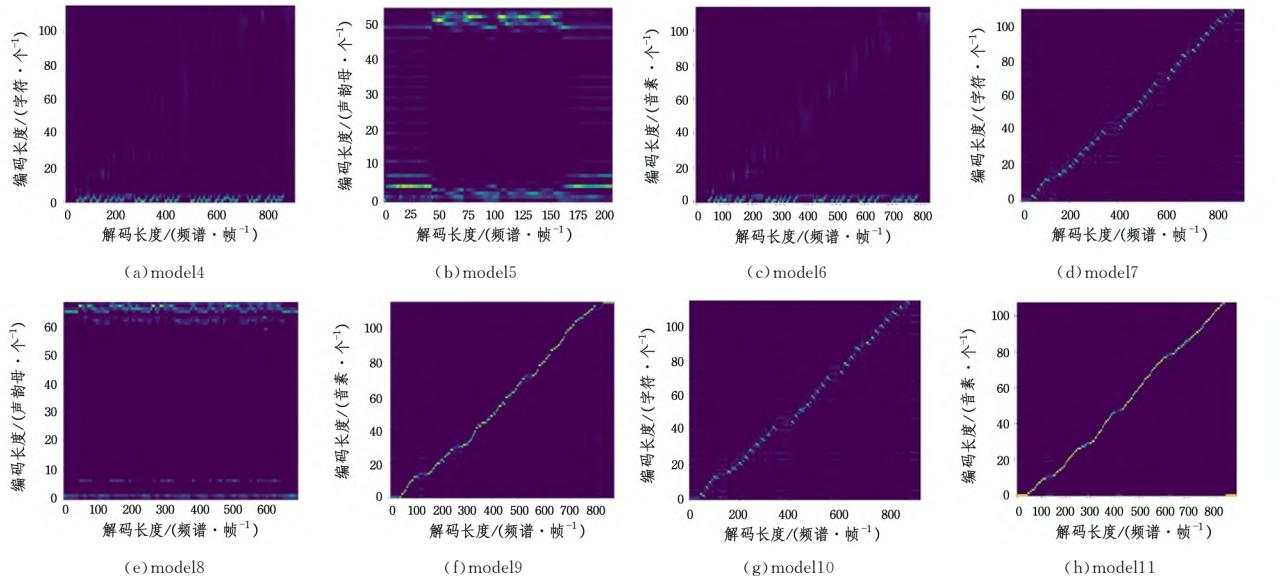


图 7　不同文本嵌入方式实验可视化注意力对齐结果

Fig. 7　Visualization attention alignment results in different text embedding methods experiments

本次评测邀请了 8 位越南语专业的同学以及 3 位非越南语专业的同学对合成的测试集中 120 句（每个模型挑选 10 句语音）语音以及原始音频中 30 句语音进行 MOS 评分。在对所有模型合成的音频以及原始音频进行主观评分之后，对获得的数据进行了分析处理，得到了如表 5 所列的结果。结论和客观评测的结论一致，使用迁移学习可以提高越南语的语音合成质量，与基线系统相比，使用迁移学习的语言 MOS 评分平均提高了 1.5 左右；使用音素嵌入的模型合成语音效果是最好的，其中 model8 的效果最好，相比 model3 合成语音 MOS 评分提高了 1.58。

表 5　语音主观评测结果

Table 5　Subjective evaluation results of voice

模型	MOS 值
原始音频	4.97
model4	2.32
model5	2.30
model6	2.53
model7	3.88
model8	2.04
model9	4.05
model10	3.76
model11	4.11

综合客观评测和主观评测的结果可以得出结论：无论在何种情况下，音素嵌入都能达到最好的结果；相比汉语普通话，选择英语作为源语言不仅能提高模型合成语音的质量，在合成语音质量相当的情况下还能减少源语言的训练数据时长。分析出现以上结果的原因可能是：1）使用字符嵌入作为文本嵌入的方式时，虽然标签集的规模最小，但是字符嵌入弱化了文本和发音之间的关系，这种简单的嵌入方式也未考虑协同发音的问题，这都会影响模型效果。2）使用声韵母嵌入作为文本嵌入的方式时，由于越南语的声韵母体系复杂，目前对于韵母的分类和数量规范程度还不够，在系统学习发音时会造成干扰；其次，庞大的声韵母体系导致标签集过大；最后，如图 4 所示，采用声韵母嵌入方式时，越南语的标签分布最不均匀，以上原因都会影响模型效果。3）不同文本嵌入时，对越南语平均音子时长（音频时长与标签集标签个数之比）进行统计，字符嵌入为 12 min，声韵母嵌入为 0.42 min，音素嵌入为 1.38 min，使用音素嵌入作为文本嵌入的方式时，平均音子时长居中，这种情况下越南语的标签集规模适中，分布也较为均匀，并且充分考虑了协同发音问题，音素与发音的关系更加紧密。因此无论使用何种源语言，使用音素嵌入作为文本嵌入方式时，模型效果都最好。

结束语　本文围绕低资源越南语语音合成，提出了基于迁移学习的越南语语音合成模型。对比研究了不用源语言、不同文本嵌入方式对迁移学习效果的影响。综合以上实验

结果可以得出以下结论：1)当越南语作为迁移学习的目标语言时，选择英语作为源语言进行迁移学习，其效果优于选择汉语普通话作为源语言的效果，源语言与目标语言之间的相似性并未直接影响迁移学习效果；2)预训练模型训练的充分度会影响迁移学习的效果，源语言模型训练越充分，迁移学习的效果就越好；3)无论是越南语单独训练还是引入迁移学习，音素嵌入的效果都是最优的，且采用迁移学习时不同文本嵌入方式对模型的影响要大于不采用迁移学习；4)在对比实验中，采用“英语音素嵌入＋越南语音素嵌入”的迁移学习方式得到的越南语语音合成效果最好。

综上所述，本文实验中，源语言与目标语言之间的相似性并未直接影响迁移学习效果，这应该与本文采用的迁移学习方式有关。后续研究中，将针对源语言和目标语言的语音特征、语音相似性，探索更为有效的字符嵌入迁移学习方法。此外，本文实验中的合成语音相比原始音频在流畅度和自然度上还有进步空间，作者将尝试通过添加韵律信息等方法进一步提高越南语语音合成的自然度。

参 考 文 献

[1] YANG J. An analysis of the linguistic family of the nanking people in Vietnam [J]. Ideological Front, 2012, 38(2): 133-134.

[2] TAN X, QIN T, SOONG F, et al. A survey on neural speech synthesis [J]. arXiv: 2106.15561, 2021.

[3] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards End-to-End Speech Synthesis[C]// Proceedings of Conference of the International Speech Communication Association. Stockholm, Sweden, 2017: 4006-4010.

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

[5] PING W, PENG K, GIBIANSKY A, et al. Deep Voice 3: 2000-Speaker Neural Text-to-Speech[C]// Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2017: 1-15.

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

[7] ARIK S Ö, CHRZANOWSKI M, COATES A, et al. Deep voice: Real-time neural Text-to-Speech[C]// International Conference on Machine Learning. PMLR, 2017: 195-204.

[8] GIBIANSKY A, ARIK S, DIAMOS G, et al. Deep voice 2: Multi-speaker neural Text-to-Speech[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-15.

[9] REN Y, RUAN Y, TAN X, et al. FastSpeech: Fast, Robust and Controllable Text to Speech[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 3171-3180.

[10] REN Y, HU C, TAN X, et al. Fastspeech 2: Fast and High-Quality End-to-End Text to Speech[C]// Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2020: 1-15.

[11] GRIFFIN D, LIM J. Signal estimation from modified short-time Fourier transform[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(2): 236-243.

[12] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks? [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems (Volume 2). 2014: 3320-3328.

[13] WANG X Z, LI Q L, LI W H. Spatio-temporal model of soil moisture prediction integrated with transfer learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(3): 675-683.

[14] WANG J f, LIU F, YANG S, et al. Dam Crack Detection Based on Multi-source Transfer Learning[J]. Computer Science, 2022, 49(6A): 319-324.

[15] PAN S J, YANG Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.

[16] ZHANG Y, WEISS R J, ZEN H, et al. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning[C]// Proceedings of Conference of the International Speech Communication Association. Graz, Austria, 2019: 2080-2084.

[17] NEKVINDA T, DUŠEK O. One Model, Many Languages: Meta-Learning for Multilingual Text-to-Speech[C]// Proceedings of Conference of the International Speech Communication Association. Shanghai, China, 2020: 2972-2976.

[18] ZHOU X, TIAN X, LEE G, et al. End-to-end code-switching TTS with cross-lingual language model[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020). IEEE, 2020: 7614-7618.

[19] HAN X, ZHANG Z, DING N, et al. Pre-trained models: past, present and future [J]. AI Open, 2021, 2: 225-250.

[20] PAPADIMITRIOU I, CHI E A, FUTRELL R, et al. Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 2522-2532.

[21] CHENG F. An Introduction to Modern Vietnamese [D]. Nanning: Guangxi Minzu University, 1988.

[22] SHI Y, BU H, XU X, et al. Aishell-3: A multi-speakerMandarin TTS corpus and the baselines[J]. arXiv: 2010.11567, 2020.

YANG Lin, born in 1999, postgraduate. Her main research interest is speech synthesis, recognition and understanding.

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

（责任编辑：何杨）