第 42 卷 第 1 期
2023 年 1 月

应用声学
Journal of Applied Acoustics

Vol. 42, No. 1
January, 2023

◇ 研究报告 ◇

汉语儿童情感语声合成*

胡航烨　王　蔚†

(南京师范大学教育科学学院机器学习与认知实验室　南京　210097)

摘要: 情感语声合成技术对于人机交互具有重要的意义。面对儿童情感语声合成所需汉语语声数据资源缺乏以及模型训练时长较长等问题，该文提出利用迁移学习实现汉语儿童情感语声合成的方法。首先基于汉语语声数据库训练深度学习模型实现中文语声端到端合成模型，再使用高质量大样本的中文情感语料库完成情感语声合成模型，最后利用自行采样的小样本汉语儿童情感语料对模型进行迁移学习实现低资源的语声合成。客观实验结果中梅尔倒谱失真指标为4.91，主观听辨实验指标分别为3.61和4.17。通过实验对比表明，该文的方法在情感语声合成技术的应用上具有良好的性能表现，并且优于现有先进的低资源情感语声合成方法。

关键词: 儿童；情感语声合成；迁移学习；低资源

中图法分类号: TP391　　**文献标识码:** A　　**文章编号:** 1000-310X(2023)01-0076-08

DOI: 10.11684/j.issn.1000-310X.2023.01.010

Affective speech synthesis of Chinese children

HU Hangye　WANG Wei

(*School of Educational Science, Nanjing Normal University, Nanjing* 210097*, China*)

Abstract: Emotional speech synthesis technology is of great significance for human-computer interaction. Facing the lack of Chinese speech data resources required for children's emotional speech synthesis and the long time of model training, this paper proposes a method of using transfer learning to realize Chinese children's emotional speech synthesis. This paper first implements the Chinese speech end-to-end synthesis model based on the Chinese speech database training depth learning model, then uses the high-quality and large sample Chinese emotional corpus to complete the emotional speech synthesis model, and finally uses the self sampled small sample Chinese children's emotional corpus to transfer the model to realize low resource speech synthesis. The objective experimental results show that the Mel cepstrum distortion index is 4.91, and the subjective auditory discrimination experimental indexes are 3.61 and 4.17 respectively. The experimental comparison shows that the method in this paper has good performance in the application of emotional speech synthesis technology, and is better than the existing advanced low resource emotional speech synthesis methods.

Keywords: Children; Emotion speech synthesis; Transfer learning; Low resource

2021-10-10 收稿; 2022-01-18 定稿

*国家社会科学基金项目 (BCA150054)

作者简介: 胡航烨 (1996–), 女, 浙江东阳人, 硕士研究生, 研究方向: 信号与信息处理。

†通信作者 E-mail: wangwei5@njnu.edu.cn

0 引言

情感语声合成技术作为近年来人机交互的热点问题受到越来越多研究者的关注。当情感语声合成技术应用至儿童的情感交互中时，会更关注到儿童情感的变化以及与成人不同的韵律特征。随着年龄的增长，儿童的韵律特征在说话速率、音高基频、共振峰方面都具有明显的变化[1] 研究表明，儿童的韵律特征相比较于成人的韵律特征具有更高的可变性[2]，并且根据对儿童定向言语的调查可以发现儿童对韵律变化明显的语声更感兴趣[3]。因此当情感语声合成技术应用至儿童情感交互中时，会更偏向于使用具有儿童本身韵律的情感语声。然而现阶段的情感语声合成技术大多基于高质量的情感语料，儿童情感语料库匮乏，其主要原因包含两点：(1) 儿童单一情感语声采集困难，不便于控制建立离散型的情感模型；(2) 在儿童的情感语声中其韵律变化范围十分广泛，建模难度较大。因此本文研究的低资源的儿童情感语声合成对于目前人机交互中日益增长的情感需求来说具有极其重要的意义。

儿童情感语声合成技术的发展可以追溯到传统的基于隐马尔可夫模型 (Hidden Markov model, HMM) 的统计参数语声合成时代[4−7]。Strömbergsson 等[8] 的研究介绍了一种通过串联不同说话人的语声来分段重新合成儿童语声的新颖方法。而随着深度神经网络的广泛应用[9−11]，情感语声合成已经发展到多种解决方案，加入变分自动编码器 (VAE) 以非监督的方式学习复杂的分布，从而得到大量不同数据[12]；Li 等[13] 的研究利用成人演绎的大量儿童情感语声，使用基于序列到序列 (seq2seq) 的 Tacotron，在编码器之前以及解码器输出之后分别插入情感分类器，以增强情感嵌入和预测梅尔谱的情感识别能力，实现可控的儿童情感语声合成。针对儿童情感语声合成现阶段存在的问题，低资源语声合成方法包括迁移学习[14]、微调和多任务学习[15] 等技术，在低资源的各种应用中被证明是有用的。例如在文献 [16] 中，研究者成功地将知识从一个被训练来区分说话人的模型转移到一个多说话人 TTS 模型。文献 [17] 中使用基于微调的说话人自适应方法，用于利用低资源的数据构建 TTS 模型。文献 [18] 中修改了 Tacotron 的结构来合成给定情感标签的语声来实现低资源化的情感语声合成。文献 [19] 中研究者利用语声转换技术实现数据增强的低资源情感语声合成。

本文基于以上研究者的工作提出利用迁移学习以及给定情感标签的有监督学习方式实现低资源的儿童情感语声合成，包括 3 个阶段，分别为汉语语声合成模型的建立、情感嵌入空间和儿童特征迁移阶段。主要工作如下：(1) 建立了包含 4 种情感的儿童离散型情感语料库；(2) 用迁移学习的方式实现低资源的儿童情感语声合成。

1 相关技术

1.1 汉语语声合成模型

汉语语声合成模型基于 Google 的 Brain 团队[20] 在 2017 年提出来的 Tacotron2 模型。该模型由 3 部分组成，一个引入了注意力机制的基于循环的 seq2seq 特征预测网络，一个基于 WaveNet 修改版的声码器以及一个利用梅尔频率声谱图的连接层。

汉语语声合成与英文语声合成相比较而言存在一定的困难，如韵律较为复杂、存在多音字及变调音等问题。针对这些问题，不少研究者对 Tacotron2 模型进行了改进，如对预训练模块、注意力机制、停止符预测等[21]。

为了解决中文较为复杂的韵律变化问题，文献 [22] 将位置敏感的注意力 (Location-sensitive attention) 扩展为多头位置注意力机制 (Multi head location-sensitive attention)，即

$$\mathrm{head}_i = \mathrm{Attention}(HW_i^H + SW_i^S + FW_i^F), \quad (1)$$

式 (1) 中，H 是编码器的输出，S 代表解码器的输出，F 为累加的注意力权重，而 W_i^H、W_i^S、W_i^F 作为待训练的一系列参数，其子注意力模块的权重是不共享的。多头注意力输出表示为

$$\begin{aligned}&\mathrm{MultiHead}(S,H,F)\\&=\mathrm{concat}(\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_\mathrm{h})W^o,\end{aligned} \quad (2)$$

其中，W^o 为待训练的参数。多头注意力机制将 S、H、F 通过矩阵映射再进行 Attention 运算，通过 Attention 运算之后再把多个子注意力的结果进行拼接，使得解码器在预测声频时，字和字之间的衔接部分，整个句子的韵律变化会更加接近于真实人声。

1.2 情感编码器

情感编码器由Skerry-Ryan等[23]在Tacotron语声合成架构的基础上进行的扩展，加入情感嵌入空间使其能够从包含想要韵律的声学表征中学习韵律的隐藏嵌入空间，实现韵律迁移。

如图1所示，模块以长度为L_R和维度为d_R的梅尔谱图作为输入，从中计算出维度为d_p的嵌入向量，在情感嵌入空间中每一个嵌入向量都对应一种情感特征。输入信号经过6层卷积神经网络(Convolutional neural networks, CNN)之后进入到单元大小为128的一层门控循环单元(Gated recurrent unit, GRU)网络[24]，通过注意力机制最后输出维度为d_p的情感嵌入向量来表示情感特征。

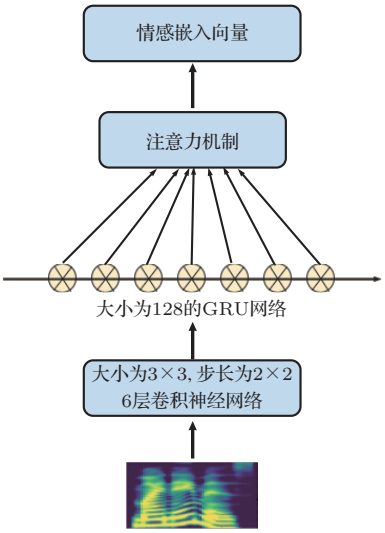


图1 情感嵌入空间

Fig. 1 Emotion embedded space

本文学习了汉语端到端语声合成模型以及加入情感编码器的情感语声合成模型，通过迁移学习的方式将两者结合进行改进。在文献[25]中，研究者利用深卷积语声合成(Deep convolutional TTS, DCTTS)模型进行低资源语声合成，不同于该方法本文采用序列到序列模型(seq2seq)加改进后的注意力机制形成的语声合成模型，以对情感编码器和说话人编码器中进行儿童的情感迁移来实现低资源的汉语儿童情感语声合成。在实验结果中可以看到本文的合成效果优于其他先进的低资源情感语声合成方法。

2 低资源情感模型构建

本文的模型在Tacotron2模型以及Skerry-Ryan等[23]的情感编码器的基础上进行构建，提出了一种低资源的汉语儿童情感语声合成模型，利用迁移学习自适应的方法缩短模型训练时长，并且使得合成语声质量和情感维度得到保证。Tacotron模型的训练基本需要24 h以上的语料[1]，而汉语儿童情感语声在语声合成上的资源极其匮乏，因此本文将使用自行采集的儿童情感语料库进行实验。

2.1 数据采集与处理

由于本研究的合成不考虑多说话人，因此录制选用单人8岁女童在成人引导下的表演型情感语声。语录文本参考由Zhou等[26]研发的情感语声数据库(ESD)，因儿童的语言表达能力与成人差异较大，从中筛选出60句左右适合于儿童口语表达的句子。实验证明8岁儿童能够清晰地用不同情感表达筛选出的语句。录制的声频包含的4种情感分别为：愤怒、开心、伤心和惊讶。所有语声数据都是16 kHz的样本，并以16位保存。考虑实际应用领域中的无法做到纯净的录声环境，达到降低情感语声合成成本的目的，模拟日常生活中安静的录声环境，对声频进行简单的剪辑降噪筛选预处理之后，4种情感录制声频的总时长约为500 s。

语料库中的文本和声频信息将进一步预处理，如图2所示。首先对文本信息进行韵律标注，本研究所采用的韵律标注规则严格遵守标贝语料库所使用的韵律标注规则。汉语的韵律标注包括对文本进行分词处理以及韵律信息的添加。利用THULAC(THU Lexical Analyzer for Chinese)工具进行分词处理，再根据韵律层级标注规则进行标注，按照不同的停顿间隔，向文本中加入韵律信息。

由于汉语与英文合成的差异，需要将原有的汉语信息转化为拼音格式进行标注[27]。本研究使用pypinyin工具来实现汉字转拼音。pypinyin是一种基于hotoo/pinyin开发的可以用于汉字注音、排序和检索工具，是一个准确率较高且较成熟的汉字拼音转换工具[21]。

在通过上述操作对文本进行处理之后，将文本与声频一一对齐。然后对情感语料库中的声频进行预处理。对每一段声频信号进行分帧、加窗，再对每一帧做快速傅里叶变换(Fast Fourier transform FFT)处理[28]，把每一帧的结果沿另一个维度堆叠起来，得到二维信号图片形式。通过短时傅里叶变换(Short-time Fourier transform, STFT)提取出每一段情感声频的频谱(spectrogram)和梅尔频谱(Melspectrogram)特征。

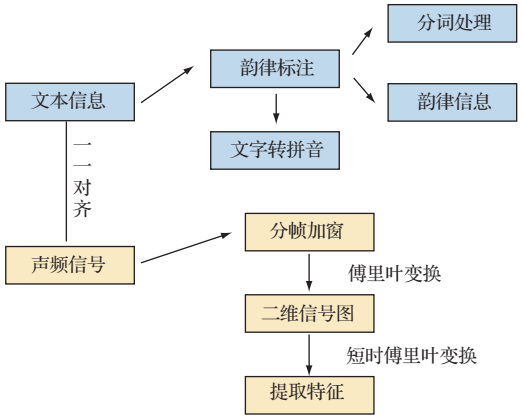


图 2　文本声频预处理

Fig. 2　Text and audio preprocessing

2.2　特征迁移

特征迁移的目的是减少在汉语儿童情感语声合成上其所需的数据量以及缩短模型的训练时间[29]。迁移学习方法是将已有的知识对于不同但是相关领域的问题进行一系列求解的新型机器学习方法[30]。文献[25]中研究者对利用低资源进行迁移学习实现情感语声合成模型的可能性进行探究，在DCTTS合成模型上具有较好的适应性。而针对本文所研究的汉语儿童情感语声合成，为解决汉语韵律复杂的问题[22]，采用其迁移学习的方法，在构建了多头注意力机制的Tacotron2模型上，加入情感编码器以及说话人编码器进行儿童情感特征的迁移。

而由于儿童的情感特征与成人相差较大，不同情感均有明显的变化。图3为在基频上成人与儿童不同情感上的差异(单位：Hz)。

儿童的其他韵律特征随着情感的不同变化也较大，因此在基于成人的情感语声合成上，对模型的情感编码器以及说话人编码器进一步进行自适应

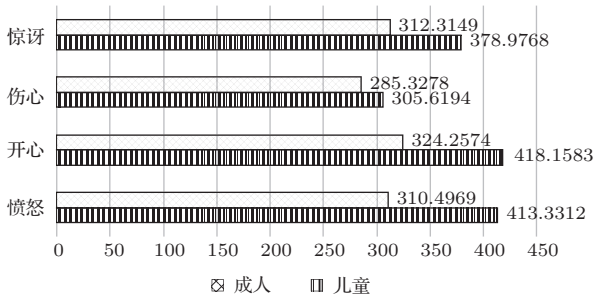


图 3　儿童与成人不同情感基频对比

Fig. 3　Comparison of different emotion fundamental frequencies between children and adults

的特征迁移。

类似于多说话人的情感语声合成，理想状态下，利用低维级的说话人嵌入向量可以实现低资源的语声合成，而在文献[31]中利用Tacotron模型进行多说话人语声合成时发现Tacotron模型的性能高度依赖模型的超参数，利用小数据集进行训练时无法进行注意力机制的学习。因此本文在不考虑多说话人语声合成的条件下，将ESD情感语料库中单人女性的4种情感语声中的说话人的相关参数存储在低维向量中，在行儿童特征迁移时，儿童的嵌入向量可以共享几乎全部的权重，解决了Tacotron模型的性能高度依赖模型的超参数的问题。成人情感语料选择成年女性作为说话者，其目的是接近儿童的韵律特征[3]。

如图4所示，进行儿童特征的迁移，首先将儿童情感语声进行数据预处理，分情感输入进不同的成人情感模型，然后通过情感编码器以及说话人编码器，实现情感特征与说话特征的迁移，再进行多头注意力机制学习，最终得到不同情感的汉语儿童语声合成模型。

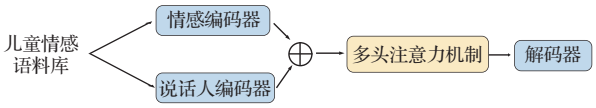


图 4　儿童特征迁移过程

Fig. 4　Transfer of children's characteristics

2.3　模型框架

本文构建的低资源儿童情感语声合成模型如图5所示。对文本进行预处理之后输入模型，最后以wav声频格式输出。

为解决汉语韵律复杂的问题，文本编码器利用标贝科技开源的汉语中性语料库进行预训练，该语料库时长约为12 h，从而降低耦合形成汉语语声合成模型对得到的模型权重等信息进行保存。在此基础上进行训练，降低其训练的时间成本。再利用ESD成人汉语情感语料库进行模型自适应的训练，使用单人4种情感语料时长约为1 h，在成人汉语情感语声合成模型的基础上进行特征迁移，如2.2节中所示，其儿童情感语声数据约为500 s，实验数据非常小，训练时长花费约为36 h时模型得到收敛，能合成出自然度以及情感度较好的语声。

模型的后处理网络基于Griffin-Lim算法[32]，在后处理网络中添加CBHG模块(其中CB表示1D

Convolutional Bank，H 表示 Highway network，G 表示 Bidirectional GRU)，能够通过正向传播和反向传播来修正每一帧的错误，以此来提高声频质量 [33]。

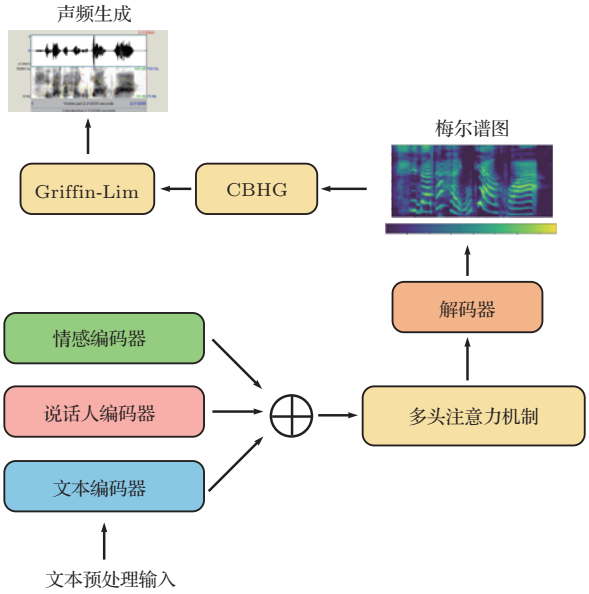


图 5　低资源儿童情感语声合成模型框架

Fig. 5　Children emotional speech synthesis model

3　实验结果与分析

本实验将从主观和客观两个维度对合成的语声进行评价。其中客观评价采用情感语声识别的不加权平均召回率 (Unweighted average recall rate, UAR)，对成人情感语声合成结果和儿童情感语声合成结果进行对比。主观听辨实验采用平均意见得分 (Mean opinion score, MOS) 和情感相似度平均意见得分 (Emotional mean opinion score, EMOS)，其目的是对合成语声从自然度和情感度两个维度上进行评价。

3.1　客观实验结果

本文的客观实验采用检测语声合成质量的梅尔倒谱失真 (Melcepstral distortion, MCD) [34] 方式以及机器学习的语声情感识别方法。

MCD 是一种用来检测语声质量的客观评价指标，衡量两个梅尔倒谱序列之间差异的量度，MCD 越小表示其合成的语声与原始语声越为相似，计算方式如下：

$$\mathrm{MCD}_K = \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sum_{k=1}^{K}\left(c_{(t,k)}-c'_{(t,k)}\right)^2}, \qquad (3)$$

其中，$c_{t,k}$、$c'_{t,k}$ 分别是来自原始和合成语声中的第 t 帧的第 k 个梅尔倒谱系数 (Mel frequency cepstral coefficient, MFCC)，其中跳过 $c_{t,0}$，因为 0 阶 MFCC 反映的是频谱能量。本实验对合成的 4 种情感分别进行 MCD 计算，其结果如表 1 所示。

表 1　合成语声不同情感的 MCD

Table 1　MCD of different emotions in synthetic speech

情感	MCD
愤怒	4.40
开心	4.59
伤心	4.36
惊讶	6.29
平均	4.91

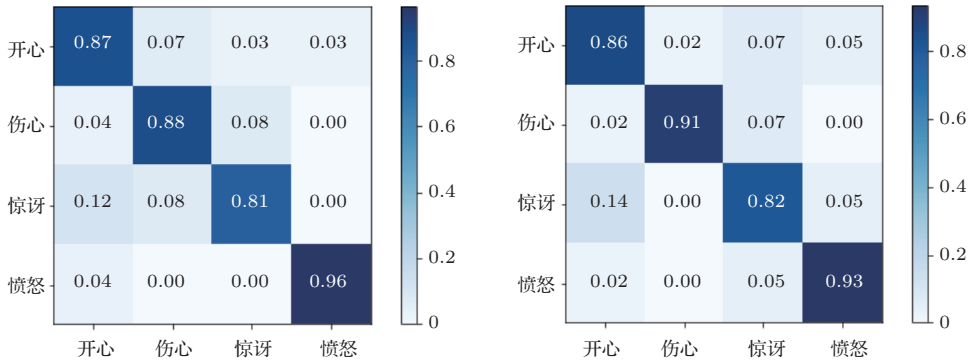
MCD 平均值为 4.91，根据文献 [35] 中的研究，当 MCD 值低于 8 时，合成的语声能被语声识别系统所识别，从而应用于语声交互中。因此本实验的语声合成质量较为优良，然而惊讶情感的合成质量较差的结果仍有待进一步探究。

对于情感语声的客观评价，在保证语声质量的条件下本实验对合成语声的情感质量使用语声情感识别的方式进行测评。由于合成的语声数据量在 100 句左右，实验采用支持向量机 (Support vector machine, SVM) 建立情感识别模型，适用于小样本的分类模型。

在语声情感识别过程中，不同的情感特征对于最终的识别效果存在重要影响。提取以及选择能够有效反映情感变化的语声特征是语声情感识别领域目前最重要的问题之一 [36]。本实验采用 eGeMAPS 声学特征集，eGeMAPS 虽仅有 88 维特征，但涵盖了多种韵律特征、基于谱的特征以及音质特征 [1]。

混淆矩阵的结果能详细表现出情感之间的分类情况及误判情况，它的横向表示实际的结果，纵向表示预测的结果，从左上到右下的对角线表示的是预测正确的值，其余是误分值，合成语声的混淆矩阵如图 6 所示。

通过对比模型合成的成人情感语声和儿童情感语声的混淆矩阵可以看出，其迁移效果较为良好。



(a) 成人　　(b) 儿童

图 6　合成情感语声混淆矩阵

Fig. 6　Synthetic emotion speech confusion matrix

3.2　主观实验结果

现如今大多数的情感语声合成测评都采用主观听辨实验来进行评价。本实验选择 MOS 和 EMOS 两种测评从自然度和情感度两个维度对合成的语声进行评价。

实验选取合成的儿童情感语声共 40 句 (每种情感 10 句)，要求 20 名听众对声频进行自然度和情感度的评价，评价分为 5 个等级，评价量表如表 2~3 所示。

表 2　MOS 评测分值标准表

Table 2　MOS evaluation score standard table

分值	评测标准
0~1	劣，极差，听不懂
1~2	差，勉强，听不太清楚
2~3	中，有延迟，可以接受
3~4	良，听得清楚，愿意接受
4~5	优，很自然

表 3　EMOS 评测分值标准表

Table 3　EMOS evaluation score standard table

分值	评测标准
0~1	劣，情感度不明
1~2	差，情感度模糊
2~3	中，情感度可以接受
3~4	良，情感度愿意接受
4~5	优，情感相似度理想

主观听辨实验的结果 (如表 4 所示) 给出了 95% 置信区间的 MOS 值，MOS 的平均值为 3.62，EMOS 的平均值为 4.17，其中愤怒和惊讶两种情感的自然度最低，儿童在表达这两种情感的时候语速通常较快，韵律特征上较为明显，时长会大幅度缩短，汉语的语声合成存在由于急促停顿而音节合成不完整的现象，这一现象在儿童语声合成上更为明显。相比较于文献 [25] 中研究，利用 (DCTTS 合成模型实现低资源情感语声合成，其各情感的平均 EMOS 值在 2.1~3.59 的范围内，本文的 EMOS 值比较可观能达到 4.17，说明对 Tacotron2 模型进行重新构建之后更适用于低资源的情感语声合成，并且在情感表达上，本实验提出的方法具有一定的可行性。

表 4　儿童合成情感语声 MOS/EMOS 评测值

Table 4　MOS/EMOS evaluation of children's synthetic emotional speech

情感	MOS	EMOS
愤怒	3.42 ± 0.13	4.1 ± 0.12
开心	3.805 ± 0.13	4.125 ± 0.13
伤心	3.745 ± 0.12	4.265 ± 0.11
惊讶	3.52 ± 0.13	4.2 ± 0.11
平均	3.62	4.17

4　结论

本文提出了一种基于迁移学习的低资源儿童汉语情感语声合成，其目的是解决儿童语料建模难度大、资源匮乏、训练时间长而导致合成模型训练效果不佳等问题。本文首先在 Tacotron2 模型的基础上实现汉语的语声合成，利用 ESD 成人汉语情感语料库实现情感语声模型，在此基础上使用小样本的儿童情感语声实现儿童特征低维迁移，在保证情

感表达度的条件下，合成出具有儿童特征的情感语声。实验结果证明本文使用的迁移学习方式可以有效得合成出自然度和情感表达度相对优良的语声。

而本文的研究仍存在一定的不足，如由于小样本录制环境以及样本量的关系，得到的实验结果会产生一定的过拟合现象。而本文的小样本的训练时长仍需 30 h 以上，相较于其他的低资源情感语声合成方法，在训练效率上仍有待提高。并且本文的语声合成仅限于单人，在多人情感语声合成的研究中，解决说话人情感特征的问题与本文的研究也存在相似之处，未来的工作中可以尝试使用相同的技术对多说话人的情感语声合成进行实验。而对于小样本的情感语声合成未来仍有很大的发展空间。

参 考 文 献

[1] Grigorev A, Frolova O, Lyakso E. Acoustic features of speech of typically developing children aged 5–16 years[C]. 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018.

[2] Shahnawazuddin S, Adiga N, Kathania H K. Effect of prosody modification on children's ASR[J]. IEEE Signal Processing Letters, 2017, 24(11): 1749–1751.

[3] House D, Bell L, Gustafson K, et al. Child-directed speech synthesis: evaluation of prosodic variation for an educational computer program[C]//European Conference on Speech Communication & Technology. DBLP, 1999.

[4] Inoue K, Hara S, Abe M, et al. An investigation to transplant emotional expressions in DNN-based TTS synthesis[C]//2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017.

[5] Yamagishi J, Onishi K, Masuko T, et al. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis[C]. IEICE Transactions on Information and Systems, 2005, 88(3): 502–509.

[6] Lorenzo-Trueba J, Barra-Chicotea R, San-Segundo R, et al. Emotion transplantation through adaptation in HMM-based speech synthesis[J]. Computer Speech & Language, 2015, 34(1): 292–307.

[7] Inoue K, Hara S, Abe M, et al. An investigation to transplant emotional expressions in DNN-based TTS synthesis[C]//2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017.

[8] Strömbergsson S, Edlund J, Götze J, et al. Approximating phonotactic input in children's linguistic environments from orthographic transcripts[C]//Interspeech 2017, 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden, 2017.

[9] Ling Z, Kang S, Zen G, et al. Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends[J]. IEEE Signal Processing Magazine, 2015, 32(3): 35–52.

[10] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//28th Conference on Neural Information Processing Systems(NIPS). Montreal, QC, Canada, 2014.

[11] Shan Y, Wu Z, Lei X. On the training of DNN-based average voice model for speech synthesis[C]//2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 2016.

[12] Zhang H, Lin Y. Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages[J]. arXiv Preprint, arXiv: 2008.04549.

[13] Li T, Yang S, Xue L, et al. Controllable emotion transfer for end-to-end speech synthesis[C]//2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2021.

[14] Li R, Wu Z, Huang Y, et al. Emphatic speech generation with conditioned input layer and bidirectional LSTMS for expressive speech synthesis[C]//ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[15] Xue L, Zhu X, An X, et al. A comparison of expressive speech synthesis approaches based on neural network[C]//Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data, 2018: 15–20.

[16] Jia Y, Johnson M, Macherey W, et al. Leveraging weakly supervised data to improve end-to-end speech-to-text translation[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 7180–7184.

[17] Inoue K, Hara S, Abe M. Module comparison of transformer-Tts for speaker adaptation based on fine-tuning[C]//2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2020: 826–830.

[18] Lee Y, Kim T. Robust and fine-grained prosody control of end-to-end speech synthesis[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5911–5915.

[19] Huybrechts G, Merritt T, Comini G, et al. Low-resource expressive text-to-speech using data augmentation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6593–6597.

[20] 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.

[21] 王国梁, 陈梦楠, 陈蕾. 一种基于 Tacotron 2 的端到端中文语音合成方案 [J]. 华东师范大学学报 (自然科学版), 2019(4):

111–119.
Wang Guoliang, Chen Mengnan, Chen Lei. An end-to-end Chinese speech synthesis scheme based on Tacotron 2[J]. Joural of East China Normal University(Natural Science), 2019(4): 111–119.

[22] 张亚强. 基于迁移学习和自学习情感表征的情感语音合成 [D]. 北京: 北京邮电大学, 2019

[23] Skerry-Ryan R J, Battenberg E, Ying X, et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron[C]//International Conference on Machine Learning. PMLR, 2018: 4693–4702.

[24] Cho K, Merrienboer B V, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. Computer Science, 2014, arXiv: 1406.1078.

[25] Tits N, Haddad K E, Dutoit T. Exploring transfer learning for low resource emotional TTS[C]//Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2019.

[26] Zhou K, Sisman B, Liu R, et al. Emotional voice conversion: theory, databases and ESD[J]. Speech Communication, 2022, 137: 1–18.

[27] 应雨婷. 基于循环神经网络的中文语音合成研究与应用 [D]. 南京: 东南大学, 2019.

[28] 曹欣怡. 基于韵律参数优化的情感语音合成 [D]. 南京: 南京师范大学, 2020.

[29] Pan S J, Qiang Y. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359.

[30] 庄福振, 罗平, 何清, 等. 迁移学习研究进展 [J]. 软件学报, 2015, 26(1): 26–39.
Zhuang Fuzhen, Luo Ping, He Qing, et al. Survey on transfer learning research[J]. Journal of Software, 2015, 26(1): 26–39.

[31] Gibiansky A, Arik S Ö, Diamos G F, et al. Deep voice 2: multi-speaker neural text-to-speech[C]//31th Conference on Neural Information Processing Systems. NIPS. Long Beach, 2017.

[32] 都格草, 才让卓玛, 南措吉, 等. 基于神经网络的藏语语音合成 [J]. 中文信息学报, 2019, 33(2): 75–80.
Dou Gecao, Cai Rangzhuoma, Nan Cuoji, et al. Neural network based tibetan speech synthesis[J]. Journal of Chinese Information Processing, 2019, 33(2): 75–80.

[33] Wu X, Cao Y, Wang M, et al. Rapid style adaptation using residual error embedding for expressive speech synthesis[C]//Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2018: 3072–3076.

[34] Kubichek R. Mel-cepstral distance measure for objective speech quality assessment[C]. In Communications, Computers and Signal Processing, 1993, IEEE Pacific Rim Conference on IEEE, 19–21 May, 1993.

[35] Yan C, Zhang G, Ji X, et al. The feasibility of injecting inaudible voice commands to voice assistants[J]. IEEE Transactions on Dependable and Secure Computing, 2019, 18(3): 1108–1124.

[36] 赵力, 黄程韦. 实用语音情感识别中的若干关键技术 [J]. 数据采集与处理, 2014, 29(2): 157–170.
Zhao Li, Huang Chengwei. Key technologies in practical speech emotion recognition[J]. Joural of Data Acqisition and Processing, 2014, 29(2): 157–170.