ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 4,Jul. 2022,pp. 909－916
DOI:10. 16337/j. 1004-9037. 2022. 04. 018

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于Tacotron模型和韵律修正的情感语音合成方法

张　昕，胡航烨，曹欣怡，王　蔚

（南京师范大学教育科学学院，南京 210097）

摘　要：语音合成技术日趋成熟，为了提高合成情感语音的质量，提出了一种端到端情感语音合成与韵律修正相结合的方法。在Tacotron模型合成的情感语音基础上，进行韵律参数的修改，提高合成系统的情感表达力。首先使用大型中性语料库训练Tacotron模型，再使用小型情感语料库训练，合成出具有情感的语音。然后采用Praat声学分析工具对语料库中的情感语音韵律特征进行分析并总结不同情感状态下的参数规律，最后借助该规律，对Tacotron合成的相应情感语音的基频、时长和能量进行修正，使情感表达更为精确。客观情感识别实验和主观评价的结果表明，该方法能够合成较为自然且表现力更加丰富的情感语音。
关键词：语音合成；端到端合成；韵律修正；情感语音
中图分类号：TP391　　**文献标志码：**A

Expressive Speech Synthesis Method Based on Tacotron Model and Prosodic Correction

ZHANG Xin, HU Hangye, CAO Xinyi, WANG Wei

(College of Education Science, Nanjing Normal University, Nanjing 210097, China)

Abstract: Speech synthesis technology is becoming more mature. In order to improve the quality of synthetic emotional speech, this study proposes a method combining end-to-end emotional speech synthesis with prosodic correction. Based on the Tacotron model, the prosodic parameters are modified to improve the emotion expression power of the synthetic system. Tacotron model is first trained with a large neutral corpus, and then a small emotional corpus is used to train and synthesize emotional speech. Then the Praat acoustic analysis tool is used to analyze the prosodic features of emotional speech in the corpus and summarize the parameters of different emotional states. Finally, with the help of this rule, the fundamental frequency, duration and energy of the corresponding emotional speech synthesized by Tacotron are modified to make the emotional expression more accurate. The results of objective emotion recognition experiment and subjective evaluation show that this method can synthesize more natural and expressive emotional speech.
Key words: speech synthesis; end-to-end synthesis; prosodic correction; emotional speech

引　　言

情感语音的合成逐步成为语音处理技术的热点方向，合成自然度高而且包含丰富情感信息的语音

基金项目：国家哲学社会科学基金（BCA150054）。
收稿日期：2021-07-23；**修订日期：**2021-10-27

对实现更自然的人机交互有着重要意义。传统的合成方法主要有以下3种:波形拼接法、韵律特征修改法以及基于隐马尔可夫模型(Hidden Markov model, HMM)的合成法。波形拼接的合成方法需要从大量的情感数据库中搜寻满足目标情感的音频片段,并根据一定的序列进行衔接,然而其语料库成本过高且语音片段拼接点处过于生硬,难以合成语料库之外的声音。韵律特征修改能有效改善合成语音情感缺乏的问题,却是以降低音频质量为代价。基于隐马尔可夫模型(Hidden Markov model,HMM)的方法,受人为干扰的影响较小,但由于其生成的是均值矢量参数序列,合成的声音过于平滑,无法有效表达需要的情感。深度学习算法的快速发展使语音合成领域的研究者们看到了希望,各类神经网络在语音合成中应用无需决策树聚类,便可从语言特征到声学特征转换的过程中学习到直接、分层和非线性的模型[1],快速提升合成语音的质量。如WaveNet[2]是基于PixelCNN架构、在不增加计算成本的条件下使用带洞卷积直接生成语音波形的一种深度学习合成模型,其合成质量都优于传统方法,但计算量大仍然是其主要缺点,而且该模型未进行前端文本的改进处理。Char2Wav整合了前端和后端,由神经声码器和读取器组成,直接从文本生成语音,但它使用的仍然是SampleRNN神经声码器之前的预测声码器参数[3]。Tacotron是一种基于注意力机制的典型端到端合成模型,它属于帧级模型,不需要在音素级别进行对齐操作,根据〈文本,音频〉对,采取随机初始化方式从零训练,方便在多种声学数据中泛化和扩展[4]。Fastspeech通过概率密度蒸馏等方法并行生成中间表征,相比较于自回归的声学模型其合成速度有了明显提升,但实现模型结构中的Pipeline比较复杂[5]。

Tacotron等经典模型由于并未清晰地实现韵律建模,合成的语音相比真实人声显得生硬呆板[6],因此研究者们一直致力于如何实现表现力更丰富的情感语音合成。合成情感语音一般有以下两种方式:

(1)先合成出中立的语音,再根据不同情感状态下的声学特征规律对中立语音进行修改,最后得到情感语音。如何凌等建立了高兴、生气、悲伤和无聊4种情感的韵律特征模板,借助时域基因同步叠加算法(Time domain pitch synchronous overlap add, TD-PSOLA)算法对中性状态的语音参数进行调整,其合成出的情感语音得到了较高的正确判别率[7]。陈洁等基于HMM可训练合成方法合成中立语音,并通过分析平静、高兴、悲伤和生气4种情感的韵律特征变化规律,用Praat软件对中立语音的特征参数进行修改,最终合成情感语音[8]。Wang等提出了一种多级韵律转换的方法,从句子、音节和韵律词3个层次对基频F_0、短时能量和语速进行修改,将中性语音转换为情感语音[9]。这种方法虽然情感表达准确,但多为人工干涉,在语音的自然度上容易有所欠缺。

(2)通过对情感数据库中的音频进行训练,直接生成目标情感语音。如Lee等在韩语情感语料库上训练Tacotron模型,能成功地为给定情感标签生成语音[10]。这种方法在语音的自然度方面具有良好的表现,但是在部分语音的情感度上,其表达的情感并不够精确,与目标情感存在一定的差异。因此一些研究者提出了对合成出的情感语音进行韵律特征修改的方法,如陈明义等建立了高兴、悲伤、中立以及愤怒4种不同情感的韵母基音模板库,从中挑选符合目标情绪的语音片段,运用基音同步叠加算法合成波形,并修改合成语音的韵律参数,得到了更理想的情感声音[11]。Zhang等采用深度神经网络(Deep neural networks,DNN)预测目标情感语音声学参数,对愉悦度、激活度、优势度(Pleasure-arousal-dominance)三维情感空间模型坐标值进行聚类并根据方差值计算训练结果与参考值之间的距离,按照不同情感权重的高低对相应韵律参数进行调整,主观印象分(Mean opinion score,MOS)结果表明,该方法的语音合成效果优于传统的DNN模型和HMM模型[12]。

考虑到Tacotron模型能够简单高效地合成出较为自然的语音但情感表达度又不够,本文在该模型的基础上,进行端到端的情感语音合成并对合成的情感语音进行韵律特征的调整,从而合成出情感表达自然且丰富的语音。

1　基于韵律修改的端到端情感语音合成

1.1　数据集

由于缺乏用于合成的情感语料,模型在进行训练时容易造成过拟合现象。该研究对LJ平静状态语音数据集进行训练,保留中性语音模型,并在训练出的权重基础上对情感语料库中的训练模型进行调整。

(1)LJ中性语料库[13]

LJ中性语料库由一位说话者朗读生成13 100个中立语音,该中性语料库含有与音频配对的文本以及同名的ID序列,存储在Metadata数据文件当中。

(2)Emotional Voices数据集

Emotional Voices数据集是由Adaeze等[13]结合2个中立数据集(CMU-Arctic英语数据集和SIWIS法语数据集)转录得到的情感语料库。该语料库包含2位男性英国说话人、2位女性英国说话人和1位法国男性说话人。本文研究选取了其中一位英国女性的语音,包括中性、气愤、疲倦、憎恶和逗乐5种情感,其中逗乐情感的语音包含笑声。

1.2　基于Tacotron模型和韵律特征修改的情感语音合成框架

端到端语音合成模型将传统语音合成系统中的3大模块集成封装于一体,既可避免不同语言学背景人员造成的文本标注差异,也降低了语音合成研究者在发声机理方面的门槛,它能够根据输入的文本信息直接合成出需要的目标音频。本文研究选取了当前广泛使用的端到端语音合成模型Tacotron,并在其基础上增加韵律调整技术,构建了端到端和韵律修正相结合的情感语音合成框架。基于Tacotron和韵律转换的情感语音合成方法框架如图1所示。本框架可分为3个模块:Tacotron训练模块;特

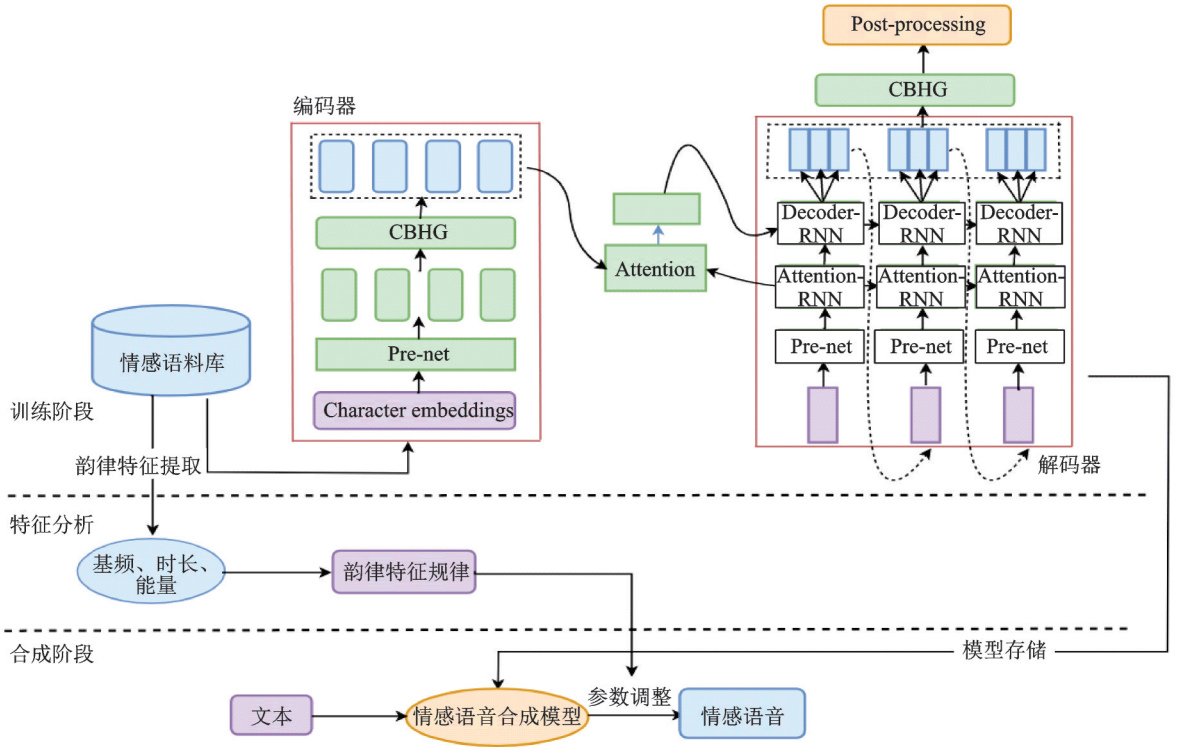


图1　基于Tacotron模型和韵律特征修改的情感语音合成框架

Fig.1　Framework for affective speech synthesis based on Tacotron model and prosodic feature modification

征分析模块和情感语音合成模块。图1中CBHG为由一维卷积滤波器、高速公路网络、双向门控递归单元组合成的模块(1-D convolution bank + highway network + bidirectional GRU model,CBHG)。

1.2.1 模型训练

在初期阶段,对待训练的文本和语音进行相应配对,实施流程化的预处理操作,分气愤、逗乐、憎恶和疲倦多种不同情感输入Tacotron模型,它可以利用深度神经网络推理模型,根据嵌入的不同情感进行迁移学习和自适应学习,训练出目标情感模型,最后将训练好的相应情感语音合成模型各自保存以便后续合成目标情感语音。

Tacotron主要由3部分组成,分别为编码器、带有注意力机制的解码器以及后处理网络[4]。编码器能够从文本中提取出稳健序列,通过Pre-net预处理结构对字符向量进行非线性操作,其中的Dropout层协助模型更快地收敛和泛化;随后连接一个CBHG模块,对所有卷积层进行批量归一化并输入到Highway net部分进行高级特征提取,再借助双向门控循环单元(Gated recurrent unit, GRU)提取出上下文序列信息。解码器采用基于内容的注意力机制将自身注意力机制循环神经网络(Attention-recurrent neural network, Attention-RNN)的输出和编码器传送到上下文矢量连接,并作为Decoder-RNN的输入,其输出又与解码器的初始帧输入结合,继而生成Mel谱帧;它包含有与编码器相同结构的Pre-net预处理网络,并采用垂直残差连接的GRU加快收敛速度。解码器没有直接将输出转化为音频,而是用不同于编码器参数的CBHG模块作为后处理结构,它可以透析完整的解码序列,提取序列特征,同时通过双向传播来更正各个帧所出现的不匹配问题,然后用Griffin-Lim算法将后处理的输出合成为语音。

1.2.2 特征分析

声学特征主要包含3大类:广泛研究的韵律特征、基于线性谱或倒谱的谱特征和声音质量特征。基频、时长和能量是研究者们关注分析的主要特征。基频指基音振动的频率,能够反映说话人的音色与腔调,决定情感语音语调的高低。激活度高的情感音调偏高,变化幅度较大;激活度低的情感则基频值相对较低,变化幅度较小。时长反映的是语速的快慢,时长与语速成反比,人们处于激活度高的情绪状态时,想要表达出来的感受更为迫切,因而一般语速较快,时长较短。声音的强度可以通过短时能量来表示,能量变化幅度越大,情感激活度越高。

本文研究利用Praat声学软件对语料库中情感语音的韵律特征进行提取及分析,通过默认的自相关方法获取不同类别情感的基频并分析其曲线变化、借助持续时长和平均发音速率分析不同情感差异、比较短时能量的均值、最大值和最小值等振幅参数值并分析不同情感的能量曲线,进而归纳出各种情感色彩韵律参数的变换规律,具体分析将在第2节中详细阐述。

1.2.3 情感语音合成及优化

根据待合成文本的上下文信息,调用训练阶段所合成的端到端不同情感模型生成目标情感语音,根据分析阶段所获得的不同情感语音的韵律参数规律,在保证整体基频曲线不变、原始合成音频速率和参照音频速率之间比例关系确定以及振动幅度比例关系确定的情况下,成倍数关系对目标情感语音进行参数调整和韵律修改,合成表现力更丰富的情感语音,提高所合成情感语音的准确性,具体调整方法将在第2节中详细阐述。

2 韵律特征参数的分析与修正

情感语音合成所研究的声学参数以基频、持续时间和振幅能量这几种韵律学参数为主。声音频率的高低变化即基频能够对不同情绪状态下的声调变换进行恰当表示,是语音韵律研究的重要指数。表达者说话速度的快慢通过时长反映,积极情境与消极状态下的时长消耗有着不同的表现。声音强度的高低通过能量进行反映,如在兴奋、激动和气愤的情况下,由于难以控制情绪,人们通常会大声说话,音量不自觉升高;而在难过、忧愁和沮丧等情绪下,声音强度相对较低。因此,该研究对这3个特征的参数进行分析与修改。

2.1　基频

在自然发音中,基频(Pitch)决定着说话者的腔调以及音色变化,因此基频对于语音研究具有重要意义。该研究使用Praat声学分析软件作为提取各种情感语音基频的工具。通过分析比较多条语句的基频参数,发现在不同情感状态下其基频参数具有相似的规律:气愤情感基频均值最高,波动幅值较次于开心情感;逗乐情感基均频值次之,但其波动幅值较高;气愤和高兴两种情感的基频变化起伏较多,另外3种情感则较少,中性情感的整体基因频率低于憎恶和疲倦两种情感,三者的基频均值从高到低依次为憎恶、疲倦和中性。表1为其中一句语音"Her own betrayal of herself was like tonic to Philip."在不同情感下基频的均值和2个极值。

表1　5种情感语音的基频值

Table 1　Fundamental frequency values of five kinds of emotional speeches　Hz

情感类型	基频均值	基频最大值	基频最小值
逗乐	329.96	500.25	133.34
疲倦	244.71	501.22	109.51
憎恶	271.82	383.02	96.55
气愤	359.12	515.18	220.87
中性	204.56	251.54	78.55

本文研究在修改基频时,主要对基频均值、基频变化区间等进行相应调整。修改基频均值时,根据合成语音的基频值与基准值之间的比值(假设为f)进行相应调整,如合成的"气愤"情感基频不够高,就将各处基频点上的值均提升到原来的f倍;反之,则减小到原来的$1/f$。在进行基频修整的时候,其整体的区间应以同样的比例进行变换,因此2个基频极值应按照对应比例缩放,同时保持原波形走向不变。

2.2　时长

时长(Duration)指说话人表述完一个完整语句所用的时间,同时表明说话速度的快慢。分析比较多条语句在不同情感状态下的时长和语速差异,可以明显地发现:憎恶情感所需时间最长,语速最慢;疲倦情感和中性情感激活程度较低,在时长和平均速率上基本没有突出变化;气愤与逗乐语音速率很快,时长偏短,尤其是气愤语音,其活跃度很高。表2为其中一句语音"Her own betrayal of herself was like tonic to Philip."在不同情感状态下所需时长和速率值,二者呈明显的反比例关系。

表2　5种情感语音的时长与平均速率

Table 2　Duration and average rate of five kinds of emotional speeches

情感类型	时长/s	平均发音速率/(音节·s^{-1})
逗乐	3.63	0.275
疲倦	5.16	0.194
憎恶	5.62	0.178
气愤	2.55	0.391
中性	4.48	0.223

根据以上分析,时长参数的修正可借助语速的调整来优化。在修改的过程中应该保持其他语音参数不变,按照合成语音的时长与基准时长之间的比例(假设为t)改变音频的速度。例如要减少时长时,可以将语速加快t倍;要增加时长时,将发音速率减慢到原来的$1/t$。

2.3　能量

能量(Energy)表示音强,即音频的强烈程度,通常采用短时能量和短时平均幅度来表示,人在不同情感情境中会有不同的发音强度。活跃状态的发音要强于平静状态,沮丧低沉时能量较弱。表3为不同情感语音"Her own betrayal of herself was like tonic to Philip."的能量值。

表3　5种情感语音的能量值

Table 3　Energy of five kinds of emotional speeches　dB

情感类型	能量均值	能量最大值	能量最小值
逗乐	66.76	86.97	36.22
疲倦	66.47	87.23	49.87
憎恶	59.34	86.21	31.11
气愤	69.04	87.97	50.13
中性	62.02	86.88	40.93

在语音调整时，优化音频信号的摆动幅度即可达到能量修正效果。如情绪较为激动时，其音强较高，可根据合成幅值与基准幅值比，将能量系数(假设为k)扩大2倍、3倍或者更大；情绪较为低落时，能量值较低，k值缩小。另外，应该使语音信号曲线保持中间幅值高，两端幅值低的走势，遵循人体本身的发音规律。

3 实验结果与分析

为了验证合成语音的效果，该研究进行了情感识别实验以及主观听辨实验。将未加入韵律修正的端到端情感语音合成方法合成的200句情感语音和加入韵律调整的合成情感语音进行比较。

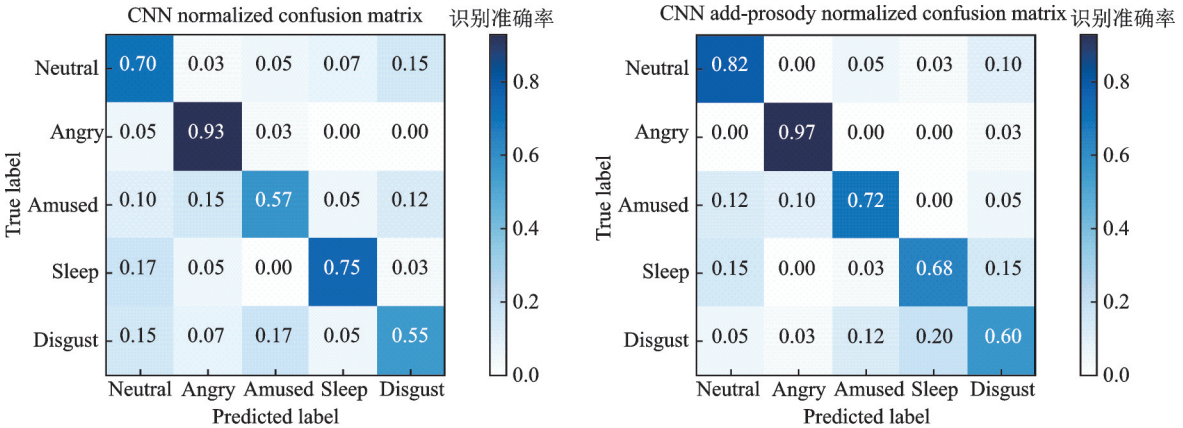
3.1 情感识别实验

借鉴已有的语音情感识别相关方法[14]，借助性能良好的卷积神经网络(Convolutional neural networks，CNN)分类器，根据eGeMAPS特征集提取特征，对5种情感进行判别。GeMAPs特征集对包括频谱特征、振幅特征、平衡参数在内的18个低水平特征进行算术均值和标准离差率计算，在音高和响度的浊音区进行其他统计操作，再加入时间参数和4个清音区特征，形成62维特征。eGeMAPS在此基础上添加7个低水平倒谱参数并对其所有区域实施算数平均和变异系数处理，在共振峰带宽、频谱流量、梅尔频率倒谱系数1～4的浊音区和频谱流量的清音区应用统计函数，再加入等效声级共26个特征参数，总共得到88维特征[15]。它涵盖多种基础声学特征，并增添了倒谱参数和更多的动态信息，在语音情感识别任务中具有较高的鲁棒性。分类结果依据不加权平均召回率(Unweighted average recall，UAR)[16]进行评比，其计算方法如式(1)所示。情感识别模型对未加入韵律修正的合成情感语音的识别率为0.7，而对加入韵律调整的情感语音的识别率为0.76，这表明对情感语音进行韵律优化确实能提升合成效果。

$$P_{\mathrm{UAR}}=\frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{n_i} \tag{1}$$

式中：N为所有情感类别；c_i为第i种情感识别准确的样本数；n_i为第i种情感总样本数。

两种方法的情感识别混淆矩阵如图2所示。从图2中可以看出，对韵律特征进行修改后，中立、气愤、逗乐和憎恶情感的分类准确度都得到了提升，尤其是降低了中性和憎恶情感、气愤和逗乐情感的混淆程度；但是疲倦和憎恶情感的易混率反而更高了，这与两种情感的特征相似性有关。在下一步的研究中需要采用更为精准的特征分析方法来区分这两种情感。



(a) Synthetic speech emotion recognition confusion matrix without prosodic feature modification

(b) Synthetic speech emotion recognition confusion matrix with prosodic feature modification

图2 两种方法情感识别混淆矩阵

Fig.2 Confusion matrices of emotion recognition in two cases

3.2　主观听辨实验

主观听辨实验在中性、逗乐、气愤、疲倦和憎恶这5种情感中随机各选取10句语音,每句语音片段从3～5 s不等,共50个样本,用MOS和AB偏好测验分别来评价所合成情感语音的自然度和情感表现力。

分别选取5名男性和5名女性实验者进行5级MOS自然性评测,1表示很不自然,2表示较不自然,3表示一般,4表示较自然,5表示很自然,平均意见结果分为3.78。如表4所示,Wang等的Tacotron模型MOS得分为3.82[4],Zen等基于长短期记忆网络的情感语音MOS得分为3.723[17]。这表明端到端合成出的情感语音进行韵律调整后会一定程度降低语音质量,但其效果仍然高于非端到端的语音合成方法,自然度仍然处于人耳可接受范围。

表4　不同方式合成的情感语音MOS评分

Table 4　MOS score of emotional speech synthesized in different ways

方法	MOS
Tacotron	3.82
Tacotron＋韵律修正	3.78
长短期记忆网络	3.723

将未进行韵律特征修改的情感语音(对照组)与修改后的情感语音(实验组)拼接在一起,同样选取5名男性和5名女性实验者进行AB偏好测试,让其在"前者更有情感、二者情感相同、后者更有情感"3个选项中进行判断,偏好结果如图3所示。在气愤和逗乐这两类情感的实验结果中,实验组的偏好占比明显超过对照组,分别为0.44和0.47,而对于疲倦和憎恶这两类情感,对照组的偏好反而更强一些,分别比实验组高出0.05和0.03,这表明韵律修改的方法能够在一定程度上增强端到端合成出的逗乐和气愤两种情感的表现力度,但在对疲倦和憎恶这两种情感的表现力上却引发了反作用,在情感识别实验中,这两种情感的易混度也有所增高,说明这两种情感的有效表达需要更深层次的分析与研究。

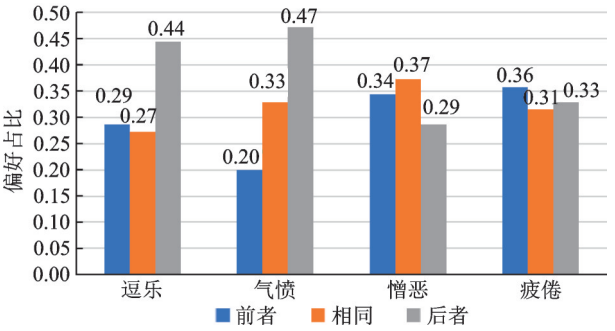


图3　韵律修正前后4种情感语音的AB偏好测试结果

Fig.3　AB preference test results of four emotion categories before and after prosodic modification

4　结束语

本文研究采用Tacotron模型进行端到端情感语音合成,并依据生成语音与基准语音之间的韵律参数比,对目标语音的基频、时长和能量等韵律特征进行更正,客观情感甄别实验和主观检验均证明该方法对逗乐情感和气愤情感合成较为有效,对疲倦与憎恶情感的调整有待进一步研究。MOS评分显示合成语音的自然度有所下降,这可能是受到人为韵律修改的影响,但依旧在可接受的范围之内。整体而言,情感分类效果提升了0.6,情感表达的准确度在一定程度上有所提高。

鉴于疲倦与憎恶情感混淆程度较高,未来的工作将会改善韵律特征的分析方法,还应进一步研究除韵律特征外的其他声学特征对于情感表达的影响。另外,本文采用5种离散情感进行实验,范围有限,下一步研究将考虑更多其他种类的情感或者从维度和连续情感的角度进行挖掘,期望能够在保证语音足够自然的前提下使其情感表现力有更明显的提升,表达的情感更加丰富准确。

参考文献:

[1]　YANG Hongwu, ZHANG Weizhao, ZHI Pengpeng. A DNN-based emotional speech synthesis by speaker adaptation[C]// Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). [S.l.]: IEEE, 2018: 633-637.

[2]　OORD A V D, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[EB/OL]. (2016-09-12)[2021-05-

26]. https://arxiv.org/abs/1609.03499v2.

[3] SOTELO J, MEHRI S, KUMAR K, et al. Char2Wav: End-to-end speech synthesis[C]//Proceedings of International Conference on Learning Representations (ICLR). [S.l.]:[s. n.], 2017.

[4] WANG Y, SKERRYRYAN R J, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[EB/OL]. (2017-03-29)[2021-05-28]. https://arxiv.org/abs/1703.10135.

[5] REN Y, RUAN Y, TAN X, et al. Fastspeech: Fast, robust and controllable text to speech[EB/OL]. (2019-05-22)[2021-05-28]. https://arxiv.org/abs/1905.09263.

[6] 潘孝勤,芦天亮,杜彦辉,等.基于深度学习的语音合成与转换技术综述[EB/OL].[2021-05-20].http://kns.cnki.net/kcms/detail/50.1075.TP.20210421.1321.024.html.
PAN Xiaoqin, LU Tianliang, DU Yanhui, et al. Overview of speech synthesis and voice conversion technology based on deep learning[EB/OL].[2021-05-20].http://kns.cnki.net/kcms/detail/50.1075.TP.20210421.1321.024.html.

[7] 何凌,黄华,刘肖珩.基于韵律特征参数的情感语音合成算法研究[J].计算机工程与设计, 2013, 34(7): 2566-2569.
HE Ling, HUANG Hua, LIU Xiaoheng. Synthesis of emotional speech based on prosody parameters[J]. Computer Engineering and Design, 2013, 34(7): 2566-2569.

[8] 陈洁, 张雪英, 孙颖.基于HMM的可训练情感语音合成研究[J].电声技术, 2012, 36(3): 43-46.
CHEN Jie, ZHANG Xueying, SUN Ying. Study for HMM-based trainable emotional speech synthesis[J]. Audio Engineering, 2012, 36(3): 43-46.

[9] WANG Z, YU Y. Multi-level prosody and spectrum conversion for emotional speech synthesis[C]//Proceedings of International Conference on Signal Processing (ICSP 2014). [S.l.]:[s. n.], 2014: 588-593.

[10] LEE Y, KIM T. Robust and fine-grained prosody control of end-to-end speech synthesis[C]//Proceedings of ICASSP. [S.l.]: IEEE, 2019: 5911-5915.

[11] 陈明义, 党培霞.基于情感基音模板的情感语音合成[J].中南大学学报(自然科学版), 2010, 41(6): 2258-2263.
CHEN Mingyi, DANG Peixia. Synthesis of emotional speech based on emotional pitch template[J]. Journal of Central South University (Science and Technology), 2010, 41(6): 2258-2263.

[12] ZHANG Weizhao, YANG Hongwu, ZHI Pengpeng. Emotional speech synthesis based on DNN and PAD emotional state model[C]//Proceedings of 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). [S.l.]: IEEE, 2018: 41-45.

[13] ADIGWE A, TITS N, HADDAD K E, et al. The emotional voices database: Towards controlling the emotion dimension in voice generation systems[EB/OL]. (2018-06-25). https://arxiv.org/abs/1806.09514.

[14] 王蔚,胡婷婷,冯亚琴.基于深度学习的自然与表演语音情感识别[J].南京大学学报(自然科学), 2019, 55(4): 660-666.
WANG Wei, HU Tingting, FENG Yaqin. Speech emotion recognition in nature and scripted state based on deep learning[J]. Journal of Nanjing University(Natural Science), 2019, 55(4): 660-666.

[15] EYBEN F, SCHERER K R, TRUONG K P, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2016, 7(2): 190-202.

[16] HERACLEOUS P, YONEYAMA A. A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme[J]. PLoS ONE, 2019, 14(8): 1-20.

[17] ZEN H, SAK H. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis[C]//Proceedings of ICASSP. [S.l.]: IEEE, 2015: 4470-4474.

作者简介:

张昕(1998-),女,硕士研究生,研究方向:情感语音合成。

胡航烨(1996-),女,硕士研究生,研究方向:情感语音合成。

曹欣怡(1996-),女,硕士研究生,研究方向:情感语音合成。

王蔚(1966-),**通信作者**,女,博士,教授,研究方向:人机交互、情感计算,E-mail: wangwei5@njnu.edu.cn。

(编辑:陈珺)