DOI:10.3979/j.issn.1673-825X.2018.05.013

利用说话人自适应实现基于 DNN 的情感语音合成

智鹏鹏,杨鸿武,宋　南
(西北师范大学 物理与电子工程学院,兰州 730070)

摘　要:为了提高情感语音合成的质量,提出一种采用多个说话人的情感训练语料,利用说话人自适应实现基于深度神经网络的情感语音合成方法。该方法应用文本分析获得语音对应的文本上下文相关标注,并采用 WORLD 声码器提取情感语音的声学特征;采用文本的上下文相关标注和语音的声学特征训练获得与说话人无关的深度神经网络平均音模型,用目标说话人的目标情感的训练语音和说话人自适应变换获得与目标情感的说话人相关的深度神经网络模型,利用该模型合成目标情感语音。主观评测表明,与传统的基于隐马尔科夫模型的方法比较,该方法合成的情感语音的主观评分更高。客观实验表明,合成的情感语音频谱更接近原始语音。所以,该方法能够提高合成情感语音的自然度和情感度。

关键词:情感语音合成;深度神经网络;说话人自适应训练;WORLD 声码器;隐马尔可夫模型
中图分类号:TN912.33　　**文献标志码**:A　　**文章编号**:1673-825X(2018)05-0673-07

DNN-based emotional speech synthesis by speaker adaptation

ZHI Pengpeng, YANG Hongwu, SONG Nan
(College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, P.R.China)

Abstract: The paper proposed a deep neural network (DNN)-based emotional speech synthesis to improve the quality of synthesized emotional speech by speaker adaptation with a multi-speaker and multi-emotion speech corpus. Firstly, a text analyzer was employed to obtain the context-dependent labels from sentences while the WORLD vocoder was used to extract the acoustic features from corresponding speeches. Then a set of speaker-independent DNN average voice models were trained with the context-dependent labels and acoustic features. Finally, the speaker adaptation was adopted to train a set of speaker-dependent DNN voice models of target emotion with target emotional training speeches. The target emotional speech was synthesized by the speaker-dependent DNN voice models. Subjective evaluations show that comparing with the traditional hidden Markov model (HMM)-based method, the proposed method can achieve higher opinion scores. Objective tests demonstrate that the spectrum of the emotional speech synthesized by the proposed method is also closer to the original speech than that of the emotional speech synthesized by the HMM-based method. Therefore, the proposed method can improve the emotion expression and the naturalness of synthesized emotional speech.

Keywords: emotional speech synthesis; deep neural network; speak adaptive training; WORLD vocoder; hidden Markov model

收稿日期:2018-01-27　**修订日期**:2018-09-13　**通讯作者**:杨鸿武　yanghw@nwnu.edu.cn
基金项目:国家自然科学基金(11664036, 61263036);甘肃省高等学校科技创新团队项目(2017C-03)
Foundation Items: The National Natural Science Foundation of China(11664036, 61263036); The High School Science and Technology Innovation Team Project of Gansu(2017C-03)

0 引 言

目前有多种方法可以有效合成情感语音,包括波形拼接合成[1],韵律特征修改[2-3]和基于隐马尔可夫模型(hidden Markov model, HMM)的统计参数语音合成[4-5]3 种。波形拼接合成方法中,需要为情感语音合成系统建立一个大型情感语料库,通过增加大量情感样本来提高情感语音合成质量。韵律特征修改方法通过韵律特征分析和参数修改实现情感语音合成。基于 HMM 的统计参数情感语音合成方法通过使用小型情感语料库进行说话人自适应训练来获得情感声学模型,或者将情感模型应用于 HMM,通过情感修正合成情感语音。以上情感语音合成方法,波形拼接方法具有不易获取大型情感语料库的缺点;韵律特征修改的情感语音合成方法合成的情感语音自然度和情感表达度差;基于 HMM 的统计参数语音合成综合了语音合成方法的多种优点,在过去十年成为最受关注的方法之一,然而同样具有合成语音的自然度较低、韵律特征平淡等缺点。

近年来随着深度学习研究的广泛深入,深度神经网路(deep neural network, DNN)在语音识别上的成功应用促使许多学者将其应用于语音合成领域[6],其中,文献[7]提出了基于深度学习的统计参数语音合成,解决了 HMM 决策树上下文状态聚类的方法在建模复杂上下文相关依赖上的局限性,文献[8-13]通过不同方面对 DNN 语音合成方法进行了探究。虽然基于 DNN 的统计参数语音合成方法取得了较好的效果,然而目前关于统计参数语音合成的研究主要依赖于说话者,需要来自单个说话人的大量数据来建立稳定的声学模型,有时训练数据的质量对合成语音的自然性有很大影响。为了解决以上问题,在 DNN 方法的基础上,文献[14-17]提出了基于 DNN 的多说话人自适应方法,实现了自适应方法的文语转换。

在情感语音合成的研究上,文献[18]采用端到端韵律转换方法合成了不同文本的语音,文献[19]采用端到端方法实现了语音情感动画的合成,文献[20-21]分别采用 DNN 和循环神经网络长短期记忆(recurrent neural network-long short term memory notwork,RNN-LSTM)方法实现了情感语音合成。以上深度学习情感语音合成方法虽然能够合成出较为自然的情感语音,但在使用多个说话人的训练语料或较小的情感语料库进行训练时,合成的语音仍然无法满足情感要求。

为了进一步提升合成的情感语音的音质和情感表达,本文将说话人自适应应用于基于 DNN 的语音合成中,首先利用多个说话人的多种情感的语料库,训练得到多个说话人的无关平均音模型(average voice models, AVM)。在此基础上,采用目标说话人的目标情感的训练语句,利用说话人自适应技术,获得说话人相关的语音模型,用该模型合成目标说话人的目标情感语音。主观评测结果表明,本文方法合成的情感语音,其情感意见得分高于基于 HMM 的情感语音合成方法合成的情感语音。客观评测结果表明,本文方法合成的情感语音的频谱更接近于原始情感语音。因此,本文方法能够提高合成语音的自然度和情感表达。

1 总体框架

本文提出的情感语音合成框架如图 1 所示,分为数据准备、平均音模型训练、说话人自适应和情感语音合成 4 个阶段。

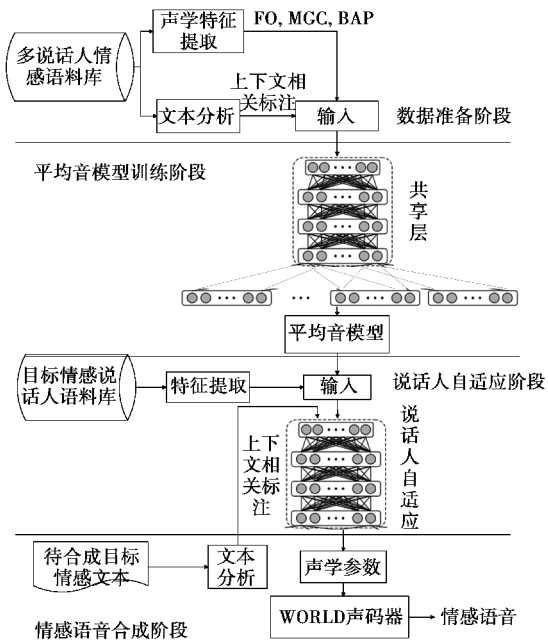

图 1 采用说话人自适应的基于 DNN 的情感语音合成框架

Fig.1 Framework of DNN-based emotional speech synthesis using speaker adaptation

1.1 数据准备

数据准备阶段,首先给出多个说话人的多种情感的情感语料库,利用 WORLD 声码器[22],从情感语音中提取模型训练所需的声学特征,包括基频(fundamental frequency, F0)、广义梅尔倒谱系数(Mel-generalized Cepstral, MGC),频带非周期分量

(band a periodical, BAP)。

本文以普通话声韵母作为语音合成的基本单元。语音对应的文本经过文本分析过程,借助于词典和语法规则库,通过文本规范化、语法分析、韵律分析、字音转换等,获得输入文本的声韵母信息、韵律结构信息、词信息和语句信息,从而获得情感语音对应文本的声韵母及其语境信息,最终得到声韵母的上下文相关标注,包括声韵母层、音节层、词层、韵律词层、短语层、语句层 6 层上下文相关标注。

1.2　平均音模型训练

在平均音声学模型的训练阶段,根据情感语音的上下文相关标注获得语言特征(二进制和数字)作为输入,声学特征包括 MGC,BAP,F0 和浊音/清音(voice/unvoiced,V/UV)作为输出,训练 DNN 模型。在训练期间,DNN 模型在不同的情感说话人之间共享各种隐藏层以建模其语言参数,经过反向传播(backpropagation, BP)算法的随机梯度下降过程(stochastic gradient descent, SGD)[23]进行时长与声学特征建模,最终训练出多个说话人的时长与声学特征的说话人无关 AVM。

在 AVM 模型的训练过程中,DNN 模型采用非线性函数对语言特征和声学特征之间的非线性关系进行建模,每个隐藏层 k 使用前一层的输出 $\boldsymbol{h}^k-1$ 计算输出向量 $\boldsymbol{h}^k$,$x=\boldsymbol{h}^0$ 是模型的输入。定义 $\boldsymbol{h}^k$ 为

$$\boldsymbol{h}^k = \tanh(\boldsymbol{b}^k + \boldsymbol{w}^k\boldsymbol{h}^{k-1}) \tag{1}$$

(1)式中:$\boldsymbol{b}^k$ 是偏移矢量;$\boldsymbol{w}^k$ 是权重矩阵。映射函数采用双曲正切函数(tanh)(以元素方式应用),也可以用其他饱和非线性函数(例如 Sigmod 形函数)代替。顶层 $\boldsymbol{h}^l$ 与监督的目标输出 y 组合成凸型损失函数$L(\boldsymbol{h}^l,\boldsymbol{y})$。

输出层使用(2)式所示的线性回归函数

$$\boldsymbol{h}_i^l = \frac{\mathrm{e}^{b_i^l + \boldsymbol{w}_i^l\boldsymbol{h}^{l-1}}}{\sum_j \mathrm{e}^{\boldsymbol{b}_j^l+\boldsymbol{w}_j^l\boldsymbol{h}^{l-1}}} \tag{2}$$

(2)式中:$\boldsymbol{w}_i^l$ 是 $\boldsymbol{w}^l$ 的第 i 行;$\boldsymbol{h}_i^l$ 是正的,并且 $\sum_i \boldsymbol{h}_i^l = 1$。在这种情况下,使用(3)式所示的条件对数似然函数作为损失函数,其在(x,y)对上的期望值最小化。

$$L(\boldsymbol{h}^l,y) = -\log P(Y=y\mid x) = -\log(\boldsymbol{h}_y^l) \tag{3}$$

训练采用基于梯度训练的 BP 算法[24],使用包括代价函数的反向传播导数,测量输入 x 和期望输出 y 之间产生的差异。然后找到每个单元的最佳权重,使得代价函数最小化。

以 *softmax*[15]作为输出函数,定义交叉熵误差函数 $E(w)$为

$$E(w) = \sum_{i=1}^{N} \log(1 + \exp(-\boldsymbol{y}^n\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^n)) \tag{4}$$

(4)式中,N 为单元数量。

通过从 w 得到相对于每个权重 w_k 的代价函数 $E(w)$来获得 $E(w)$的梯度,其定义为

$$\boldsymbol{g}_k = \frac{\partial}{\partial(\boldsymbol{w}_k)}E(w) \tag{5}$$

根据文献[14],梯度下降可以写成

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \boldsymbol{\eta}_k g_k \tag{6}$$

(6)式中,$\boldsymbol{\eta}_k$ 是步长或学习速率。

最后,根据(4)—(6)式,使用对应的随机梯度下降算法来确定最小化代价函数的最优参数,从而得到最佳权重。

平均音声学模型训练中,采用包括输入层、输出层和 6 个隐藏层的 DNN 结构,在隐藏层中使用 tanh 函数,在输出层处进行线性激活,训练语料库中的所有说话人的训练数据共享隐藏层,因此,隐藏层是所有说话人共享的全局语言特征映射。每个说话人具有自己的回归层,以对自己的特定声学空间进行建模。经过多批次 SGD 训练,得到最优的多说话人的 AVM 模型(平均时长模型和平均声学特征模型)。

1.3　说话人自适应

在说话人自适应阶段,给定目标说话人的目标情感的小语料库,提取其声学特征。声学特征的提取与 AVM 训练过程相同,包括 F0,MGC,BAP 和 V/UV。将平均音模型训练阶段获得的多说话人平均音模型放入目标情感说话人的 DNN 模型中,进行说话人自适应变换,得到说话人相关的自适应模型,即说话人相关的时长模型与说话人相关的声学模型。说话人自适应获得的说话人相关模型与平均音模型的 DNN 结构相同,采用 6 个隐藏层结构,映射函数与(1)式相同。

1.4　情感语音合成

在情感语音合成阶段,首先输入待合成的情感语音的文本,经过文本分析过程获得的上下文相关标注,并将上下文相关标注作为语言特征作为说话人相关模型的输入,采用最大似然参数生成(maximum likelihood parameter generation, MLPG)算法[25]生成目标情感语音的声学参数,然后采用 WORLD 声码器,合成得到目标说话人的目标情感语音。

2 情感语音合成实验

由于男女录音人的基频差异较大,为了避免性别差异对合成语音质量的影响,本文选取 9 位女性大学生作为录音人录制情感语料。情感语音的录制采用诱导激发方式,在录制某种情感的语音时,首先利用故事、视频或图片激发出录音人的情感,然后让录音人在激发的情感状态下朗读录音文本。每个说话人录制 11 种情感的语音,每种情感录制 300~350 句,总共录制 3 500 句。录音在专业录音棚中进行,每句录音的时长为 5~8 s,采样率为 16 kHz,量化精度为 16 位。

实验中,用 8 位女性说话人的情感语料训练 DNN 平均音模型(AVM),用 1 位女性说话人的情感语料作说话人自适应的目标说话人训练语料。其中,每种情感使用 250~300 句情感语音进行平均音模型训练,50 句情感语音进行目标情感说话人自适应训练,其中,40 句用作训练集,10 句用做测试集,总共对 11 种典型情感(放松、惊奇、温顺、喜悦、愤怒、焦虑、厌恶、轻蔑、恐惧、悲伤,中性)进行自适应训练。对语音文件,使用 WORLD 声码器以 5 ms 的步长以对数标度提取 60 维广义梅尔倒谱系数,1 维频带非周期分量和 1 维基频。

DNN 的输入包含 416 个二进制语言特征、9 个数字特征和 1 个表示性别的二进制特征。语言特征包括声韵母、音节、词性,以及韵律词、韵律短语、语句长度、位置等信息。9 个数字特征涉及 HMM 状态、声韵母中的帧位置、声韵母中的状态位置和声韵母持续时间。输出声学特征包括 60 维 MGC,1 维 BAP,1 维 F0,以及它们的一阶差分和二阶差分特征,以及 1 个 V/UV 特征,构成 187 维的声学特征向量。对 F0 进行线性内插以提取动态特征,V/UV 特征用于在语音合成时决定有声和无声区域。输入特征被归一化到[0.01,0.99],输出特征用说话人相关的均值和方差归一化。说话人自适应的声学特征也用同样的方法进行归一化。在参数生成阶段,将 MLPG 算法应用于输出特征,以产生平滑的参数轨迹,之后在倒谱域中进行频谱增强和后滤波。

DNN 模型有 6 个隐藏层,每个隐藏层有 1 024 个结点。在隐藏层中使用 tanh 函数,在输出层处进行线性激活。在 AVM 训练和说话人自适应期间,mini-epoch 大小设置为 256,并采用动量加速收敛。对于前 10 个 epoch,首先将动量设置为 0.6,然后增加到 0.9。在 AVM DNN 训练的前 10 个时期中使用 0.000 8 的固定学习速率。在说话人自适应期间,说话人的学习率设置为 0.02。在 10 个 epoch 之后,学习速率在每个 epoch 减半。将 L2 正则化应用于具有罚分因子为 0.000 01 的权重,AVM DNN 训练和说话人自适应的最大 epoch 数设置为 25。实验使用 Merlin[26] 工具箱,在训练中使用 CUDA8.0,theano 框架及 Python 模块在 GPU 上进行矩阵计算。

为了评估本文提出的方法,采用了以下 3 组实验进行对比。

1) DNN(adp):利用 9 位说话人进行 DNN 说话人自适应训练;

2) DNN:利用 1 位说话人进行 DNN 模型训练;

3) HMM(adp):利用 9 位说话人进行基于 HMM 的说话人自适应训练[27-28]。

在对比实验中,2) 的训练过程与 1) 相同,但 2) 只使用 1 位说话人的训练语料,而 1) 使用 9 位说话人的训练语料,并且进行了说话人自适应。

3 实验结果评测

本文采用客观和主观测量对合成语音进行评测,根据原始说话者的自然语音与合成语音之间的失真来测量情感语音合成质量。

3.1 客观评测

客观评测通过计算原始语音与合成语音在基频和时长上的均方根误差(root mean square error, RMSE),以及梅尔倒谱失真(Mel-cepstral distortion, MCD)、频带非周期性失真(band a periodicity distortion, BAPD) 和 V/UV 交换误差来评测合成语音的质量,评测结果分别如表 1 和表 2 所示。可以看出,在大部分情感中,说话人自适应方法训练的 DNN 和 HMM 模型比 DNN 模型数值更小,有更好的性能。几种情绪浮动较大的情感如愤怒、焦虑的 RMSE 值在 DNN 模型上浮动较大,其 RMSE 值有时会接近或小于基于 HMM 的说话人自适应训练方法的结果,而轻蔑和中性等弱情感在 3 种对比实验中的 RMSE 值比较接近。表 1 与表 2 结果表明,本文提出的方法合成的情感语音要优于其他 2 种方法合成的情感语音。

3.2 主观评测

主观评测方法采用在不同方法合成的语音句子对之间的 AB 偏好测试、平均意见得分(mean opin-

ion score, MOS)和情感相似度平均意见得分 (emotional mean opinion score, EMOS)来评估合成语音的质量。

表 1　3 种情感语音合成方法的基频(F0)和时长(Dur)的 RMSE

Tab.1　RMSE of F0 and duration (Dur) for three emotional speech synthesis methods

情感	F0/Hz			Dur/s		
	HMM (adp)	DNN	DNN (adp)	HMM (adp)	DNN	DNN (adp)
悲伤	58.6	65.5	52.0	0.232	0.238	0.202
愤怒	66.6	62.8	60.5	0.118	0.111	0.090
放松	25.0	31.6	23.1	0.176	0.183	0.106
焦虑	67.5	67.6	47.5	0.124	0.136	0.087
惊奇	66.4	71.3	64.3	0.171	0.153	0.132
恐惧	68.6	78.8	62.8	0.133	0.138	0.088
轻蔑	51.8	51.3	49.1	0.132	0.146	0.095
温顺	33.8	42.2	32.4	0.149	0.153	0.105
喜悦	72.2	75.0	61.9	0.106	0.113	0.106
厌恶	61.8	64.6	58.8	0.143	0.145	0.094
中性	41.4	43.6	39.8	0.143	0.139	0.102

表 2　3 种情感语音合成方法的客观评测结果

Tab.2　MCD, BAPD and V/UV error of three emotional speech synthesis methods

情感		MCD/dB	BAPD/dB	V/U/%
悲伤	HMM(adp)	6.21	0.14	8.66
	DNN	7.15	0.18	10.96
	DNN(adp)	5.75	0.14	8.35
愤怒	HMM(adp)	9.01	0.39	23.85
	DNN	9.71	0.47	25.49
	DNN(adp)	8.58	0.36	20.78
放松	HMM(adp)	6.94	0.23	16.75
	DNN	7.56	0.28	18.17
	DNN(adp)	6.77	0.22	15.08
焦虑	HMM(adp)	7.07	0.32	8.91
	DNN	7.92	0.41	13.51
	DNN(adp)	7.01	0.36	9.81
惊奇	HMM(adp)	7.78	0.40	17.33
	DNN	8.20	0.42	19.14
	DNN(adp)	6.99	0.32	13.69
恐惧	HMM(adp)	8.04	0.26	20.62
	DNN	8.64	0.29	22.32
	DNN(adp)	7.18	0.19	10.19
轻蔑	HMM(adp)	6.70	0.26	14.76
	DNN	8.14	0.34	19.82
	DNN(adp)	6.65	0.25	11.30
温顺	HMM(adp)	7.10	0.24	8.02
	DNN	7.67	0.34	12.29
	DNN(adp)	6.60	0.25	7.23
喜悦	HMM(adp)	7.69	0.32	12.70
	DNN	8.46	0.41	14.57
	DNN(adp)	7.58	0.31	10.14
厌恶	HMM(adp)	7.25	0.30	10.38
	DNN	8.14	0.39	15.78
	DNN(adp)	6.60	0.28	10.72
中性	HMM(adp)	7.87	0.31	10.06
	DNN	9.08	0.44	17.52
	DNN(adp)	7.53	0.28	9.42

本次实验合成了 3 个组别相同说话人的 11 种情感的语料,每种情感各 10 句,总共 330 句情感语音用于评价,邀请 5 个评测者进行评测。在偏好测试中,有 3 种偏好选择:①前者更好;②后者更好;③没有偏好或中性(成对句子之间的差异不能被感知或可以被感知,但难以选择哪一个更好)。在 MOS 和 EMOS 评测中,在听到语音之后,评测者在 5 级评分(1:差,2:较差,3:一般,4:好,5:优)中评价语音的自然度和情感度。结果如图 2—图 4 所示。从图 2—图 4 中可以看出,本文提出的方法(DNN adp)相较于其他 2 种方法分别获得了 62%和 44%的偏好度,MOS 和 EMOS 评测中得到了 3.7 和 3.5 的平均分,优于其他 2 种方法,说明本文提出的情感语音合成方法具有更好的偏好度、自然度和情感度。

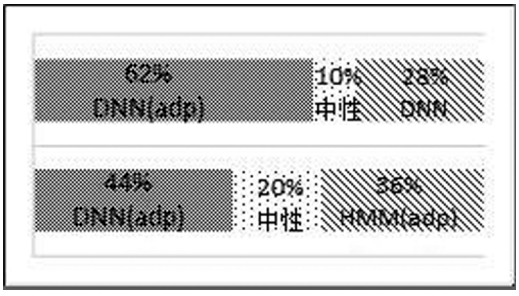

图 2　情感语音合成偏好得分

Fig.2　Preference score of emotional speech synthesis

4　结束语

本文提出了一种利用多个说话人的多情感训练语料,采用说话人自适应方法实现基于 DNN 的情感语音合成方法,实现了情感语音合成。主客观实验均表明,本文提出的情感语音合成方法要优于传统

的基于 HMM 的情感语音合成方法和基于 DNN 的情感语音合成方法。下一步工作将采用递归神经网络、长短时记忆网络等不同的深度学学习方法实现情感语音合成,扩充情感语料库的规模,并对不同方法、不同语料库合成的情感语音进行深入评测。

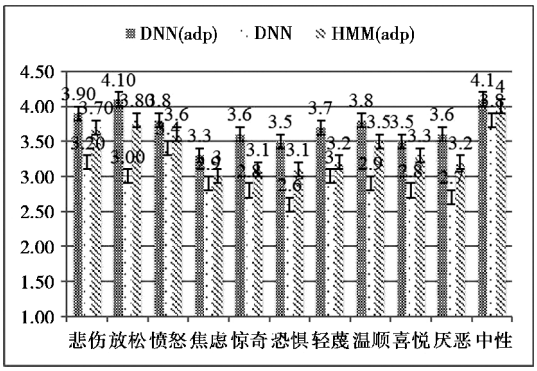

图 3 95%置信区间下 3 种方法合成的情感语音的 MOS 得分

Fig.3 MOS score of emotional speech synthesized by three methods with 95% confidence interval

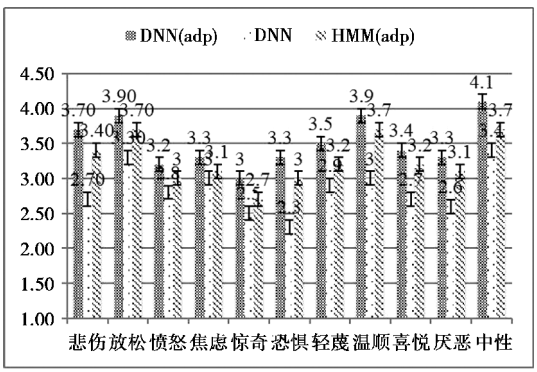

图 4 95%置信区间下 3 种方法合成的情感语音 EMOS 得分

Fig.4 EMOS score of emotional speeches synthesized by three methods with 95% confidence interval

参考文献:

[1] FAN B, LEE S W, TIAN X, et al. A waveform representation framework for high-quality statistical parametric speech synthesis [C] //Asia-Pacific Signal and Information Processing Association Summit andConference. New York:IEEE Press, 2015:530-536.

[2] 鲁小勇, 杨鸿武, 郭威彤,等. 基于 PAD 三维情绪模型的情感语音韵律转换[J].计算机工程与应用, 2013, 49(5):230-235.
LU Xiaoyong, YANG Hongwu, GUO Weitong, et al. Prosody conversion of emotional speech based on PAD three dimensional emotion model[J].Computer Engineering & Applications,2013,49(5):230-235.

[3] 李贤, 於俊, 汪增福. 面向情感语音转换的韵律转换方法[J].声学学报,2014(4):509-516.
LI Xian, YU Jun, WANG Zengfu. Prosody conversion for mandarin emotional voice conversion[J].Shengxue Xuebao/acta Acustica,2014,39(4):509-516.

[4] TOKUDA K, NANKAKU Y, TODA T, et al. Speech synthesis based on hiddenmarkov models [J]. Proceedings of the IEEE,2013,101(5):1234-1252.

[5] 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.

[6] 王坚, 张媛媛.基于深度神经网络的汉语语音合成的研究[J].计算机科学,2015,42(s1):75-78.
WANG Jian, ZHANG Yuanyuan.Title research on deep neural network based chinese speech synthesis[J].Computer Science, 2015, 42(s1):75-78.

[7] ZEN H, SENIOR A, SCHUSTER M. Statistical parametric speech synthesis using deep neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE Press, 2013: 7962-7966.

[8] LING Z H, DENG L, YU D. Modeling spectral envelopes using restrictedboltzmann machines and deep belief networks for statistical parametric speech synthesis[J].IEEE Transactions on Audio Speech & Language Processing, 2013, 21(10):2129-2139.

[9] ZEN H, SENIOR A. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis [C]//IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE Press, 2014:3844-3848.

[10] QIAN Y, FAN Y, HU W, et al. On the training aspects of deep neural network (DNN) for parametric TTS synthesis[C]//IEEE International Conference on Acoustics, Speech and SignalProcessing. New York: IEEE Press, 2014:3829-3833.

[11] WU Z, VALENTINI B, WATTS O, et al. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. New York:IEEE Press,2015:4460-4464.

[12] WATTS O, HENTER G, MERRITT T, et al. From HMMS to DNNS: Where do the improvements come from? [C]//IEEE International Conference on Acoustics, Speech and SignalProcessing. New York: IEEE

Press,2016:5505-5509.

[13] HASHIMOTO K, OURA K, NANKAKU Y, et al. The effect of neural networks in statistical parametric speech synthesis[C]//IEEE International Conference on Acoustics, Speech and SignalProcessing. New York: IEEE Press,2015:4455-4459.

[14] FAN Y, QIAN Y, SOONG F K, et al. Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis [C] //IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE Press,2015:4475-4479.

[15] WU Z, SWIETOJANSKI P, VEAUX C, et al. A study of speaker adaptation for DNN-based speech synthesis [C]//Interspeech.Lyon:ISCA,2015:879-883.

[16] FAN Y, QIAN Y, SOONG F K, et al. Unsupervised speaker adaptation for DNN-based TTS synthesis[C]// IEEE International Conference on Acoustics, Speech and SignalProcessing. New York: IEEE Press, 2016: 5135-5139.

[17] SHAN Y, WU Z, LEI X. On the training of DNN-based average voice model for speech synthesis[C]//Signal and Information Processing Association Summit andConference.New York:IEEE Press,2017:1-6.

[18] SKERRY R R J, ERIC B, YING X, et al.Towards end to end prosody transfer for expressive speech synthesis with Tacotron[EB/OL].(2018-03-24)[2018-04-19].https://arxiv.org/abs/1803.09047.

[19] SARAH T, TAEHWAN K, YISONG Y, et al. A deep learning approach for generalized speech animation[J]. ACM Transactions on Graphics, 2017, 36(4):93.

[20] INOUE K, HARA S, ABE M, et al. An investigation to transplant emotional expressions in DNN-based TTS synthesis [C] //Asia-Pacific Signal and Information Processing Association Summit andConference. New York: IEEE Press,2017:1253-1258.

[21] AN Shumin, LING Zhenhua, DAI Lirong. Emotional statistical parametric speech synthesis using LSTM-RNNs [C]// Asia-Pacific Signal and Information Processing Association Summit and Conference. New York: IEEE Press,2017:1613-1616.

[22] MORISE M, YOKOMORI F, OZAWA K. WORLD: Avocoder-based high-quality speech synthesis system for real-Time applications[J].Ieice Transactions on Information & Systems,2016, 99(7):1877-1884.

[23] DENG L, YU D. Deep learning: methods and applications[J].Foundations & Trends © in Signal Processing, 2014,7(3):197-387.

[24] LAROCHELLE H, BENGIO Y, LOURADOUR J, et al. Exploring strategies for training deep neural networks[J]. Journal of Machine Learning Research, 2009, 1(10):1-40.

[25] HWANG H T, TSAO Y, WANG H M, et al. A probabilistic interpretation for artificial neural network-based voice conversion [C]// Asia-Pacific Signal and Information Processing Association Summit andConference. New York:IEEE Press,2015:552-558.

[26] WU Z, WATTS O, KING S. Merlin: an open source neural network speech synthesis system [C]//Speech Synthesis Workshop. Lyon:ISCA,2016:202-207.

[27] WU P, YANG H, GAN Z. Towards realizing mandarin-tibetan bi-lingual emotional speech synthesis with mandarin emotional training corpus[C]//International Conference of Pioneering Computer Scientists, Engineers and Educators. Singapore:Springer,2017:126-137.

[28] YAMAGISHI J, KOBAYASHI T, NAKANO Y, et al. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm[J].IEEE Transactions on Audio Speech and Language Processing,2009, 17(1):66-83.

作者简介:

智鹏鹏(1992—),男,河南洛阳人,硕士研究生,主要研究方向为语音信号处理。E-mail:327005363@qq.com。

杨鸿武(1969—),男,甘肃合作人,教授,博士,博士生导师,主要研究方向为自然语言处理、语音信号处理。E-mail:yanghw@nwnu.edu.cn。

宋　南(1990—),男,河北迁安人,硕士研究生,主要研究方向为语音信号处理。E-mail:904782919@qq.com。

(编辑:王敏琦)