ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 5,Sep. 2019,pp. 753－770
DOI:10. 16337/j. 1004-9037. 2019. 05. 001

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

语音转换技术研究现状及展望

张雄伟　苗晓孔　曾　歆　孙　蒙　曹铁勇

(陆军工程大学,南京,210007)

摘　要:语音转换通常是指将一个人的声音个性化特征通过“修改变换”,使之听起来像另外一个人的声音,同时保持说话内容信息不变。近年来,随着信息处理和机器学习技术的快速发展,语音转换技术也得到了突飞猛进的进步。为此,在简要介绍语音转换基本概念的基础上,重点综述了近几年语音转换的典型模型和方法,分析了语音转换的关键技术,列举了语音转换技术的主要应用场景,梳理了目前语音转换中仍存在的若干技术问题,并展望了语音转换研究的发展方向。

关键词:语音处理;语音转换;神经网络;频谱转换;韵律转换

中图分类号:TN912　　**文献标志码**:A

Voice Conversion:The State of the Art and Prospects

Zhang Xiongwei, Miao Xiaokong, Zeng Xin, Sun Meng, Cao Tieyong

(Army Engineering University, Nanjing, 210007, China)

Abstract: Voice conversion usually refers to the process of modifying and transforming the personalized features of a person's voice to make it sound like another person's voice while keeping the linguistic information unchanged. In recent years, with the rapid development of information processing and machine learning, the technology of voice conversion has also achieved great progress. On the basis of introducing the basic concepts of voice conversion, the typical models and methods of voice conversion researched in recent years are summarized in this paper, the key technologies of voice conversion are reviewed, the main application scenarios of voice conversion technology are listed, and some technical problems still existing in voice conversion at present are briefly introduced. Finally, the prospects for the directions of the research and development of voice conversion are given.

Key words: speech processing; voice conversion; neural networks; spectrum conversion; prosody conversion

引　　言

随着人工智能应用领域的不断扩大和发展,智能语音交互、个性化语音生成等技术逐步受到人们的关注。语音转换作为个性化语音生成的一种重要技术和手段,涉及语音信号处理、人工智能、模式识别、语音学等多方面学科领域,是当今语音处理研究领域的热点和难点,近年来越来越引起学者的重视[1]。

收稿日期:2019-02-10;**修订日期**:2019-05-20

广义上讲,人们把改变语音中说话人个性特征的语音处理技术统称为语音转换[2-5],广义的语音转换可分为非特定人语音转换和特定人语音转换两大类。非特定人语音转换是指通过技术处理,使得转换后的语音不再像原说话人的声音。而在实际研究和应用中,语音转换通常是指改变一个说话人,即源说话人(Source speaker)的语音个性特征,如频谱、韵律等,使之具有另外一个特定说话人,即目标说话人(Target speaker)的个性特征[6-7],同时保持语义信息不变的技术。一般来说,特定人语音转换的技术难度要高于非特定人语音转换。

语音转换研究的相关工作最早可追溯至20世纪六七十年代,至今已经有50多年的研究历史,但真正受到学术界和产业界广泛关注则是近十多年的事情。近年来,语音信号处理和机器学习等技术的进步,以及大数据获取能力和大规模计算性能的提高有力地推动了语音转换技术的研究及发展[8]。特别是基于人工神经网络(Artificial neural network, ANN)的语音转换方法的兴起,使得转换语音的质量进一步得到提升。国内较早进行语音转换研究的机构包括中国科学院、中国科学技术大学、国防科技大学、亚洲微软研究院、IBM中国研究院等[9]。近年来,东南大学、南京邮电大学、华南理工大学、苏州大学、哈尔滨工业大学、西北工业大学、陆军工程大学等多所高校以及腾讯、科大讯飞和百度等多家企业也开始此项技术研究,并相继取得了一些的研究成果。2016年,来自中、日、英等国语音转换领域的科学家组织了VCC2016语音转换竞赛,在统一的数据集上,对17个国际著名的语音研究小组提交的系统做了统一的评价和分析,为语音转换研究提供了数据平台和性能标尺。2018年VCC2018也如期举办,语音转换方法再次推陈出新,且转换语音的质量也得到明显提升。

本文在简要介绍语音转换原理的基础上,重点梳理了语音转换的相关方法和研究进展,归纳了语音转换的关键技术和应用,并总结了目前语音转换中仍存在的问题和挑战,对语音转换未来的发展方向作出展望。

1 语音转换的基本原理及框架

研究表明,语音中的声道谱信息、共振峰频率和基音频率等参数是影响语音个性特征的主要因素[10]。了解语音转换的原理和过程,将有助于提取语音成分中的个性特征,更好地实现语音转换。语音转换的基本原理如图1所示。

通常一个完整的语音转换方案由反映声源特性的韵律转换和反映声道激励特性的频谱(或声道谱)

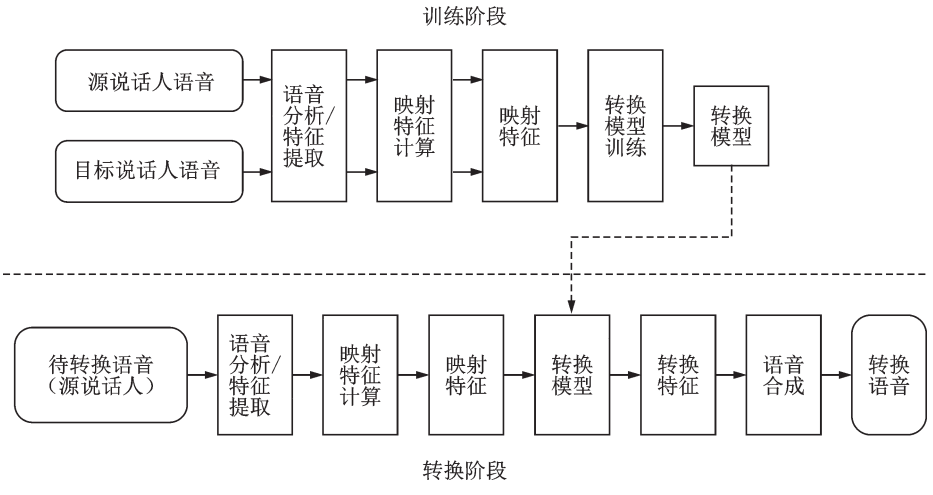

图1 语音转换原理图

Fig.1 Schematic design of voice conversion

转换两部分组成[11-13]。韵律的转换主要包括基音周期的转换、时长的转换和能量的转换[9],而声道谱转换表现为共振峰频率、共振峰带宽、频谱倾斜等转换。因为声道谱包含人更多的声音个性特征,且转换建模相对复杂,是制约语音转换效果的主要原因。因此,目前的语音转换研究也主要集中在对声道谱的转换方面[11]。

根据图1的语音转换原理可知,一个语音转换系统通常包含训练和转换两个阶段。训练阶段,首先对源说话人和目标说话人的语音进行分析和特征提取,然后对提取特征进行映射处理,最后对这些映射特征进行模型训练,进而得到转换模型。转换阶段,对待转换源语音进行分析、特征提取和映射,然后用训练阶段获取的转换模型对映射特征进行特征转换,最后将转换后的特征用于语音合成得到转换语音。

目前实现语音转换的方法中多数采用的是源和目标说话人语音特征参数间的匹配映射方式,且均在同一个语音分析合成的模型框架下进行。随着信息技术的不断进步,出现了序列到序列、波形到波形等语音转换方法,而且可用于语音转换的模型也越来越多。

2 语音转换的典型模型和方法

为了便于实现语音转换,本节介绍了声道谱转换和韵律转换两方面的研究现状。通过对研究现状的分析可知,目前语音转换研究主要集中在对声道谱的建模和转换规则方面,而对韵律的建模和转换研究尚不够充分。

2.1 声道谱转换方法

声道谱转换中较为常用的参数有幅度谱、对数谱、倒谱、线性预测系数等基本参数以及动态差分、本征空间短时谱[14]等变换参数。目前,对声道谱转换模型的研究通常是在对源和目标说话人语音进行统计分析的基础上,通过参数映射方式实现。声道谱转换研究经历了从离散映射到连续映射、从单帧映射到音段映射、从线性映射到非线性映射、从单一方法到多方法融合的过程,转换性能不断提升。训练条件也从大数据量、平行语音到小数据量、非平行语音过渡[9]。声道谱转换是语音转换中的重点和难点,也是目前语音转换需重点解决的问题。

2.1.1 基于码书映射的转换方法

码书映射的方法首先通过矢量量化的方法有效减少源与目标语音的特征数量,然后将最接近源码书的质心矢量通过聚类方法转换成相应的目标码书,从而实现语音转换。Abe等[15]于1988年首次提出将该方法用于语音转换,方法的转换流程图如图2所示[16]。但由于这种方法在量化时会造成特征空间的不连续,且忽略了帧间信息,所以转换效果不够理想。

此后,针对上述问题的改进方案陆续被提出,1991年,Shikano等[17]提出模糊矢量量化算法的码书映射;1997年,Kim[18]提出利用隐马尔科夫解决帧间信息的码书映射;1999年,Arslan[19]构建了一种码字线性加权的码书映射方法。这些方法均是通过解决码书映射不连续问题来提高转换语音的质量,但同时也造成了过平滑等其他问题的产生。此后,2005年,Wang等[20]通过引入分级码书的方法来解决转换精度的问题。2011年,Eslami等[21]提出在连续语音段上利用码书映射,在解决码书不连续问题的同时保留较好谱结构信息。2015年,胡芳等[22]提出基于码书映射的语音转换改进算法,通过基于转换权重预测重构码书映射关系的方法提升转换语音质量。总的来说,码书映射的方法实现原理简单,量化矢量均来自目标特征空间,语音频谱信息保留较为完整,但存在过平滑问题,且语音转换的精度还有待进一步提升。

2.1.2 基于高斯混合模型的转换方法

高斯混合模型(Gaussian mixture model, GMM)是一种采用若干个基于高斯概率密度的函数来精确

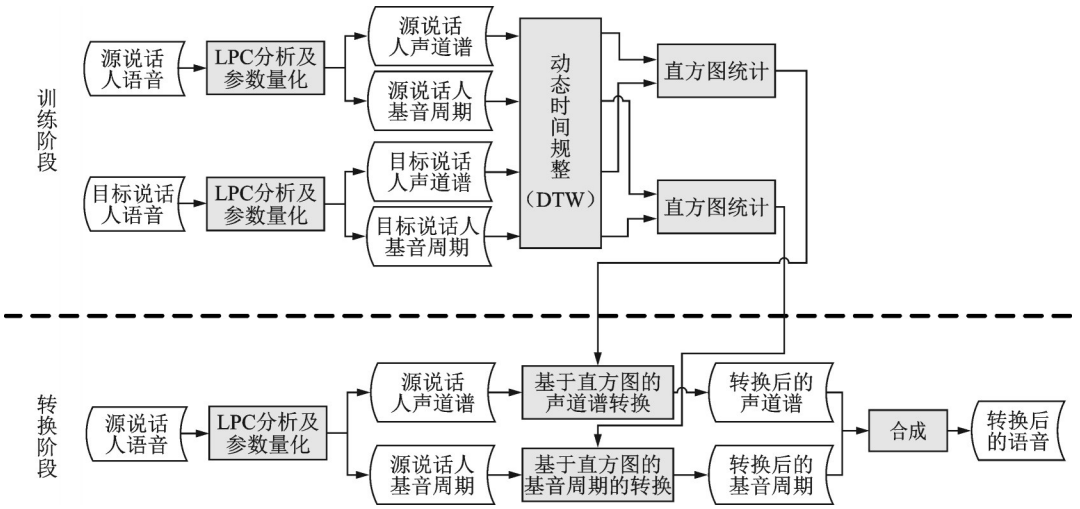

图2 基于码书映射的语音转换流程图[16]

Fig.2 Voice conversion flow chart based on codebook mapping[16]

量化事物的概率模型，即采用一组高斯函数的加权求和结果来表达观测数据的概率分布，如式(1)所示[16,23-24]

$$p(\boldsymbol{x}) = \alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i) \tag{1}$$

式中：x为观测数据，p为其维度；α_i为权重系数(需要满足$\alpha_i \geqslant 0$且$\sum_{i=1}^{m}\alpha_i = 1$)，其中$m$为高斯分量数目，则$N(\boldsymbol{x}; \mu_i, \Sigma_i)$表示均值为$\mu_i$、协方差矩阵为$\Sigma_i$的高斯分布。

针对码书映射中特征空间不连续的问题，Stylianou等[25]提出引入GMM对声道谱特征进行建模，使用基于概率的“软”聚类代替基于矢量量化的“硬”聚类，该方法获得了很好的效果，提升了语音转换的质量。但这种方法仅在源特征矢量上进行估计，而不是联合特征矢量估计，也就是说帧间信息考虑不足，所以极易出现过拟合和过平滑问题。受此启发，对基于高斯混合的统计映射模型改进的研究逐步展开。

在经典GMM的基础上，Chen等[26]提出采用最大后验概率自适应来转换声道函数；申毅等[27]提出依据后验概率大小和前后语音相关性来改进基于GMM模型的转换系统。Toda等[28]提出采用最大似然估计来改进；Helander等[29]则将最小二乘法引入GMM统计映射方法中来解决训练中的过拟合问题。近年来，针对改进GMM的方法仍在研究和发展中，例如在2012年，Helander等[30]提出了Dynamic Kernel PLS(DKPLS)转换方法，把传统的GMM方法中的线性映射拓展到非线性空间，提高了转换精度；2013年，宋鹏等[31]提出基于混合Gauss归一化的语音转换方法，成功将该模型运用到非对称语料库中；2015年，王明明[32]提出基于GMM和码书映射相结合的语音转换方法，减少GMM产生的过平滑问题；2016年，Kobayashi等[33]提出了倒谱残差的GMM，将频域变换转变为时域滤波，提高了转换时的计算效率。虽然基于GMM的方法不断完善和发展，但是由于GMM本身存在非一一映射情况，导致的过平滑问题一直没有得到根本解决，所以也限制该方法的进一步普及和应用。

2.1.3 基于隐马尔科夫模型的转换方法

隐马尔科夫模型(Hidden Markov model, HMM)是常用的统计分析模型之一，在语音识别、行为识别、自然语言处理等领域得到广泛应用。与GMM相比较，HMM在声道谱转换上的最大优势在于可利用自身的隐含状态及状态转移概率矩阵来对语音信号的动态变化进行建模[16]。1997年，Kim等[18]提出

并实现了基于HMM的语音信号的频谱建模和转换。随后基于HMM的转换方法得到进一步的研究和发展,2004年,Duxans等[34]提出将GMM的转换思路拓展到HMM中,采用联合HMM的方法将动态信息纳入考虑范畴,因而转换过程的鲁棒性更强。2006年,Wu等[35]提出一种基于Bi-HMM模型的语音转换算法,该方法利用HMM中的状态持续时间来刻画音素的时长信息,并采用Gamma函数分布来描述状态持续时间变量。基于Bi-HMM的语音转换不仅降低了转换后语音与目标语音的谱距离,而且极大地改善了语音韵律特性的转换,特别有利于语音情感特性的控制和转换。2010年,Qiao等[36]提出了一种基于HMM的帧序列到单帧的转换方法,有效解决了转换过程中帧间不连续问题。2011年,Zen等[37]提出了基于轨迹HMM的连续概率映射方法,有效地解决了传统HMM考虑动态特征参数后训练和转换不一致的问题。虽然HMM的语音转换方案也在不断更新完善,但由于HMM的隐含状态数目受限,造成了语音信号的动态变化范围受限,进而制约了转换处理精度,故该方法在实际转换应用中也有一定局限。

2.1.4 基于频率弯折转换方法

基于频率弯折的语音转换方法是指通过沿频率轴拉伸或压缩频谱,来调整共振峰的位置和带宽,并通过幅度缩放来调整每个频率中的能量大小,从而实现源到目标说话人的频谱映射。其转换示意图如图3所示。

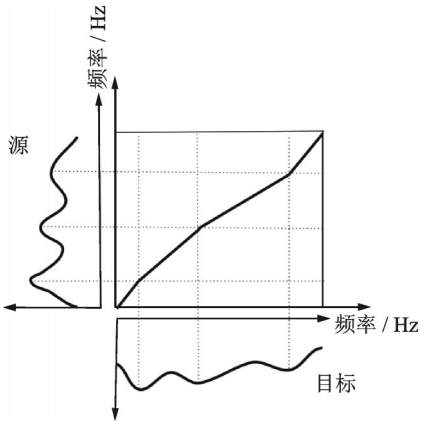

图3 分段线性频率弯折示意图[8]

Fig.3 Diagram of piecewise linear frequency warping[8]

基于频率弯折的语音转换方法主要分为基于动态规划的频率弯折(Dynamic frequency warping,DFW)和基于映射共振峰的频率弯折[38]两类。采用频率弯折的语音转换方法最早于1992年由Valbret等[39]提出,虽然该方法最大程度地保留了频谱的结构信息,但由于频谱调整幅度的限制,导致转换语音自然度高,但相似度较低。针对该问题,Sündermann等[40]结合声道谱归一化技术提出了单参数弯折函数和多参数分段线性函数的频率弯折方法,对源说话人语音声道谱进行弯折处理。双志伟等[41]提出将频率弯曲与单元挑选相结合来提高相似度的方法。后续关于频率弯折的研究还有很多,近些年将频率弯折方法与GMM、字典映射、最大谱相关等方法结合起来实现语音转换的方法也陆续出现[42-46]。综上,采用基于频率弯折来实现语音转换的方法,能够最大程度地保持语音自然度,且转换语音质量较高,但是其在相似度方面略显不足,还需结合其他方法以获得进一步提升。

2.1.5 基于神经网络的转换方法

ANN模仿人类神经网络行为特征,为一种模仿动物神经网络行为特征并进行分布式并行信息处理的数学模型。这种网络依靠系统的复杂程度,通过调整内部大量节点之间相互连接的关系,从而达到处理信息的目的[47]。因为神经网络对非线性转换具有良好的效果,所以在语音转换过程中也得到了应用。早期采用神经网络实现语音转换的方法主要集中于对神经网络隐层个数和逻辑单元上的改进,但一般不超过3个隐层且大多为前馈神经网络[48-50]。近几年在深度学习技术的推动下,能有效表示高维序列数据的深度神经网络不断发展,如全卷积神经网络(Fully convolutional network, FCN)、生成对抗网络(Generative adversarial network, GAN)、双向长短时记忆网络(Bidirectional long short term memory, BLSTM)等均被用来实现谱序列到序列的高精度转换。例如:Huang等[51]提出基于结合变分自动编码器和全卷积网络的语音转换研究,Kaneko等通过序列到序列(seq2seq)的GAN模型初步研究了语音转换[52]以及语音质量增强中的过平滑问题[53],Huang等[54]提出的自动化评价指标可作为GAN的判别器,

Esteban 等[55]提出了适用于时间序列预测的循环式 GAN 用于生物数据建模，Takuhiro 等[56]在循环 GAN 的基础上进行改进，进一步提升语音转换效果。由于 BLSTM 网络兼顾了语音序列的上下文信息同时也是一种序列映射的神经网络，所以其转换的语音质量相对较好[57]。同时，将神经网络与其他转换模型结合的方法也不断涌现，Hsu 等[58]提出了一种针对谱转换任务中高维数据的非负矩阵分解问题的字典更新方法，Seyed 等[59]提出了具有独立于说话人预训练的深度神经网络语音转换。通过预训练好的深层自编码器和 ANN 权值构造了一个深层神经网络，然后利用反向传播对网络权值进行微调最终实现对特征的转换。Chen 等[60]提出一种新的基于深度神经网络的谱包络转换方法，通过 DNN 对级联不同模型的网络进行训练进而实现语音转换。

随着神经网络模型的不断改进和发展，结合不同语音特征采用不同的网络转换模型方法不断提出。神经网络转换的本质是参数的多元回归模型，通过增加网络训练层数、添加高维特征序列和增大训练数据量等多种手段可以有效提升转换语音的质量。随着参数的增多，模型的表示能力不断增强。但当前表现优异的深度学习模型，所依赖参数过多，在非合作模式下当训练数据不充分时，就会发生过拟合现象，导致性能急速下降。这也是基于神经网络实现语音转换方法所面临的共性问题。同时，对数据训练量的依赖也成为制约此类方法转换效果的一个重要因素。

2.1.6 基于波形生成的转换方法

针对转换合成语音产生过平滑的问题，2016 年 Google 公司的 Deepmind 团队提出采用 WaveNet 网络直接生成音频波形样本点的方法来解决[61]。该方法主要基于一条件概率建模的深度自回归模型，将语音的各种特征作为条件，如式(2)所示，通过训练找到合适的自回归模型。同时网络中还采用因果卷积、扩张卷积等多种模型。

$$P(X|\theta)=P(x_t|x_1,\cdots,x_{t-r},\theta) \tag{2}$$

式中：T 为样本点总数，θ 为条件特征向量，t 和 r 分别为采样数量和接收域大小，x_t 为当前时刻样本点。

该方法最初被用于文本转换为语音的系统(Text to speech, TTS)。通过这种条件建模方法产生的语音清晰度和自然度高、质量好且没有过平滑问题，但网络生成速度较慢。随着 Fast-WaveNet 网络的提出，该网络开始具有实用价值。2018 年，Niwa 等[62]首次提出将该网络用于语音转换，转换流程图如图 4 所示。由图 4 可知，转换过程中无需语音合成的单独步骤即可直接生成转换语音。其后，采用该网络的语音转换方法不断产生。

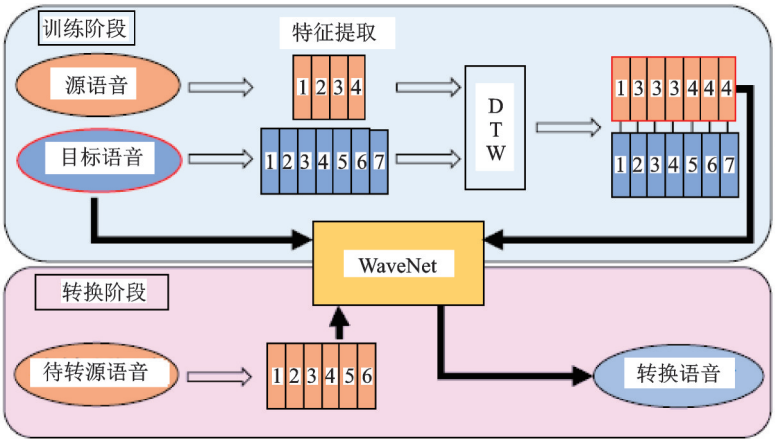

图 4 基于 WaveNet 的语音转换流程[62]

Fig.4 Overview of the method of voice conversion based on WaveNet[62]

中国科学技术大学在VCC-2018大赛上提出了结合Bi-LSTM先转换特征再进行特征条件概率建模的方法,有效提升了语音质量。Chen等[63]提出结合不同特征作为条件训练产生转换语音,文献[64-67]分别在此基础上对WaveNet的合成速度和合成条件展开了进一步研究。目前,基于WaveNet网络的语音转换正在发展中,还需不断优化结构以提升转换效率,针对该网络中逐样点生成语音波形易造成语音崩塌的现象,以及如何进一步提升转换语音自然度的问题仍有待深入研究。

2.1.7 *其他转换方法*

随着深度学习的进一步发展,多种多样的网络结构被提出,除了上述介绍的几类常见的声道谱转换方法外,还有一些其他转换方法同样值得关注,如:Wu等[68]提出了改进了基于时频模板的方法,既有效地保存了频谱细节,又减轻了转换负担。李娜等[69]将动态核方法、宋鹏等[70]将主成分回归用于声道谱参数映射,孙健等[71]采用基于卷积非负矩阵分解实现了语音转换,孙新建等[72]采用隐变量模型进行的语音转换,马振等[73]提出基于语音个人特征信息分离的语音转换等。可见字典映射转换、特征融合转换以及支持向量回归等方法都还有一定的应用空间。随着神经网络的不断发展,多网络模型融合的方法日益成为主流的转换方法,因此如何基于小样本数据,高效率实时实现声道谱转换,在既保证转换语音的高自然度和高相似度的同时,又保证转换算法的鲁棒性是一个需要重点关注的问题。

2.2 韵律转换模型方法

除了声道谱转换外,激励源转换对整个语音转换系统性能也举足轻重,主要包括韵律和非周期分量的转换。非周期分量作为激励的一部分,不少转换模型直接将其复制不做任何变换。Chen等[13]研究了基于深度神经网络的转换方法。对于此部分,为尽可能减少对转换效果贡献较小的运算,也可采用直接复制的方式,不对其进行转换。而韵律在反映目标说话人特征、情感状态、口音等特征上具有重要作用,所以韵律转换也值得关注。韵律建模通常是通过基频包络来实现,当前的研究主要体现在情感转换方面。由于韵律信息的复杂性及不稳定性,所以目前韵律转换主要集中在对基频包络F_0的转换,可直接取对数后做线性变换,如式(3)所示[74-76],或结合声道谱参数做回归分析。

$$p_t^{(Y)} = \frac{p_t^{(X)} - u^{(X)}}{\sigma^{(X)}} \times \sigma^{(Y)} + u^{(Y)} \tag{3}$$

式中:$p_t^{(Y)}$和$p_t^{(X)}$分别为转换后的$\log F_0$和原始$\log F_0$,$u^{(X)}$和$u^{(Y)}$为各自的均值,$\sigma^{(X)}$和$\sigma^{(Y)}$为标准差。线性转换的方法虽然简单,但映射稳定,不易产生“阴阳怪调”的问题,因此目前韵律转换仍基本采用上述方式。

除上述方法外,也有一些算法将语音分析后的残差信号或激励信号统一考虑,进行处理以求获取更精确的韵律模型。例如2004年,Ye等[77]在训练阶段成对保存目标说话人的激励源残差和声道谱参数,在转换阶段,以最近邻方式寻找与转换得到的声道谱参数对应的目标说话人声道谱参数,进而得到合成语音所需的激励源信号。2007年,Percybrooks等[78]提出通过建立残差转移概率矩阵来模拟残差信号帧间变化特性,收到良好效果。Kang等[79]提出采用音高目标模型来实现转换,通过GMM的方法训练转换模型,但是由于并未考虑上下文信息影响,所以存在一定弊端。Ming等[80]提出使用DBLSTM-RNN网络转换韵律中的F_0,这个网络能够兼顾上下文信息,但原始F_0的结构信息未保留,会导致其与频谱的合成过程中产生杂音,影响语音转换的质量。2015年,凌震华等[81]提出基于目标逼近特征和双向联想贮存器的情感语音基频转换方法,可以在目标情感数据较少的情况下取得更高的情感表现力。还有一些学者提出了F_0的其他转换方法[82-83],但是针对语音转换中,有时候除了单纯的韵律考虑外还需要考虑其频谱信息,不考虑F_0的结构信息,依然得不到理想的语音转换效果。

当然韵律转换还包含其他多方面的转换内容,单纯地对基频包络的转换达不到真正意义上的转换

效果,所以韵律转换目前还存在很多难点和问题。

3 语音转换关键技术

为更好地实现语音转换,提升转换语音的质量,一个完整的语音转换系统通常会涉及以下几项关键技术:语音分析与合成、语音特征参数提取、语音时间对齐、转换模型和规则训练以及转换效果评价[84]等,转换模型和规则训练第2节已介绍,此处不再赘述。

3.1 语音分析与合成

为了实现语音转换,语音信号分析与合成必不可少。早期的分析合成模型包括线性预测编码(Linear prediction coding, LPC)[85]、基音同步叠加(Pitch synchronous overlap and add, PSOLA)[86]和波形相似叠加(Waveform similarity overlap and add, WSOLA)等[87]。这些方法中LPC建模的语音信号质量较差,已逐渐淘汰,而后两者方法建模语音质量较高,但通常不具备语音分析和参数化能力,不适用于对语音个性特征的转换。谐波噪声模型(Harmonic noise model, HNM)[88]和STRAIGHT(Speech transformation and representation using adaptive interpolation of weighted spectrum)[89]因为重构语音质量高、参数容易控制而被广泛采用。HNM不能灵活处理相位,不像源滤波器那样可以灵活修改,但AHOCODER编码器提供高质量的HNM合成的工具包[90]。而STRAIGHT提出了一种基音自适应时频频谱平滑算法模型,能够减轻信号周期和频谱之间的干扰,在此基础上后来扩展到TANDEM-STRAIGHT[91],而后CheapTrick和WORLD又在该模型的基础上提出了一些改进。HNM和STRAIGHT这两种方法是目前主要的分析合成手段。

3.2 语音特征参数提取

经过语音分析合成器而获得的语音特征,理论上是可以直接用作特征对进行训练的,但是为了使所得信息能够更好地表征语音信号的个性特征,往往还需要对其进行进一步处理,从而获得更合适的映射特征。

表征语音个性的特征主要体现在3个层次上:一是音段信息,描述的是语音的音色特征,主要包括共振峰位置、共振峰带宽、频谱倾斜(Spectral tilt)、基音频率、能量等;二是超音段特征,描述的是语音的韵律特征,主要包括音素的时长、基音频率的变化(音调)、能量等;三是语言特征(Linguistic cues):包括习惯用语、方言、口音等[92]。目前语音转换中所提取的映射参数特征主要是音段信息的局部特征和超音段信息的上下文特征两类。局部特征主要是谱包络、倒谱和共振峰等参数,除此之外,较为常用的还有线谱对(Line spectrum frequency, LSF)参数[93]和考虑了人耳听觉特性的梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)[94]。上下文特征主要是指语音帧间的动态信息。

3.3 时间对齐

语音转换过程中通常是利用源和目标特征向量来训练源和目标特征之间的映射函数。从不同说话人语音特征空间把那些具有相同语义信息的特征参数进行匹配,然后利用这些配对参数设计和训练出转换模型。这里配对就是时间对齐,也叫时间规整。

3.3.1 平行语料

平行语料指的是源和目标说话人包括相同语言内容的语句。平行语句确保了源和目标语音具有时序一致、内容相同的语义信息,只是在各音素的持续时间上呈现不同。因此,必须使用时间对准方法来解决时间差异。最常用的方法是动态时间规整(Dynamic time warping, DTW)[95],计算每个话语对或在每个音素对之间的最佳时间对准。动态时域规整后,最终产生的是一对相等长度的源和目标特征序列。图5为2个时间序列的规整示意图。

图5中,上下两条实线分别代表源特征序列$X=[x_1,x_2,\cdots,x_m]$和目标特征序列$Y=[y_1,y_2,\cdots,y_n]$。2个序列之间的虚线代表2个时间序列之间的相似点。DTW使用所有这些相似点之间的距离的和,即归整路径距离,来衡量2个时间序列之间的相似性。

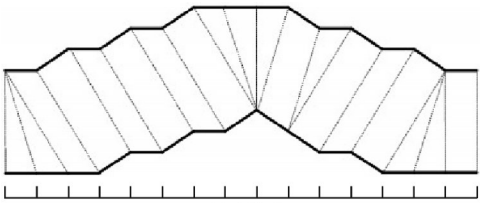

图5 2个时间序列的规整示意图

Fig.5 Warping of two time series

DTW对准策略是假定源和目标说话人的相同音素具有相似的特征(当使用特定距离度量)时成立[96]。然而,这种方法有时也会导致次优对准问题的产生。为了改善对准输出,可以迭代地执行目标特征和转换特征(而不是源特征)之间的对准,然后进行训练和转换,直至满足收敛条件。

3.3.2 非平行语料

对于非平行语料,由于语义信息不同或者语义信息虽有重叠,但时间顺序存在差异,因此此情况下的时间对齐算法相对复杂得多。但由于非平行语料相对于平行语料更易获取,故针对非平行语料的对齐研究也在不断发展。Sündermann等[97-98]针对非平行语料对齐问题先后提出基于分类的语音对齐和基于单元选择的语音对齐。Salor等[99]也提出类似单元选择的动态编程方法来实现语音帧的时间对齐。此后,Saito等[100]提出了一种噪声信道模型解决该问题。简志华等[101-102]提出的基于混合线性变换法和区域最邻迭代训练法等都是为了解决非平行语料中时间对齐的问题。近些年随着神经网络的不断发展,通过设计网络模型或中间变量直接实现音素级别的语音对齐方法也逐步发展起来。例如,Tian等[103]提出基于WaveNet的无声码器语音转换方法,该方法不需要处理中间特征,而是利用波形网直接将语音后验器映射到波形样本,这样就避免了声码器和特征转换引起的估计误差。Kameoka等[104]提出了一种非并行多对多语音转换方法,该方法使用了一种称为辅助分类器的条件变量变分自动编码器来实现非平行语料对齐。Yeh等[105]提出利用循环一致性对抗网络和变分自动编码器等模型应用于无并行数据的语音转换任务中。其他通过各式网络解决该问题的方法也不断出现,非平行语料的时间对齐问题已经逐步成为非平行语料的转换问题,从单一步骤中的对齐问题转化为由整个网络设计解决,这也使该问题得到了更好的处理。

3.4 转换效果评价

对语音转换方法性能的测试和评价是语音转换研究的重要组成部分之一,设计一个可信、高效的评价方案对于提高转换性能具有重要意义。目前,对语音转换方法性能优劣的测试和评价主要通过客观和主观2种手段来实现。

3.4.1 客观评价

客观评价建立在语音数据失真测度基础上,利用某种距离准则来测量转换后语音和原始目标语音间的相似程度,并由此得出对转换方法优劣的评价方法。主要的客观评价指标有均方误差(Mean square error, MSE)、谱失真(Spectral distortion, SD)和梅尔倒谱失真(Mel cepstral distortion, MCD),MSE、SD和MCD的值越小,说明失真越小,转换精度越高。

近几年,随着语音转换挑战赛事的举办,MCD成为评价转换语音质量的主要客观衡量指标,MCD的表达式如下[106]

$$\mathrm{MCD}=\frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{I}(m_{i\,\mathrm{con}}-m_{i\,\mathrm{tar}})^2} \tag{4}$$

式中:$m_{i\,\mathrm{con}}$和$m_{i\,\mathrm{tar}}$分别为第i维转换特征(con: converted)和目标特征(tar: target)的梅尔倒谱系数,I为梅

尔倒谱总维度,MCD的单位为dB。

3.4.2 主观评价

主观评价就是以人为主体,通过人的主观感受来对语音进行测试。由于语音信号最终是用来给人聆听的,因而人对语音转换效果好坏的感受是最为重要的评价结果。相对于客观评价来说,主观评价结果更具有可信度。主观方法对转换效果的评价一般从语音质量和说话人特征相似度2个角度进行,采用的方法主要是平均意见分(Mean opinion score, MOS)和ABX。

(1) MOS测试

MOS测试的主要原理是让测评人根据5个等级划分对测试语音的主观感受进行打分,它既可以用于对语音质量进行主观评价,也可以用于对说话人特征相似度的评价。MOS 分是对所有测试语句和所有测评人的综合平均结果。其具体的评测标准请参考文献[107]。

(2) ABX测试

ABX测试主要针对转换后语音的说话人特征相似度进行转换效果评价,借鉴了说话人识别的原理。测试过程中,测评人分别测听3段语音A、B和X,并判断在语音的个性特征方面语音A还是B更接近于X。其中,X是转换后得到的语音,而A和B分别为源语音和目标语音。最后统计所有测评人员的判决结果,计算出听起来像目标语音的百分比。

4 语音转换的应用领域

语音转换之所以能够得到重视和发展,是因为它有较大的现实应用和需求,尤其是人们对个性化语音的需求越来越广泛,如导航、听书软件中希望听到自己喜欢明星的声音,希望能够和已故的亲人说说话,渴望与自己无法直接接触的偶像对话等。具体来说,语音转换可以应用在以下几方面。

4.1 在文本到语音转换系统中的应用

多数语音可以通过TTS系统来合成,但是TTS合成的语音通常比较单一,缺乏情感和个性化特征。随着科大讯飞在高德语音导航上实现合成林志玲等人的特色语音,使得个性化语音生成备受关注。但在TTS系统中想要实现个性化语音,就必须录制大量相应的语音数据,建立起与个性化语音对应的语音库,而建立一个完整的语音库,则涉及对所录语音进行切分、韵律标注、韵律调整(如对音高、音长、音强进行调节)等,费时费力,工作量相当大。而且多个个性化特征的语音库也会大大增加系统检索及数据处理的时间。使用语音转换技术可以将TTS系统生成的说话人转换为所需要的具有特定人语音特征的语音,而转换过程中所需参数及系统开销将被极大降低,可以大大提高效率[108]。目前,谷歌、微软、科大讯飞等公司都在进行相关研究,日、美、中等国的多所大学也在进行技术跟踪研究。例如:2019年5月,搜狗输入法上线了“变声功能”,能将每个人不同的声音转换成系统内置的明星、动漫、游戏等形象的声音,通过语音转换技术轻松实现了对个性化语音的合成。

4.2 在影视配音方面的应用

互联网时代自媒体高度发达,个人制作和发布的视频占有越来越大的比重,而这些视频所影响的范围也越来越广泛。利用语音转换技术可为虚拟人物角色设定个性化语音,加深角色和情景带入。同时很多动画片、电视节目等都是通过配音、讲解等赋予人物角色或节目特色的,随着配音、讲解人员的更迭,同一卡通形象或者节目难以维持一贯的风格或形象。而通过语音转换则可使经典的声音常在,实现对经典语音的重现,保持卡通形象或节目的风格。

4.3 在信息安全和情报获取方面的应用

在情报作战中,按照预期任务制作假新闻、假视频的需求不断出现。2016年,德国马克斯-普朗克

研究所(Max Planck Institute, MPI)、埃朗根-纽伦堡大学和美国斯坦福大学的研究人员提出了表情捕捉复现技术(Face reenactment)。该技术能够实时重现一个人说话时的动作和表情,并将其映射到视频中另外一个人的脸上[109]。如果将该技术与语音转换技术结合在一起,则能制作出以假乱真的目标说话人的视频,达到欺骗通信另一方的目的。可以说通过语音转换技术既可以隐藏说话人身份、保护说话人自身和信息安全,也可以迷惑敌方正常通信、扰乱敌方的情报收集,如:2017年5月,由蒙特利尔大学深度学习实验室提供技术支持的加拿大初创公司琴鸟(Lyrebird)发布了一段由其新款人工智能语音系统合成的音频 Demo,模仿了特朗普、奥巴马和希拉里3人对话的声音。据称,该软件仅需目标说话人1 min的语音,就可模仿其音色发声。如果语音转换技术可轻松实现这一功能将会对情报探测和分析带来严重的影响。

4.4 在语音识别和语音认证领域的应用

随着声纹识别认证的普及和应用,越来越多的手机软件开始采用语音认证作为用户安全登录的密码,例如微信、支付宝、手机银行等APP中均增加了语音认证模块。通过语音转换技术可以将源说话人语音转换成目标说话人语音,利用转换语音攻击语音认证系统,以此来检验系统模块的防护性和安全性,可有效提升网络防护的安全性能,如:2017年10月24日,极棒(GeekPwn)安全实验室在国际安全极客大赛期间组织了一场AI仿声验声攻防赛——AI PWN,第一次在公开场合演示了通过特色合成或语音转换制作的声音在短时间内破解预设声纹锁的设备的能力,也从侧面说明了语音转换给认证领域所带来的冲击和挑战。此外,在语音识别时,也可利用语音转换技术检验语音识别的准确性等。

4.5 其他领域的应用

除上述应用外,语音转换还广泛存在于其他领域,如:语音情感的转换,比如将悲伤情绪转换为高兴情绪;生物医学的应用,通过语音转换将患有语音障碍的人的话语转换为更易理解和交流的语音[110];在电信领域可以即时将说话人语音转换为标准语音,再对转换语音进行相应的压缩编码等处理,减少语音处理难度等。

5 存在问题及发展趋势

5.1 存在问题

语音转换虽然已经过了几十年的发展,直到近几年在神经网络和深度学习等技术发展的推动下,才有了进一步突破。新的技术带来了新的问题和挑战。目前,语音转换所面临的挑战归纳起来主要有以下几个方面:首先,对训练语音数据要求较大。无论是通过语音合成还是通过语音转换来获取带有特定目标说话人音色的语音,在技术实现上多采用统计学习中有监督学习的手段,传统的有监督学习一般需要较大规模的数据才能获得较好的效果。而在很多实际应用场景中往往难以大规模获取目标说话人语音,转换语音质量对大数据量的依赖,一定程度上制约着语音转换技术的发展,所以如何解决对训练数据量的依赖是目前乃至今后一段时间所要面临的问题。其次,对目标说话人的语料质量要求过于苛刻。对于大多数语音转换方法理论上都需要有高质量的语音数据库,但将语音转换技术真正“落地”到生活或实际应用中时,人们会发现,通常情况下难以获取高质量的目标语音,除非特定的目标说话人十分配合来制取其高质量的语音数据库,否则无论是日常生活中的随意录制,还是窃取该说话人的语音,都不可避免地受到各种噪声的污染,难以直接获得高质量的语音库。大多数情况下,由于录音环境未知,噪声及其统计特性都难以获取,给语音转换带来了新的问题。再次,语音转换算法的实时性有待提高,目前能产生高质量语音的转换算法实时性差,为使语音转换能实际应用于特定场合,既需要转换后的语音质量足够高,还需要转换过程足够快。而当前语音转换方案在转换质量和实时处理的

平衡性方面还存在一定的发展空间。此外,算法软件在不同设备、不同平台之间的可移植性也是制约语音转换实际应用的一个关键问题。同时,转换语音质量还可进一步提高,虽然目前语音转换质量相较于之前有了较大的提升和改善,但是与真正目标语音相比仍存在一定差距,语音中的情感、韵律等描述特定说话人的属性,还很难在近期的特色语音合成中被有效地量化建模。而这恰恰也是人们直观感受上觉得某种语音与目标说话人发音相似的关键之处。因此,如何进一步提升转换语音与目标语音的相似度是一个值得持续关注的问题。最后还有非平行语料的转换问题。目前大多数语音转换算法依赖于平行语料,然而要使语音转换成为一种主流,非平行语料的转换则必不可免,因为在实际生活中,更多情况下人们获取的是非平行语料,录制大规模平行语料显然不符合实际要求。虽然目前针对非平行语料的语音转换已经有相当一部分研究,但是相比于平行语料的转换还有一定差距,所以如何解决非平行语料的转换,也是一个需要重视的方面。

总之,除了上述几个主要方面外,语音转换还存在其他有待提高和完善的地方,如:动态模型的转换,能够将一个转换模型迅速调整使其适应其他转换场景;多对一的转换,将不同说话人通过模型转换成统一目标说话人语音;完善的韵律建模,构建一个更加符合实际的韵律转换模型,实现情感语音的转换等。这些都是语音转换中尚未解决并值得进一步研究的问题。

5.2 发展趋势

虽然目前语音转换中还存在很多问题和挑战,但语音转换实际需求也不断扩大,未来语音转换将着力解决当前语音转换中存在的现实问题,朝着下述方向不断发展。

5.2.1 鲁棒语音转换

当前关于源说话人或目标说话人语音中含噪声的语音转换的直接研究不多。Masaka 等[111]在非负矩阵分解框架下,借助唇动等视觉信息对源说话人语音中含噪声情况下的语音转换进行了研究。Aihara 等[112]在 Masaka 工作的基础上,将源说话人语音中的噪声模型和少量平行语料的自适应仿射变换相结合,研究了小样本情况下,源说话人语音含噪声情况下的语音转换。针对含噪语音,或许先通过处理得到较纯净语音,然后再进行语音转换,目前已有相关学者进行这方面的研究。

5.2.2 小样本训练的语音转换

前文提到目前多数语音转换效果在一定程度上依赖于训练语音数据库的规模,训练数据集规模大则转换语音效果高,否则转换效果较差。未来的研究方向必定是小样本语音的转换,通过较少的数据实现高质量的语音转换,通过半监督或者无监督的网络来训练生成新的样本数据,然后提升语音转换效果。

5.2.3 实时语音转换

训练数据越多,提取映射函数的时间越久,转换语音耗费的时间越长。减小网络规模,实现语音实时转换将成为必要之选。因此,神经网络或深度学习模型的瘦身和加速是未来语音转换模型发展不可或缺的环节。近年来,通过减枝、权重共享等技术,深度神经网络模型的压缩取得了较大进展,相信针对这方面的研究也会逐步深入。

此外,轻量化模型、多对多建模及非合作式等语音转换等也将成为未来语音转换发展的趋势。

6 结束语

随着智能语音交互应用的不断发展,语音转换技术的不断提高,人们对于特定说话人语音的生成有着越来越高的需求和期望;在人工智能时代,个人媒体制作、声纹认证等越来越普及,语音代表个人身份特征的场景日益广泛。未来的语音转换技术一定会朝着转换模型更小、转换效率更高、转换效果更好、转换速度更快的方向发展,也必将进一步推动语音转换技术在其他领域的应用和发展。

参考文献:

[1] 岳振军. 语音变换关键技术与算法研究[D]. 南京:解放军理工大学,2009.
Yue Zhenjun. Research on key technologies and algorithms of voice conversion[D]. Nanjing: PLA University of Science and Technology, 2009.

[2] Kuwabara H, Sagisak Y. Acoustic characteristics of speaker individuality: Control and conversion[J]. Speech Communication, 1995, 16(2): 165-173.

[3] Stylianou Y. Voice transformation[M]// Handbook of Speech Processing.[S.l.]: Springer, 2008: 489-504.

[4] Stylianou Y. Voice transformation: A survey[C]// Proc IEEE ICASSP 09. Taipei, China: [s.n.], 2009: 3585-3588.

[5] Machado A F, Marcelo Q. Voice conversion: A critical survey[C]// SMC. [S.l.]: [s.n.], 2010.

[6] Abe M. A segment - based approach to voice conversion[C]// Acoustics, Speech, & Signal Processing, International Conference . [S.l.]: IEEE Computer Society, 1991: 765-768.

[7] Moulines E, Sagisaka Y, Sorinetc C. Voice conversion: State of the art and perspectives[J]. Special Issue of Speech Communication, 1995, 16(2): 125-224.

[8] Mohammadi S H, Kain A. An overview of voice conversion systems[J]. Speech Communication, 2017, 88: 65-82.

[9] 孙新建. 基于说话人特征替换的语音转换技术研究[D]. 南京:解放军理工大学,2013.
Sun Xinjian. Research on voice conversion based on speaker characteristics replacement[D]. Nanjing: PLA University of Science and Technology, 2013.

[10] Matsumoto H, Hiki S, Sone T, et al. Multidimensional representation of personal quality of vowels and its acoustical correlates[J]. IEEE Trans, 1973, AU-21: 428-436.

[11] Furui S. Research on individuality features in speech waves and automatic speaker recognition techniques[J]. Speech Communication, 1986, 5(2): 183-197.

[12] 李波. 语音转换的关键技术研究[D]. 长沙:国防科学技术大学,2005.
Li Bo. Studies on key technologies of voice conversion[D]. Changsha: National University of Defense Technology, 2005.

[13] Chen L H, Liu L J, Ling Z H, et al. Neural network based approaches for spectrum, Aperiodicity and F0 conversion[C]// Interspeech. [S.l.]: [s.n.], 2016.

[14] 李杨春, 俞一彪. 倒谱本征空间结构化高斯混合模型语音转换方法[J].声学学报, 2015(1): 12-19.
Li Yangchun, Yu Yibiao. Voice conversion using structured Gaussian mixture model in eigen space[J]. Acta Acustica, 2015(1): 12-19.

[15] Abe M, Nakamura S, Shikano K, et al. Voice conversion through vector quantization[C]// Proc IEEE ICASSP 88. New York, USA:[s.n.], 1988.

[16] 孙健. 语音转换中声道谱技术研究[D]. 南京:解放军理工大学,2012.
Sun Jian. Research on vocal track spectrum modification technology in voice conversion[D]. Nanjing: PLA University of Science and Technology, 2012.

[17] Shikano K, Nakamura S, Abe M. Speaker adaptation and voice conversion by codebook mapping[C]//Proc IEEE ISCS. [S.l.]: [s.n.], 1991: 594-597.

[18] Kim E K, Lee S, Oh Y H. Hidden Markov Model based voice conversion using dynamic characteristics of speaker[C]//Proc Eurospeech. Rhodes, Greece: [s.n.], 1997.

[19] Arslan L M. Speaker transformation algorithm using segmental codebooks (STASC)[J]. Speech Communication, 1999, 28(3): 211-226.

[20] Wang Y P, Ling Z H, Wang R H. Emotional speech synthesis based on improved codebook mapping voice conversion[C]// Proc ACII. Beijing, China: [s.n.], 2005.

[21] Eslami M, Sheikhzadeh H, Sayadiyan A. Quality improvement of voice conversion systems based on trellis structured vector quantization[C]// Proc Interspeech. Florence, Italy: [s.n.], 2011: 665-668.

[22] 胡芳, 徐宁, 李海燕. 基于码书映射的语音转换算法改进[J]. 微处理机, 2015(2): 35-38.
Hu Fang, Xu Ning, Li Haiyan. Improvement of voice conversion algorithm based on codebook mapping[J]. Microprocessors,

2015(2): 35-38.

[23] Duda R O, Hart P E. Pattern classification and scene analysis[M]. [S.l.]: John Wiley & Sons, 1973.

[24] 简志华, 杨震. 语声转换技术发展及展望[J]. 南京邮电大学学报(自然科学版), 2007, 27(6): 88-94.
Jian Zhihua, Yang Zhen. An overview of voice conversion[J]. Nanjing University of Posts and Telecommunications (Natural Science), 2007, 27(6): 88-94.

[25] Stylianou Y, Cappe O, Moulines E. Continuous probabilistic transform for voice conversion[J]. Speech and Audio Processing, IEEE Transactions on, 1998, 6(2): 131-142.

[26] Chen Y, Chu M, Chang E, et al. Voice conversion with smoothed GMM and MAP adaption[C]// Proc EUROSPEECH. Geneva, Switzerland: [s.n.], 2003: 2413-2416.

[27] 申毅, 简志华, 杨震. 改进的GMM模型语声转换系统[J]. 南京邮电大学学报(自然科学版), 2007, 27(5): 11-15.
Shen Yi, Jian Zhihua, Yang Zhen. A modified method for voice conversion based on GMM[J]. Nanjing University of Posts and Telecommunications (Natural Science), 2007, 27(5): 11-15.

[28] Toda T, Black A W, Tokuda K. Voice conversion based on Maximum-likelihood estimation of spectral parameter trajectory[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(8): 2222-2235.

[29] Helander E, Virtanen T, Nurminen J, et al. Voice conversion using partial least squares regression[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010,18(5): 912-921.

[30] Helander E, Silën H, Virtanen T, et al. Voice conversion using dynamic kernel partial least squares regression[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(3): 806-817.

[31] 宋鹏, 王浩, 赵力. 基于混合Gauss归一化的语音转换方法[J]. 清华大学学报(自然科学版), 2013(6): 757-761.
Song Peng, Wang Hao, Zhao Li. Voice conversion based on a mixture of Gaussian normalization methods[J]. Journal of Tsinghua University (Science and Technology), 2013(6): 757-761.

[32] 王明明. 基于GMM和码本映射相结合的语音转换方法[D]. 西安:西安建筑科技大学,2015.
Wang Mingming. Voice conversion based on GMM and codebook mapping[D]. Xi'an: Xi'an University of Architecture and Technology, 2015.

[33] Kobayashi K, Toda T, Nakamura S. Implementation of F0 transformation for statistical singing voice conversion based on direct waveform modification[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2016.

[34] Duxans H, Bonafonte A, Kain A, et al. Including dynamic and phonetic information in voice conversion systems[C]// Proc ICSLP. Jeju Island, Korea: [s.n.], 2004.

[35] Wu C H, Hsia C C, Liu T H, et al. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis[J]. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(4): 1109-1116.

[36] Qiao Y, Saito D, Minematsu N. HMM-based sequence-to-frame mapping for voice conversion[C]// Proc ICASSP. Dallas, Texas, USA: [s.n.], 2010: 4830-4833.

[37] Zen H, Nankaku Y, Tokuda K. Continuous stochastic feature mapping based on trajectory HMMs [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(2): 417-430.

[38] Shuang Z, Bakis R, Qin Y. Voice conversion based on mapping formants[J]. TC-STAR Workshop on Speech-to-Speech Translation, 2006.

[39] Valbret H, Moulines E, Tubach J P. Voice transformation using PSOLA technique[J]. Speech Communication, 1992, 11(2/3): 175-187.

[40] Sündermann D, Strecha G, Bonafonte A, et al. Evaluation of VTLN-based voice conversion for embedded speech synthesis[C]// INTERSPEECH. Lisbon, Portugal: [s.n.], 2005.

[41] 双志伟, 张世磊, 秦勇. 基于频谱弯曲的语音转换相似度改进[C]//全国人机语音通讯学术会议. [S.l.]: [s.n.], 2009.
Shuang Zhiwei, Zhang Shilei, Qin Yong. Similarity improvement of voice conversion based on spectrum warping[C]// National Conference on Man-Machine Speech Communication, NCMMSC2009. [S.l.]: [s.n.], 2009.

[42] Erro D, Agustín Alonso, Serrano L, et al. Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations[J]. Computer Speech Lang, 2015,30 (1): 3-15.

[43] Erro D, Agustín Alonso, Serrano L, et al. Towards physically interpretable parametric voice conversion functions[J]. Advances in Nonlinear Speech Processing, 2013(3): 75-82.

[44] Tian X, Wu Z, Lee S W, et al. Sparse representation for frequency warping based voice conversion[C]// Proceedings of the ICASSP. [S.l.]: [s.n.], 2015.

[45] 李金中, 李贤, 汪增福. 基于声道长度对齐的年龄语音转换[J]. 中国科学技术大学学报, 2015(7): 575-581.
Li Jinzhong, Li Xian, Wang Zengfu. Vocal tract length aligning based mandarin age voice conversion[J]. Journal of University of Science and Technology of China, 2015(7): 575-581.

[46] Tian X, Wu Z, Lee S W, et al. System fusion for high-performance voice conversion[C]// Proceedings of INTERSPEECH. [S.l.]: [s.n.], 2015.

[47] He Ming. University computer foundation[M]. Nanjing: Southeast University Press, 2015.

[48] Drioli C. Radial basis function networks for conversion of sound spectra[J]. EURASIP Journal on Applied Signal Processing, 2001(1): 36-44.

[49] Desai S, Black A, Yegnanarayana B, et al. Spectral mapping using artificial neural networks for voice conversion[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(5): 954-964.

[50] Narendranath M, Murthy H A, Rajendran S, et al. Transformation of formants for voice conversion using artificial neural networks [J]. Speech Communication, 1995, 16(2): 207-216.

[51] Huang W C, Wu Y C, Lo C C. et al. Investigation of F0 conditioning and fully convolutional networks in variational autoencoder based voice conversion[EB/OL]. (2019-05-02). https://arXiv:1905.00615.

[52] Kaneko T, Kameoka H, Hiramatsu K, et al. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks[C]// INTERSPEECH . [S.l.]: [s.n.], 2017.

[53] Kaneko T, Kameoka H, Hojo N, et al. Generative adversarial network-based post-filter for statistical parameter synthesis[C]// ICASSP . [S.l.]: [s.n.], 2017.

[54] Huang D Y, Xie L, Lee S W, et al. An automatic voice conversion evaluation strategy based on perceptual background noise distortion and speaker similarity[C]// The 9th ISCA Workshop on Speech Synthesis (SSW9). [S.l.]: [s.n.], 2016.

[55] Esteban C, Hyland S L, Ratsch G. Real-valued (medical) time series generation with recurrent conditional GANs[EB/OL]. (2017-06-08). https://arxiv.org/abs/1706.02633v2.

[56] Takuhiro K, Hirokazu K, Kou T, et al. CycleGAN-VC2: Improved cycleGAN-based non-parallel voice conversion[EB/OL]. (2019-04-09). https://arxiv.org/abs/1904.04631.

[57] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09). https://arxiv.org/abs/1508.01991.

[58] Hsu C C, Hwang H T, Wu Y C, et al. Dictionary update for NMF-based voice conversion using an encoder-decoder network[C]//International Symposium on Chinese Spoken Language Processing (ISCSLP). 2016, doi: 10.1109/iscslp.2016.7918382.

[59] Seyed H M, Alexander K. Voice conversion using deep neural networks with speaker-independent pre-training[C]//Spoken Language Technology Workshop. [S.l.]: [s.n.], 2014.

[60] Chen L H, Ling Z H, Liu L J, et al. Voice conversion using deep neural networks with layer-wise generative training[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1859-1872.

[61] Oord A V D, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio[EB/OL]. (2016-09-12). https://arxiv.org/abs/1609.03499v1.

[62] Niwa J, Yoshimura T. Statistical voice conversion with WaveNet vocoder[C]// Proc IEEE ICASSP. [S.l.]: [s.n.], 2018: 5289-5293.

[63] Chen K, Chen B, Lai J H, et al. High-quality voice conversion using spectrogram-based WaveNet vocoder[J]. Interspeech, 2018, 9 (6): 1993-1997.

[64] Oord A V D, Li Y, Babuschkin I, et al. Parallel WaveNet: Fast high-fidelity speech synthesis[EB/OL]. (2017-11-28). https://arxiv.org/abs/1711.10433.

[65] Hayashi T, Tamamori A, Kobayashi K, et al. An investigation of multi-speaker training for WaveNet vocoder[C]// Automatic

Speech Recognition and Understanding Workshop (ASRU). [S.l.]: IEEE, 2017: 712-718.

[66] Shen J, Pang R, Weiss R J, et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions[EB/OL]. (2017-12-16). https://arxiv.org/abs/1712.05884.

[67] Wu Y C, Kobayashi K, Hayashi T, et al. Collapsed speech segment detection and suppression for WaveNet vocoder[EB/OL]. (2018-04-30). https://arxiv.org/abs/1804.11055v2.

[68] Wu Z, Virtanen T, Chng E S, et al. Exemplar-based sparse representation with residual compensation for voice conversion[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1506-1521.

[69] Li Na, Zeng Xiangyang, Qiao Yu, et al. Voice conversion using Bayesian analysis and dynamic kernel feature[J]. Acta Acustica, 2015(3): 455-461.

[70] 宋鹏. 个性化语音生成及其相关问题的研究[D]. 南京:东南大学, 2014.
Song Peng. Research on personalized speech generation and its related factors[D]. Nanjing: Southeast University, 2014.

[71] 孙健, 张雄伟, 曹铁勇, 等. 基于卷积非负矩阵分解的语音转换方法[J]. 数据采集与处理, 2013, 28(2): 141-148.
Sun Jian, Zhang Xiongwei, Cao Tieyong, et al. Voice conversion based on convolutive nonnegative matrix factorization[J]. Journal of Data Acquisition & Processing, 2013, 28(2): 141-148.

[72] 孙新建, 张雄伟, 杨吉斌, 等. 基于隐变量模型的语音转换方法研究[J]. 信号处理, 2012(3): 344-351.
Sun Xinjian, Zhang Xiongwei, Yang Jibin, et al. Voice conversion using latent variable model[J]. Signal Processing, 2012(3): 344-351.

[73] 马振, 张雄伟, 杨吉斌. 基于语音个人特征信息分离的语音转换方法研究[J]. 信号处理, 2013, 29(4): 513-519.
Ma Zhen, Zhang Xiongwei, Yang Jibin. A speech conversion method based on the separation of speaker-specific characteristics[J]. Signal Processing, 2013, 29(4): 513-519.

[74] Zhou C, Horgan M, Kumar V, et al. Voice conversion with conditional SampleRNN[EB/OL]. (2018-08-24). https://arxiv.org/abs/1808.08311v1.

[75] Sun L, Kang S, Li K, et al. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2015: 4869-4873.

[76] 解伟超. 语音转换中声道谱参数和基频变换算法的研究[D]. 南京:南京邮电大学, 2013.
Xie Weichao. The research on vocal tract spectrum and pitch frequency transformation in voice conversion[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2013.

[77] Ye H, Young S. High quality voice morphing[C]// IEEE ICASSP 04. Montreal, Quebec, Canada: IEEE, 2004.

[78] Percybrooks W S, Moore E. New algorithm for LPC residual estimation from LSF vectors for a voice conversion system[C]// Proc ICSLP. [S.l.]: [s.n.], 2004: 85-88.

[79] Kang Y, Tao J, Xu B. Applying pitch target model to convert F0 contour for expressive mandarin speech synthesis[C]// IEEE International Conference on Acoustics. [S.l.]: IEEE, 2006.

[80] Ming H P, Huang D Y, Xie L, et al. Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion[J]. Interspeech, 2016: 2453-2457.

[81] 凌震华, 高丽, 戴礼荣. 基于目标逼近特征和双向联想贮存器的情感语音基频转换[J]. 天津大学学报(自然科学与工程技术版), 2015(8): 670-674.
Ling Zhenhua, Gao Li, Dai Lirong. F0 transformation for emotional speech synthesis using target approximation features and bi-directional associative memories[J]. Journal of Tianjin University (Science and Technology), 2015(8): 670-674.

[82] Inanoglu Z. Transforming pitch in a voice conversion framework[D]. Cambridge, U.K.: Univ of Cambridge, 2003.

[83] Helander E, Nurminen J. A novel method for prosody prediction in voice conversion[C]// Proc ICASSP. Honolulu, Hawaii, USA: [s.n.], 2007: 509-512.

[84] Erro D. Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models[D]. Barcelona: Universitat Politècnica de Catalunya, 2009.

[85] Markel J D, Gray A H. Linear prediction of speech [M]. New York: Springer-Verlag, 1976.

[86] Kortekaas R W, Reinier W L. Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform

manipulation technique using single-formant stimuli[J]. The Journal of the Acoustical Society of America, 1997, 101(4): 2202.

[87] Verhelst W, Roelands M. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech[C]// Proc ICASSP 93. Minneapolis, Minnesota, USA:[s.n.], 1993: 554-557.

[88] Stylianou Y, Laroche J, Moulines E. High-quality speech modification based on a harmonic+noise model[C]// Proc EUROSPEECH. [S.l.]: [s.n.], 1995.

[89] Kawahara H. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited[C]// IEEE ICASSP 97. Munich, Bavaria, Germany: IEEE, 1997: 1303-1306.

[90] Erro D, Navas E, Hernáez I. Iterative MMSE estimation of vocal tract length normalization factors for voice transformation[C]// IEEE International Conference on Image Processing. [S.l.]: IEEE, 2012: 175-178.

[91] Morise M, Yokomori F, Ozawa K. A vocoder based high-quality speech synthesis system for real-time applications[J]. Leice Transactions on Information & Systems, 2016, 99 (7) : 1877-1884.

[92] 李波,王成友,蔡宣平,等. 语音转换及相关技术综述[J]. 通信学报,2004(25): 109-118.
Li Bo, Wang Chengyou, Cai Xuanping, et al. A survey of voice conversion and its relevant technology[J]. Journal on Communications, 2004(25): 109-118.

[93] Arslan L M, Talkin D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum[C]// Proc of the Eurospeech '97. Rhodes, Greece:[s.n.], 1997.

[94] Bae H S, Lee H J, Lee S G . Voice recognition based on adaptive MFCC and deep learning[C]// 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA). [S.l.]: IEEE, 2016.

[95] Kain A. High resolution voice transformation[D]. [S.l.]: OGI School of Science and Engineering, 2001.

[96] Duxans H, Erro D, Perez J, et al. Voice conversion of non-aligned data using unit selection[C]// Proc TC-STAR Workshop on Speech-to-Speech Translation. Barcelona, Spain: [s.n.], 2006.

[97] Sündermann D, Bonafonte A, Ney H, et al. A first step towards text-independent voice conversion[C]// Proc ICSLP'04. Jeju Island, South Korea:[s.n.], 2004.

[98] Sündermann D, Höge H, Bonafonte A, et al. Text-independent voice conversion based on unit selection[C]// Proc IEEE ICASSP 06. Toulouse, France:[s.n.], 2006.

[99] Salor Ö, Demirekler M. Dynamic programming approach to voice transformation[J]. Speech Communication, 2006, 48(10): 1262-1272.

[100] Saito D, Watanabe S, Nakamura A, et al. Statistical voice conversion based on noisy channel model[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(6): 1784-1794.

[101] 简志华, 杨震. 基于混合线性变换的语声转换算法[J]. 电子与信息学报, 2007, 29(2): 1700-1702.
Jian Zhihua, Yang Zhen. An algorithm for voice conversion based on mixtures of linear transformation[J]. Journal of Electronics & Information Technology, 2007, 29(2): 1700-1702.

[102] 简志华, 王向文. 一种用于语音转换的区域最近邻迭代训练算法[J]. 电子与信息学报, 2012, 34(9): 2091-2096.
Jian Zhihua, Wang Xiangwen. An iterative training algorithm based on local nearest neighbor for voice conversion[J]. Journal of Electronics & Information Technology, 2012, 34(9): 2091-2096.

[103] Tian X, Chng E S, Li H. A vocoder-free WaveNet voice conversion with non-parallel data[EB/OL]. (2019-02-11). https://arxiv.org/abs/1902.03705.

[104] Kameoka H, Kaneko T, Tanaka K, et al. ACVAE-VC non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder[EB/OL]. (2018-06-06). https://arxiv.org/abs/1806.02169.

[105] Yeh C C, Hsu P C, Chou J C, et al. Rhythm-flexible voice conversion without parallel data using cycle-GAN over Phoneme posteriorgram sequences[EB/OL]. (2018-08-01). https://arxiv.org/abs/1808.0311.

[106] Wu Y, Tobing P L, Hayashi T, et al. NU non-parallel voice conversion system for the voice conversion challenge 2018[C]// Proc Odyssey Speaker Lang Recognit Workshop. [S.l.]: [s.n.], 2018.

[107] Shuang Z, Bakis R, Qin Y. IBM voice conversion systems for 2007 TC-STAR evaluation[J]. Tsinghua Science & Technology, 2008, 13(4): 510-514.

[108] Tamura M, Masuko T, Tokuda K, et al. Speaker adaptation for HMM-based speech synthesis system using MLLR[C]// Proc

of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis. Jenolan Caves, Australia: [s.n.], 1998.

[109] Thies J, Zollhofer M, Stamminger M, et al. Face2Face: Real-time face capture and reenactment of RGB videos[C]// ACM SIGGRAPH 2016 Emerging Technologies. [S.l.]: ACM, 2016.

[110] Toda T, Nakamura K, Saruwatari H, et al. Alaryngeal speech enhancement based on one-to-many eigenvoice conversion[J]. IEEE/ACM IEEE Trans on Audio Speech Lang. Process, 2014, 22(1): 172-183.

[111] Masaka K, Aihara R, Takiguchi T, et al. Multimodal voice conversion using non-negative matrix factorization in noisy environments[C]// IEEE International Conference on Acoustics. [S.l.]: IEEE, 2014.

[112] Aihara T, Fujii T, Nakashika T, et al. Small-parallel exemplar-based voice conversion in noisy environment using affine non-negative matrix factorization[J]. Audio Speech & Music Processing, 2015(1): 75.

作者简介:

张雄伟(1965-),男,教授,博士生导师,研究方向:语音与图像处理、智能信息处理,E-mail:xwzhang9898@163.com。

苗晓孔(1991-):男,博士研究生,研究方向:语音转换,E-mail:miao_xk@163.com。

曾歆(1995-):男,硕士研究生,研究方向:语音转换,E-mail:245104441@qq.com。

孙蒙(1984-):男,副教授,研究方向:智能语音处理、机器学习,E-mail:sunmengccjs@163.com。

曹铁勇(1971-):男,教授,研究方向:智能信息处理、图像处理,E-mail:cty_ice@sina.com。

(编辑:张彤)